International Journal of Engineering Trends and Technology
ISSN: 2231–5381 / https://doi.org/10.14445/22315381/IJETT-V73I12P120

Volume 73 Issue 12, 244-265, December 2025
© 2025 Seventh Sense Research Group®

Original Article

A Novel Multifaceted and Multitargeted Approach to Predict the Efficacy of New SMILE for NSCLC using Graph Attention Networks

Sandhi Kranthi Reddy[1], S V G Reddy[2]

[1,2]Department of CSE, GST, GITAM (Deemed to be University), Visakhapatnam, Andhra Pradesh, India.
[1]Department of CSE, Vignan Institute of Technology and Science, Telangana, India.

[1]Corresponding Author : kranthi.research@gmail.com

Received: 02 October 2025 Revised: 05 December 2025 Accepted: 09 December 2025 Published: 19 December 2025

Abstract - *NSCLC - Non-Small Cell Lung Cancer, which holds almost 85% cases of lung cancer, is one of the deadliest diseases worldwide and a leading cause of death related to cancer. Types of NSCLC are Adenocarcinoma, Large cell carcinoma, and Squamous cell carcinoma. Among these, adenocarcinomas account for 40%-50% of NSCLC cases that occur more among youngsters, non-smokers, and East Asians and are often diagnosed at advanced stages, which remains a challenge for their better treatment. NSCLC occurs due to a wide range of targetable alterations, among which EGFR, ALK, KRAS, and PDGFR account for numerous cases. The emergence of artificial intelligence has accelerated the early detection of NSCLC using various machine learning and deep learning models based on numerical or image datasets, but there is a huge requirement to shift the focus to identifying a novel drug that could work effectively at an early or advanced stage. Existing drugs may become resistant after some time, and there will always be a huge requirement to develop a new drug, which perhaps requires a lengthy amount of time and more cost using traditional approaches, and it is even a risky process since 97% of drug discoveries fail. Hence, it is necessary to build and use machine learning or deep learning models to estimate the ability of a new drug as a part of lead identification before moving to further processing. To address this, a multifaceted and multitargeted approach using Graph Attention Networks has been proposed, designing a model that is trained using 15 FDA-approved drugs and a vast library of 1.048 million drug molecules to predict the efficiency of a new drug, which achieved 89% accuracy. In the drug discovery process, this highlights the potential of deep learning, which provides enhanced, cost-effective, and efficient means to identify novel drugs for the treatment of NSCLC.*

Keywords - *NSCLC, EGFR, ALK, KRAS, PDGFR, Deep Learning, Graph Attention Network.*

1. Introduction

Lung cancer(LC) is one of the leading causes of cancer-related deaths worldwide [1]. In 2020, LC was the second most common cancer in India, accounting for 11.4% cancer cases and 18% cancer-related deaths [2].

In 2022, it had the highest number of new cases globally, accounting for 2.5 million, representing 12.4% of all cancer cases, and also led to deaths of about 1.8 million, representing 18.7% of all cancer-related deaths, as depicted in Figures 1(a) and 1(b). Among males, lung cancer ranked as the most commonly diagnosed cancer, while among females, it held the second position. Also, it had the highest rates in Asia according to the region-wise analysis [3], and if it continues at the same rate as in 2022, the number of new lung cancer cases may increase to about 4.62 million and deaths to about 3.55 million by 2050 [4]. Smoking remains the leading cause of LC, while other risks include exposure to biomass smoke, asbestos, arsenic, and radon, particularly in poorly ventilated homes or unsafe workplaces [5]. Non-Small Cell LC (NSCLC) and Small Cell LC (SCLC) are the common types of LC, among which the most common is NSCLC, representing approximately 85% of cases, and grows slowly [6]. Historically, NSCLC has been associated with poor outcomes due to limited options for treating as well as late diagnosis [7].

NSCLC comprises three subtypes: adenocarcinoma, squamous cell carcinoma, and large cell carcinoma. Adenocarcinoma represents 40% of NSCLC cases and 34% of LC, squamous cell carcinoma accounts for about 25-30% and 23% of LC, and large cell carcinoma accounts for about 5-10% and 6% of LC [8]. Among these subtypes, adenocarcinoma's high prevalence highlights the need to focus more on it to improve the treatment and management of NSCLC.

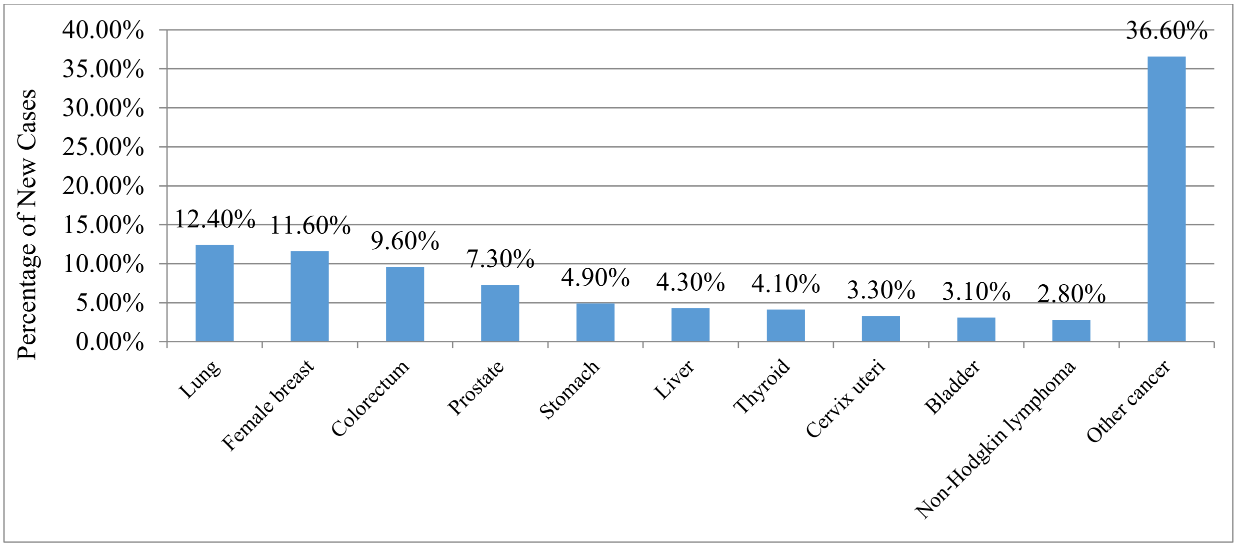


Fig. 1(a) Distribution of 20 million global cancer cases in 2022

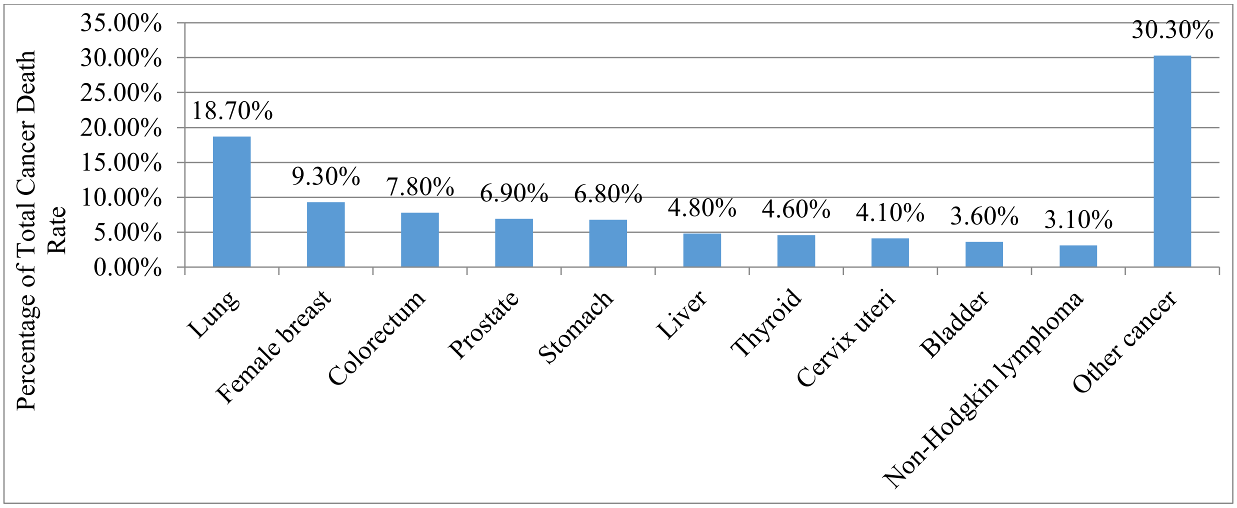


Fig. 1(b) Distribution of 9.7 million global cancer-related deaths in 2022

Molecular analysis is a key aspect in NSCLC management, transforming the process of diagnosis, prognosis, and treatment. Next-generation sequencing (NGS)-based molecular testing has identified alterations in NSCLC, such as EGFR, ALK, KRAS, MET, ROS1, BRAF, RET, and ERBB2, that influence both disease progression and response to treatment over time, particularly with FDA-approved targeted drugs [9, 10].

The rapid growth in AI, especially deep learning, has enhanced the process of drug discovery, predictions of cancer alteration and survival rate [11-15]. These technological improvements allow us to do many tasks in minimal time compared to traditional approaches, such as analyzing volumes of molecular data and predicting how a new drug might work. However, despite these advancements, treating NSCLC remains a major challenge because of the difficulty in identifying the alterations and development of resistance among patients over a period of time [16]. Hence, there is a requirement to discover new drugs that are effective and capable of targeting multiple alterations. To address this, two identifications are required:

i) Identifying the key alterations that cause NSCLC,
ii) identifying the most effective FDA-approved drugs, and determining which new drugs can be developed.

Traditional approaches to developing new drugs take several years and are expensive, and the majority of drugs fail during clinical trials [17]. However, the recent integration of molecular biology and computational techniques, especially in implementing Deep Learning (DL), has greatly advanced drug discovery. By analyzing huge amounts of data, DL models identify the complex hidden patterns and find potential drug candidates, which enables rapid and efficient drug discovery, especially in the critical phase of lead identification. Among the various DL techniques, Graph Neural Networks (GNNs) are gaining significance because of their ability to model drug molecules in a graph-based structure, where atoms are represented as nodes and chemical bonds as edges. GNNs are used to identify the important features from molecular graphs that are critical to predict the drug effectiveness. Graph Attention Networks (GATs) are a type of GNNs that are capable of finding insights from huge, complex data represented in graphs due to their attention mechanisms [18, 19]. In the field of drug discovery, they are widely used for various tasks, such as the prediction of the effectiveness of drug combinations and drug-target interactions [20, 21].

This paper proposes an efficient approach using GATs to predict the efficacy of a new SMILE/drug that targets multiple alterations causing NSCLC, using the corresponding drugs approved by the FDA as references before proceeding to further analysis. This helps in reducing the time taken when compared to traditional approaches.

2. Literature Review

To develop an efficient framework that predicts the efficacy of an SMILE, it is necessary to identify the most common key alterations that cause NSCLC, and FDA-approved drugs that are effective against those key alterations. It is also required to review the recent and similar works in NSCLC using ML or DL techniques.

2.1. Key Alterations in NSCLC

The IASCLC conducted a survey in 2020, from the survey, it was identified that among several alterations, three are high in number(i.e., KRAS-68%, ALK-83% and EGFR-94%). These findings highlight the consideration of alterations as potential targets [23]. The importance of considering these three alterations is further supported by the many recent studies. Kumar and Kumar (2022) have studied various alterations in NSCLC, including the above three top alterations, and specified that target-specific treatments provide better survival outcomes. They also specified the need to identify new drugs [24]. De Jong et al. (2023) specified that identifying the key alterations provides improved diagnosis as well as treatment with better survival in NSCLC [25]. Similarly, Choi and Chang (2023) specified that there are better survival rates for the approved treatments that have targeted EGFR and ALK [26]. Sharma et al. (2025) analyzed 5,219 lung cancer patients' data in India, by considering 13 alterations, and found that TP53, EGFR, KRAS, and ALK were the most occurring alterations [27]. The FDA has not approved drugs for TP53 alteration for any type of cancer [28]. All the above studies confirmed that EGFR, ALK, and KRAS are the most important alterations in NSCLC. Platelet-Derived Growth Factor Receptor (PDGFR) is emerging as a key alteration in NSCLC [29, 30]. Various Studies have proved its importance in NSCLC, specifying its role in tumor progression and treatment resistance [31]. Even in preclinical models, blocking PDGFR has shown good results in reducing the growth of the tumor and also improving its effectiveness towards resistance [32-34]. Clinical trials have also shown encouraging results, further supporting their potential use in NSCLC therapy [35].

All the above findings highlight that the EGFR, ALK, KRAS, and PDGFR are key alterations in NSCLC.

2.2. FDA-Approved Drugs that Effectively Target Key Alterations in NSCLC

2.2.1. EGFR

1st generation TKIs (Tyrosine Kinase Inhibitors), Gefitinib and Erlotinib, that target EGFR mutations often develop resistance; their 2nd and 3rd generation TKIs, Afatinib, Dacomitinib, and Osimertinib, have improved survival rates compared to 1st generation TKIs [36].

2.2.2. ALK

Similarly, 1st generation TKI, Crizotinib, targeting ALK rearrangements also acquired resistance, and their 2nd and 3rd generation TKIs, Ceritinib, Alectinib, Brigatinib, Lorlatinib, and Ensartinib, have improved survival rates compared to 1st generation TKIs [37, 38].

2.2.3. KRAS

KRAS mutations are historically considered undruggable, but now have Sotorasib and Adagrasib as FDA-approved inhibitors that are equally potent [39].

2.2.4. PDGFR

There are FDA-approved drugs targeting PDGFR in other cancers but not in NSCLC. So we reviewed the potential of PDGFR-targeting drugs that could efficiently help in treating NSCLC. Avapritinib is a potential drug targeting PDGFR in Gastrointestinal Stromal Tumors (GISTs). It has proven its efficiency in response rates and survival rates. Its structural features could help the design of next-generation inhibitors capable of overcoming drug resistance with minimal side effects. [40, 41], Hence, it can be considered as one of the reference drugs for targeting PDGFR in NSCLC. Anlotinib was approved by the FDA in China but not by the FDA in the USA, and it has shown better antitumor activity in patients with previously treated advanced NSCLC [42, 43]. Nintedanib has shown a survival benefit when compared to other TKIs targeting PDGFR in phase III trials conducted in advanced NSCLC [44]. In an NSCLC mouse model, crenolanib [45] and imatinib [46] have shown antitumor activity.These findings highlight that the selected FDA-approved TKIs targeting EGFR, ALK, KRAS, and PDGFR offer a strong foundation for designing a model to evaluate the efficacy of new multitargeted drugs.

2.3. Recent Advancements in NSCLC using AI

Li et al. (2021) and Christie et al. (2021) studies show that AI improved diagnosis, treatment planning, and evaluating response and prediction of survival [47, 48].

Wang (2022) performed a review on the Deep learning methods used for diagnosing lung cancer, especially in classifying nodules. The research indicates that the deep learning models, particularly convolutional neural networks, enhance precision, sensitivity, and overall diagnostic efficacy through quicker and specific image analysis [49].

Wankhade and Vigneshwari (2023) have proposed a method called CCDC-HNN (Cancer Cell Detection using Hybrid Neural Network). This method combines 3D-CNN (Convolutional Neural Networks) for identifying features and RNN (Recurrent Neural Networks) for extracting and classifying lung nodules as non-cancerous or cancerous. The method was performed on the LUNA16 dataset, which contains 888 patient CT scan images, and it has achieved an accuracy-95%, a specificity-90%, and a sensitivity-87% [50].

Nakarin et al. (2022) designed a deep learning model using principal neighborhood aggregation to predict the binding affinity of SMILES across seven targets (ALK, EGFR, ERBB2, ERBB4, MET, RET, and ROSI) of NSCLC. The model is trained on a labeled dataset of 16,345 unique SMILES with their binding affinities to targets. The model generated R^2 scores ranging from 0.4344-ERBB4 to 0.7280-ALK [51]. This work highlights the importance of multitargeted drugs for NSCLC, but it has not focused on the emerging targets KRAS and PDGFR.

Gogoshin and Rodin (2023) conducted a review on how Graph Neural Networks (GNNs) are used in cancer, focusing on radiographic images, molecular structures, and gene expression data. The study illustrated the strength of GNNs in predicting the effectiveness of drug combinations, classifying types of cancer, and planning personalized treatments [52].

Zhang et al. (2023), Wang et al. (2023), and Chen et al. (2024) designed models based on GNNs and illustrated their efficiency in predicting how a drug will bind well to its target [53-55].

Wang et al. (2025) designed a hybrid model by combining CNNs and GATs on gene expression and molecular structure to predict drug response. GAT's mechanism helps the model to analyze complex data and achieve the highest Pearson correlation coefficient of 0.923[56].

All the above works used for predicting survival or drug-target binding, drug combination effectiveness, or drug response have used either gene expression data, small labeled datasets of binding affinities, or drug-related features. They also specified the importance of considering multitargeted drugs and the use of GNNs, especially GATs. GATs can effectively analyze complex molecular structures and extract meaningful insights. This makes them the best choice for predicting drug efficacy. However, the usage of small labelled datasets with known affinities would limit the ability to predict the drug efficacy, whereas analyzing the large unlabelled datasets using GATs can be efficient to find new patterns based on structural similarities over FDA-approved drugs.

3. Materials and Methods

3.1. Dataset

A dataset of 1.048 million drug-like molecules was downloaded from the ZINC database (https://zinc12.docking.org/subsets/clean-drug-like). This dataset includes chemically diverse molecules with drug-like properties, making it highly suitable for virtual screening and predictive modeling in drug discovery. Each compound in this dataset is available in a textual representation of chemical structures, known as SMILES (Simplified Molecular Input Line Entry System). It encodes molecular graphs into readable strings that are easily interpreted by both humans and machines.

3.2. Reference Drugs

A total of 15 FDA-approved drugs, 3 targeting EGFR (coded as E1, E2, E3), 5 targeting ALK (coded as A1, A2, A3, A4, A5), 2 targeting KRAS (coded as K1, K2), and 5 targeting PDGFR (coded as P1, P2, P3, P4, P5) were selected as reference drugs as depicted in Table 1, since they have proven their efficiency in suppressing tumor growth and improved survival rates. Each reference drug is represented in SMILES format.

Table 1 summarizes the key TKIs targeting EGFR, ALK, KRAS, and PDGFR, including their approval status and their code used in the development of a model to assess the efficacy of a new multitargeted drug.

Table 1. FDA-approved reference TKIs targeting identified key alterations to assess the efficacy of multitargeted drugs in NSCLC

S. No.	Drug / TKI Name	FDA approved in	Key Alteration	CODE
1	Afatinib	July 2013	EGFR	E1
2	Dacomitinib	September 2018		E2
3	Osimertinib	November 2015		E3
4	Ceritinib	April 2014	ALK	A1
5	Alectinib	December 2015		A2
6	Brigatinib	May 2020		A3
7	Lorlatinib	March 2021		A4
8	Ensartinib	December 2024		A5
9	Sotorasib	May 2021	KRAS	K1
10	Adagrasib	December 2022		K2
11	Avapritinib	January 2020	PDGFR	P1
12	Anlotinib	May 2018 (Not by the		P2

		U.S. FDA, but approved by the China FDA)		
13	Nintedanib	October 2014		P3
14	Crenolanib	December 2017		P4
15	Imatinib	February 2001		P5

3.3. Pharmacophore Fingerprints (PFs)

Molecular fingerprints are used to represent chemical structures for similarity searching, clustering, and predictive modelling. They encode the chemical structure of molecules/SMILES as bit vectors or numerical arrays, where each bit holds either 1 or 0. 1 indicates the presence, and 0 indicates the absence of a particular property [57]. Many types of fingerprints are available to digitally represent chemical structures for various cheminformatics applications, among which PFs have been used in this work, as they offer an efficient representation with a bit length of 39,972, focusing on identifying key functional groups and spatial arrangements responsible for a compound's biological activity, enhancing virtual screening by capturing the essential chemical features linked to drug-target interactions, and improving the identification of biologically active compounds with diverse chemical scaffolds [58, 59].

3.4. Jaccard Similarity (JS)

Similarity metrics are mathematical tools used to evaluate how closely two different entities resemble each other. In cheminformatics and drug discovery, these metrics are crucial for comparing molecular structures and identifying potential leads [60]. Commonly used metrics include Cosine Similarity, Dice Coefficient, and Jaccard Similarity. Among them, JS is particularly well-suited for binary molecular fingerprints supported by RDKit. Its computational simplicity and effectiveness in handling sparse binary data make it ideal for virtual screening and identifying structurally similar compounds.

In this framework, JS is applied using RDKit [61], which calculates similarity as the proportion of the common features (intersection) to the complete set of unique features (union) across two molecules. By focusing on the overlap of key functional features, JS supports more accurate compound comparison [62, 63].

JS is defined by the formula:

$$T(A,B) = \frac{|A \cap B|}{|A \cup B|} \qquad (1)$$

Where A and B are two binary PFs.

3.5. K-Means Clustering

Clustering is an ML algorithm that groups the data into different clusters depending on their similar structures. There are various clustering algorithms that work effectively on large datasets, such as model-based clustering, density-based clustering, K-means clustering, and hierarchical clustering [64]. We choose K-means clustering to process the dataset obtained after computing Phase I, which contains high-dimensional, unlabeled molecular data, because it provides quick and efficient computations [65, 66].

K-means clustering works as follows:

- Determine the K value, which represents the ideal number of clusters.
- Choose K centroids randomly from the dataset.
- Form k clusters by assigning each data point to the nearest centroid based on the Euclidean distance.
- Compute the new centroid for each cluster.
- Reassign the data point based on the new centroid.
- Repeat steps 4 and 5 until the centroid remains constant or predefined iterations are completed.

The efficiency of this algorithm depends on the K value.

3.6. Elbow Method - Determining K value

The elbow method is efficient to determine the number of clusters (K), with the main idea of balancing the number of clusters. In this method, K varied from 1 to 15, and WCSS (Within-Cluster Sum of Squares) is calculated for each value of K. WCSS measures the total squared distance between each data point and its corresponding cluster centroid. As the number of clusters increases, WCSS decreases, and it is largest when K = 1 (all data in a single cluster). When we plot WCSS with the K value, it will have a rapid change at one point, making the plot look like an elbow. The K-value at that point gives the exact number of clusters [67]. Once the clustering is done, it is essential to evaluate the quality of clustering and the effectiveness of each and every cluster.

3.7. Silhouette Score - Evaluating Quality of Clusters

The Silhouette Score is used to evaluate the clustering's quality by measuring how well all the data points fit in their respective cluster compared to others [68, 69]. It ranges from -1 to +1.

If the Silhouette score is close to

- +1 means that the point is fitted well to its cluster.
- -1 means that the point is not fitted to its cluster.
- 0 means that the point may fit two clusters.

3.8. Cluster Profiling – Evaluating Effectiveness of Each and Every Cluster

The uniqueness of each and every cluster can be analyzed by using cluster profiling. It includes statistics such as range, mean, quartiles, etc. Based on statistical analysis, a label that reflects its priority is assigned [70]. Cluster profiling and labelling are used in identifying the potential compounds [71].

3.9. GATs

GATs are the advanced GNN techniques that use message passing, aggregation, and attention mechanisms, which help the model to analyze complex data and improve the performance [18-21].

The following are the steps that are performed at each layer [72]:

- Linear Transformation: This step transforms the node features using a learnable weight matrix to enable the algorithm to understand complex patterns and allows the model to adjust the importance of node features during the process of training. The linear transformation of a node is computed as

$$h_i' = wh_i \quad (2)$$

Where,
h_i is the original feature vector of node i.
w is the learnable weight matrix.
h_i' is the transformed feature vector of node i.

- Compute Attention Scores: This step helps the model to understand the importance of neighbors to each node by considering both node and edge features, which is done by comparing their transformed features.

The attention score between node i and its neighbor node j is computed as

$$e_{ij} = LeakyReLU(a^T[h_i' || h_j']) \quad (3)$$

Where,
LeakyReLU is an activation function that helps the model effectively continue its learning by assigning small non-zero values to negative inputs.

a^T is the transpose of a learnable attention vector a
|| is concatenation.
h_i' and h_j' are the transformed feature vectors of the nodes i and j.

- Softmax Normalization: This step normalizes the attention scores using the softmax function, making all attention scores sum to 1, allowing the model to focus on important neighbors.

The normalized attention score between node i and node j is computed as

$$\alpha_{ij} = \frac{exp(e_{ij})}{\sum_{k\varepsilon N(i)} exp(e_{ik})} \quad (4)$$

Where,
N(i) is a Set of neighbors of node i (including itself).

- Weighted Aggregation of Node Features: This step combines the node features with the features of its neighbor using attention weights. The attention weights are different for each neighbor node, reflecting their importance.

The weighted aggregation of a node is computed as

$$h_i'' = \sum_{j\varepsilon N(i)} \alpha_{ij} h_j' \quad (5)$$

- Final Aggregation of Transformation: This step is important to further process the combined features using another learnable weight matrix.

It is computed as

$$h_i''' = w' h_i'' \quad (6)$$

Where,
w' is a learnable weight matrix for the final transformation, which will be the same for all nodes.

3.10. Evaluating GAT Model using Performance Metrics

The deep learning classification models are evaluated using different performance metrics; these metrics are computed using the confusion matrix [73, 74].The following are the metrics that are used to evaluate the performance of the GAT model effectively:

- Accuracy: It gives the proportion of correctly predicted predictions, and it is computed as

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

- Precision: It gives the proportion of correctly predicted positive predictions among all positive predictions, and it is computed as

$$Precision = \frac{TP}{TP+FP} \quad (8)$$

- Sensitivity or Recall: It gives the proportion of correctly predicted positive predictions among all actual positives, and it is computed as

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (9)$$

- F1 Score: It shows how the model balances both sensitivity and precision, and it is computed as

$$\text{F1} = 2 * \frac{\text{Precision*Recall}}{\text{Precision+Recall}} \quad (10)$$

3.11. Proposed Methodology

To provide a multitargeted drug efficacy prediction, offering a computationally efficient and faster approach by reducing the need to consider gene expression data, this paper introduces MMDEP-GAT, a Multifaceted and Multitargeted Drug Efficacy Prediction Leveraging Graph Attention Networks, that leverages the features of FDA-approved drugs targeting key alterations in NSCLC. The schematic representation of MMDEP-GAT is depicted in Figure 2.

MMDEP-GAT works in three phases:

Phase-I: SMILES Processing and Computing Similarity Metric.
Phase-II: Clustering and Labeling of Unlabeled Data
Phase-III: Predicting Efficacy of a New Drug/SMILE using GAT Model.

3.11.1. Phase-I: SMILES Processing and Computing Similarity Metric

SMILES of 1.048 million drug-like compounds and 15 reference drugs were converted into molecular structures using RDKit, an open-source cheminformatics toolkit that provides robust tools for conversion [61].

After converting the SMILES to molecular structures, PFs will be generated for further processing. This is also done by using RDKit.

Finally, JS is computed for 1.048 compounds with respect to 15 reference drugs. Figure 3 depicts the complete overview of the process of converting SMILES to molecular structures, generating PFs, and computing JS.

After the computation of Phase I, a 16-column unlabeled dataset is obtained, with the first column holding a SMILE and the remaining 15 being the JS value of a SMILE with reference to 15 drugs.

3.11.2. Phase-II: Clustering and Labeling of Unlabeled Data

The elbow method is applied on an unlabeled dataset obtained in Phase I to produce the K-value for performing K-means clustering, which helps in converting the unlabeled dataset into a labeled dataset.

After K-means clustering, the Silhouette score is applied to evaluate its quality. Finally, cluster profiling is done to evaluate the effectiveness of each cluster formed by K-means and to label them.

Figure 4 depicts the complete overview of the process of applying the elbow method, K-means clustering, evaluating the quality of clusters, and labelling clusters through profiling.

After the computation of Phase II, a 19-column labeled dataset is obtained by adding three more columns to the dataset obtained in Phase-I.

The first one is the cluster number, the second one is the label assigned to the cluster, and the third is the class value assigned to the cluster label, which is the outcome variable.

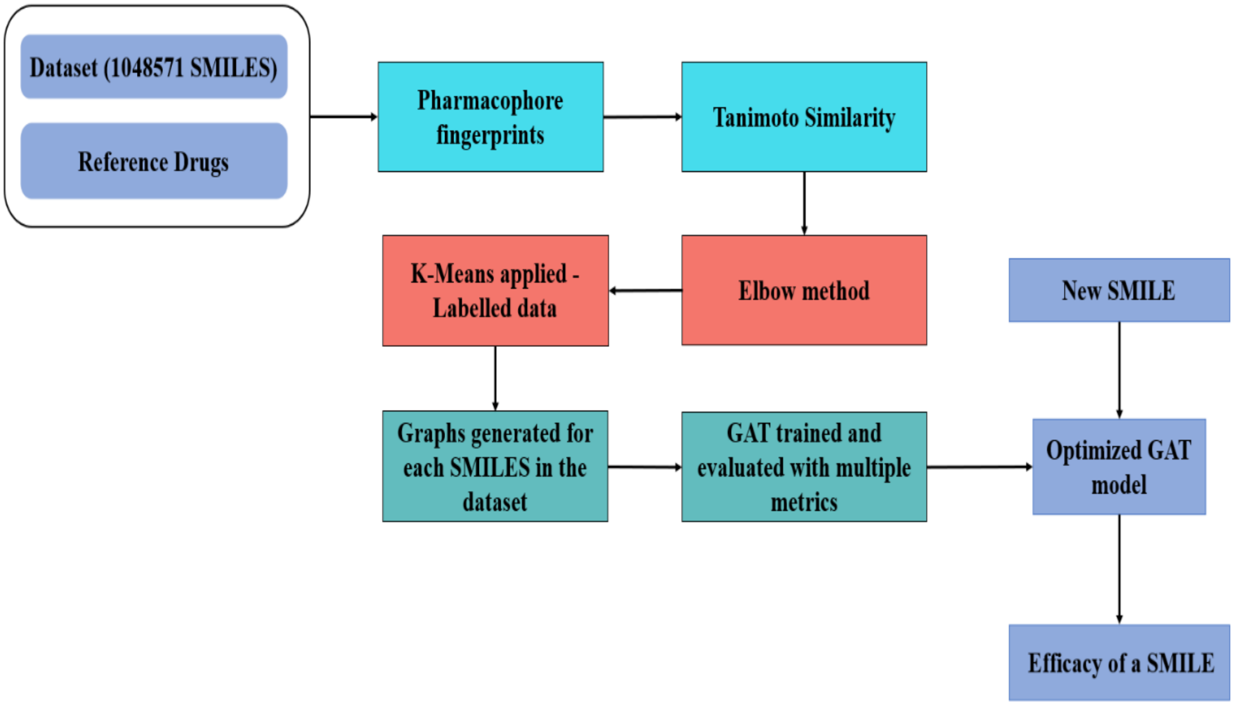
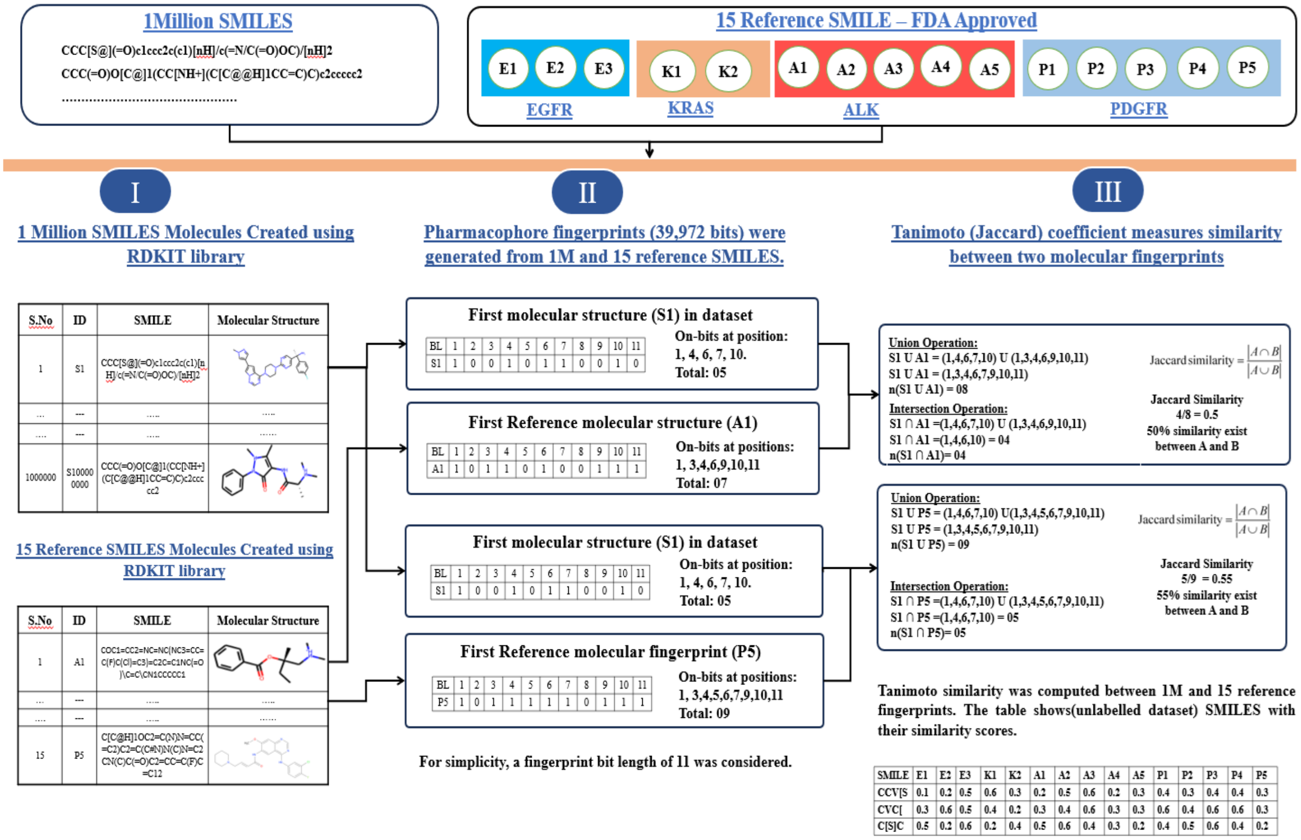

Dataset (1048571 SMILES)
Reference Drugs
Pharmacophore fingerprints
Tanimoto Similarity
Elbow method
K-Means applied - Labelled data
Graphs generated for each SMILES in the dataset
GAT trained and evaluated with multiple metrics
New SMILE
Optimized GAT model
Efficacy of a SMILE

Fig. 2 Schematic representation of the MMDEP-GAT



1Million SMILES

CCC[S@](=O)c1ccc2c(c1)[nH]/c(=N/C(=O)OC)/[nH]2

CCC(=O)O[C@]1(CC[NH+](C[C@@H]1CC=C)C)c2ccccc2

....

15 Reference SMILE – FDA Approved

1 Million SMILES Molecules Created using RDKIT library

S.No	ID	SMILE	Molecular Structure
1	S1	CCC[S@](=O)c1ccc2c(c1)[nH]/c(=N/C(=O)OC)/[nH]2	
...	---	.....	.....
....	---	.....	......
1000000	S100000000	CCC(=O)O[C@]1(CC[NH+](C[C@@H]1CC=C)C)c2ccccc2	

15 Reference SMILES Molecules Created using RDKIT library

S.No	ID	SMILE	Molecular Structure
1	A1	COC1=CC2=NC=NC(NC3=CC=C(F)C(Cl)=C3)=C2C=C1NC(=O)\C=C\CN1CCCCC1	
...	---	.....	.....
....	---	.....	......
15	P5	C[C@H]1OC2=C(N)N=CC(=C2)C2=C(C#N)N(C)N=C2CN(C)C(=O)C2=CC=C(F)C=C12	

Pharmacophore fingerprints (39,972 bits) were generated from 1M and 15 reference SMILES.

First molecular structure (S1) in dataset

BL	1	2	3	4	5	6	7	8	9	10	11
S1	1	0	0	1	0	1	1	0	0	1	0

On-bits at position:
1, 4, 6, 7, 10.
Total: 05

First Reference molecular structure (A1)

BL	1	2	3	4	5	6	7	8	9	10	11
A1	1	0	1	1	0	1	0	0	1	1	1

On-bits at positions:
1, 3,4,6,9,10,11
Total: 07

First molecular structure (S1) in dataset

BL	1	2	3	4	5	6	7	8	9	10	11
S1	1	0	0	1	0	1	1	0	0	1	0

On-bits at position:
1, 4, 6, 7, 10.
Total: 05

First Reference molecular fingerprint (P5)

BL	1	2	3	4	5	6	7	8	9	10	11
P5	1	0	1	1	1	1	1	0	1	1	1

On-bits at positions:
1, 3,4,5,6,7,9,10,11
Total: 09

For simplicity, a fingerprint bit length of 11 was considered.

Tanimoto (Jaccard) coefficient measures similarity between two molecular fingerprints

Union Operation:
S1 U A1 = (1,4,6,7,10) U (1,3,4,6,9,10,11)
S1 U A1 = (1,3,4,6,7,9,10,11)
n(S1 U A1) = 08

Intersection Operation:
S1 ∩ A1 =(1,4,6,7,10) U (1,3,4,6,9,10,11)
S1 ∩ A1 =(1,4,6,10) = 04
n(S1 ∩ A1)= 04

$$\text{Jaccard similarity} = \frac{|A \cap B|}{|A \cup B|}$$

Jaccard Similarity
4/8 = 0.5
50% similarity exist between A and B

Union Operation:
S1 U P5 = (1,4,6,7,10) U(1,3,4,5,6,7,9,10,11)
S1 U P5 = (1,3,4,5,6,7,9,10,11)
n(S1 U P5) = 09

Intersection Operation:
S1 ∩ P5 =(1,4,6,7,10) U (1,3,4,5,6,7,9,10,11)
S1 ∩ P5 =(1,4,6,7,10) = 05
n(S1 ∩ P5)= 05

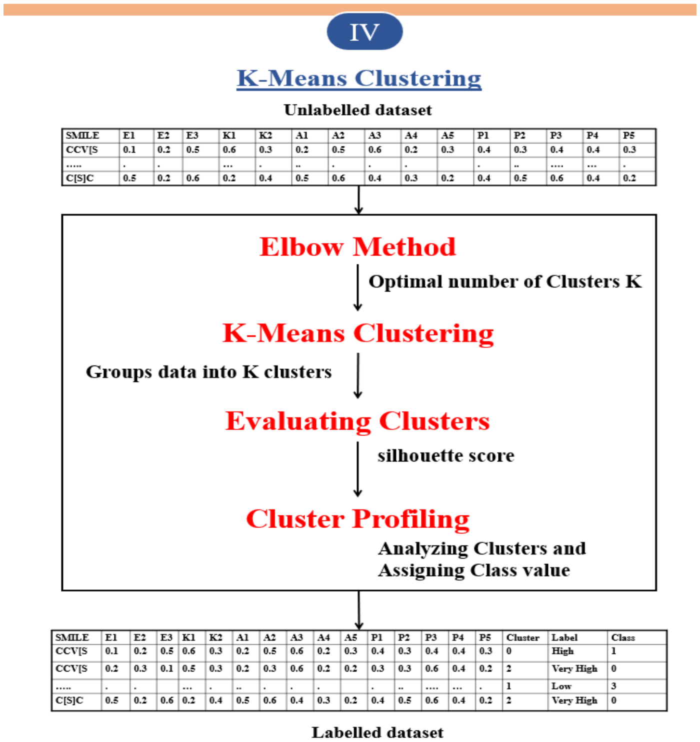
$$\text{Jaccard similarity} = \frac{|A \cap B|}{|A \cup B|}$$

Jaccard Similarity
5/9 = 0.55
55% similarity exist between A and B

Tanimoto similarity was computed between 1M and 15 reference fingerprints. The table shows(unlabelled dataset) SMILES with their similarity scores.

SMILE	E1	E2	E3	K1	K2	A1	A2	A3	A4	A5	P1	P2	P3	P4	P5
CCV[S	0.1	0.2	0.5	0.6	0.3	0.2	0.5	0.6	0.2	0.3	0.4	0.3	0.4	0.4	0.3
CVC[	0.3	0.6	0.5	0.4	0.2	0.3	0.4	0.6	0.3	0.3	0.6	0.4	0.6	0.6	0.3
C[S]C	0.5	0.2	0.6	0.2	0.4	0.5	0.6	0.4	0.3	0.2	0.4	0.5	0.6	0.4	0.2

Fig. 3 Overview of phases-I in MMDEP-GAT - the process of converting SMILES to PFs and then generating a dataset with the values of JS



K-Means Clustering

Unlabelled dataset

SMILE	E1	E2	E3	K1	K2	A1	A2	A3	A4	A5	P1	P2	P3	P4	P5
CCV[S	0.1	0.2	0.5	0.6	0.3	0.2	0.5	0.6	0.2	0.3	0.4	0.3	0.4	0.4	0.3
.....	.	.		...	.	..	.	.	.		.	..	....	...	.
C[S]C	0.5	0.2	0.6	0.2	0.4	0.5	0.6	0.4	0.3	0.2	0.4	0.5	0.6	0.4	0.2

Elbow Method

Optimal number of Clusters K

K-Means Clustering

Groups data into K clusters

Evaluating Clusters

silhouette score

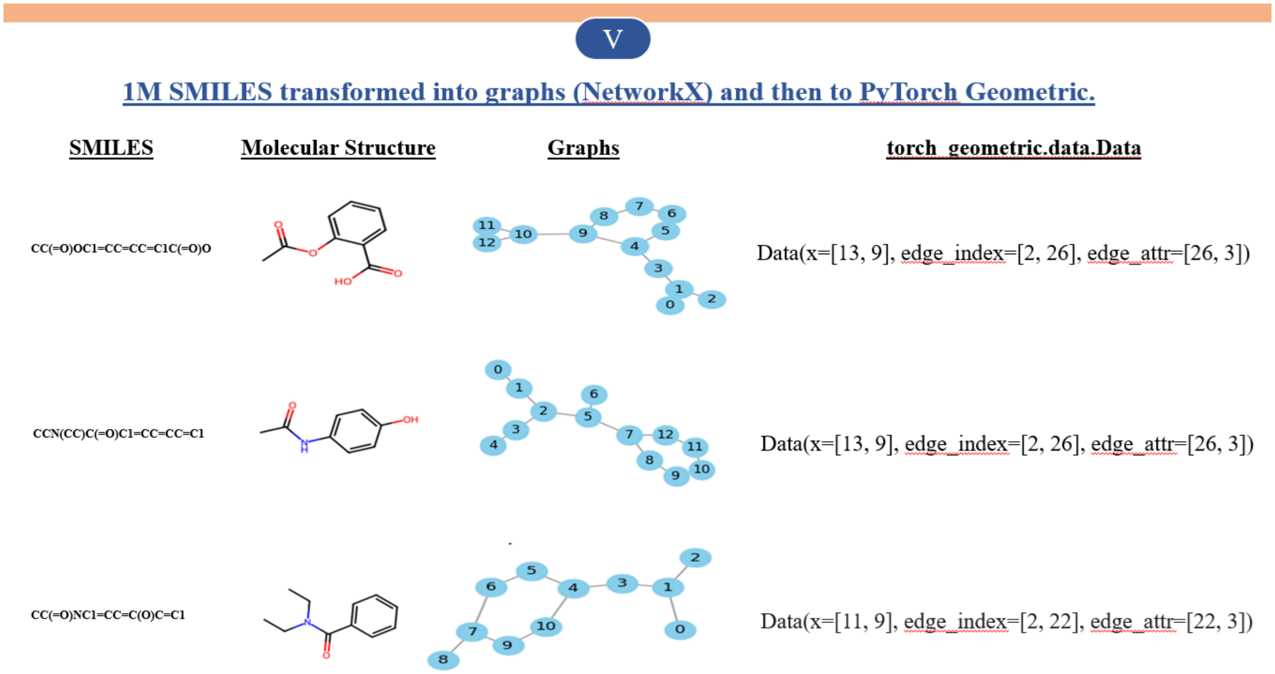
Cluster Profiling

Analyzing Clusters and Assigning Class value

SMILE	E1	E2	E3	K1	K2	A1	A2	A3	A4	A5	P1	P2	P3	P4	P5	Cluster	Label	Class
CCV[S	0.1	0.2	0.5	0.6	0.3	0.2	0.5	0.6	0.2	0.3	0.4	0.3	0.4	0.4	0.3	0	High	1
CCV[S	0.2	0.3	0.1	0.5	0.3	0.2	0.3	0.6	0.2	0.2	0.3	0.3	0.6	0.4	0.2	2	Very High	0
.....	.	.		...	.	..	.	.	.		.	..	....	...	.	1	Low	3
C[S]C	0.5	0.2	0.6	0.2	0.4	0.5	0.6	0.4	0.3	0.2	0.4	0.5	0.6	0.4	0.2	2	Very High	0

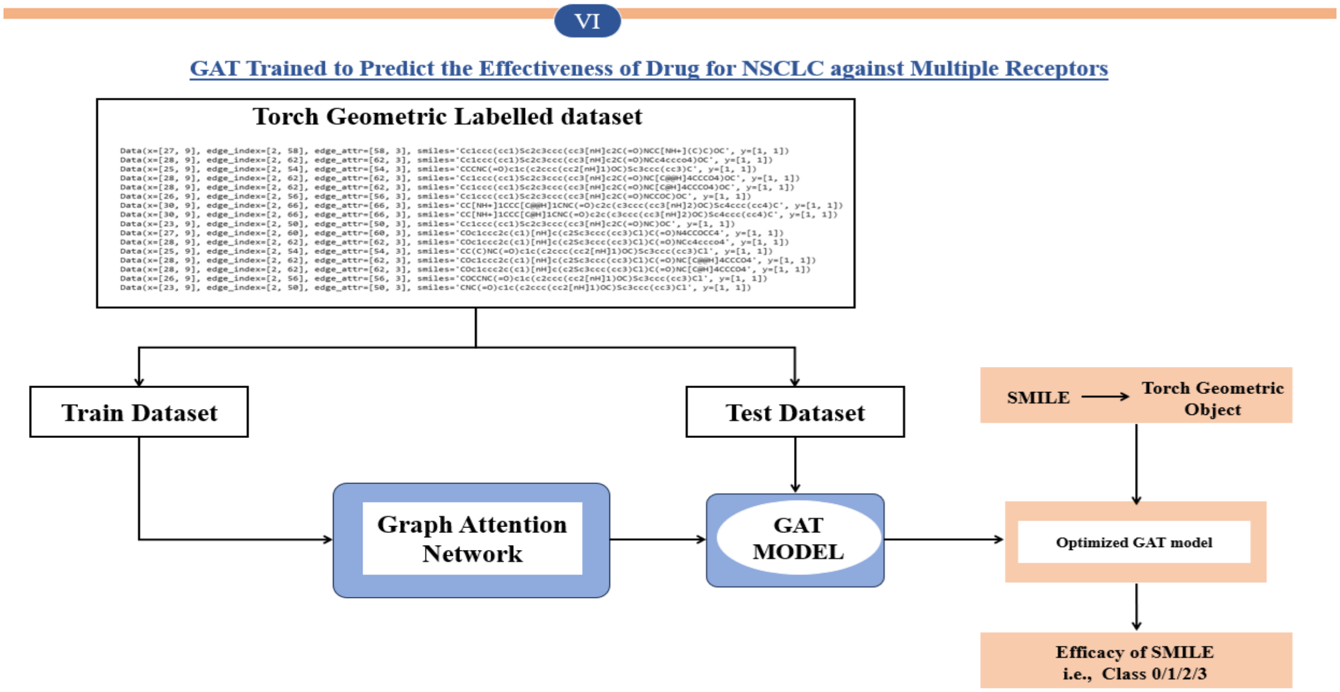
Labelled dataset

Fig. 4 Overview of the phase II in MMDEP-GAT – the process of converting an unlabeled dataset obtained in phase I to a labelled dataset



```
Data(x=[13, 9], edge_index=[2, 26], edge_attr=[26, 3])
Data(x=[13, 9], edge_index=[2, 26], edge_attr=[26, 3])
Data(x=[11, 9], edge_index=[2, 22], edge_attr=[22, 3])
```

Fig. 5 Generating PyTorch geometric objects from SMILES


VI
GAT Trained to Predict the Effectiveness of Drug for NSCLC against Multiple Receptors
Torch Geometric Labelled dataset
Train Dataset
Test Dataset
Graph Attention Network
GAT MODEL
SMILE
Torch Geometric Object
Optimized GAT model
Efficacy of SMILE i.e., Class 0/1/2/3

```
Data(x=[27, 9], edge_index=[2, 58], edge_attr=[58, 3], smiles='Cc1ccc(cc1)Sc2c3ccc(cc3[nH]c2C(=O)NCC[NH+](C)C)OC', y=[1, 1])
Data(x=[28, 9], edge_index=[2, 62], edge_attr=[62, 3], smiles='Cc1ccc(cc1)Sc2c3ccc(cc3[nH]c2C(=O)NCc4ccco4)OC', y=[1, 1])
Data(x=[25, 9], edge_index=[2, 54], edge_attr=[54, 3], smiles='CCCNC(=O)c1c(c2ccc(cc2[nH]1)OC)Sc3ccc(cc3)C', y=[1, 1])
Data(x=[28, 9], edge_index=[2, 62], edge_attr=[62, 3], smiles='Cc1ccc(cc1)Sc2c3ccc(cc3[nH]c2C(=O)NC[C@@H]4CCCO4)OC', y=[1, 1])
Data(x=[28, 9], edge_index=[2, 62], edge_attr=[62, 3], smiles='Cc1ccc(cc1)Sc2c3ccc(cc3[nH]c2C(=O)NC[C@H]4CCCO4)OC', y=[1, 1])
Data(x=[26, 9], edge_index=[2, 56], edge_attr=[56, 3], smiles='Cc1ccc(cc1)Sc2c3ccc(cc3[nH]c2C(=O)NCCOC)OC', y=[1, 1])
Data(x=[30, 9], edge_index=[2, 66], edge_attr=[66, 3], smiles='CC[NH+]1CCC[C@@H]1CNC(=O)c2c(c3ccc(cc3[nH]2)OC)Sc4ccc(cc4)C', y=[1, 1])
Data(x=[30, 9], edge_index=[2, 66], edge_attr=[66, 3], smiles='CC[NH+]1CCC[C@H]1CNC(=O)c2c(c3ccc(cc3[nH]2)OC)Sc4ccc(cc4)C', y=[1, 1])
Data(x=[23, 9], edge_index=[2, 50], edge_attr=[50, 3], smiles='Cc1ccc(cc1)Sc2c3ccc(cc3[nH]c2C(=O)NC)OC', y=[1, 1])
Data(x=[27, 9], edge_index=[2, 60], edge_attr=[60, 3], smiles='COc1ccc2c(c1)[nH]c(c2Sc3ccc(cc3)Cl)C(=O)N4CCOCC4', y=[1, 1])
Data(x=[28, 9], edge_index=[2, 62], edge_attr=[62, 3], smiles='COc1ccc2c(c1)[nH]c(c2Sc3ccc(cc3)Cl)C(=O)NCc4ccco4', y=[1, 1])
Data(x=[25, 9], edge_index=[2, 54], edge_attr=[54, 3], smiles='CC(C)NC(=O)c1c(c2ccc(cc2[nH]1)OC)Sc3ccc(cc3)Cl', y=[1, 1])
Data(x=[28, 9], edge_index=[2, 62], edge_attr=[62, 3], smiles='COc1ccc2c(c1)[nH]c(c2Sc3ccc(cc3)Cl)C(=O)NC[C@@H]4CCCO4', y=[1, 1])
Data(x=[28, 9], edge_index=[2, 62], edge_attr=[62, 3], smiles='COc1ccc2c(c1)[nH]c(c2Sc3ccc(cc3)Cl)C(=O)NC[C@H]4CCCO4', y=[1, 1])
Data(x=[26, 9], edge_index=[2, 56], edge_attr=[56, 3], smiles='COCCNC(=O)c1c(c2ccc(cc2[nH]1)OC)Sc3ccc(cc3)Cl', y=[1, 1])
Data(x=[23, 9], edge_index=[2, 50], edge_attr=[50, 3], smiles='CNC(=O)c1c(c2ccc(cc2[nH]1)OC)Sc3ccc(cc3)Cl', y=[1, 1])
```

Fig. 6 Overview of the phase III in MMDEP-GAT – dividing labeled dataset and the process of obtaining GAT model

3.11.3. Phase-III: Clustering and Labeling of Unlabeled Data

To build a GAT model for the evaluation of drug efficacy, it is necessary to transform the data so that it can be easily processed using GATs. For this, all the SMILES of 1.048 million drug-like compounds, along with the 15 reference drugs, are transformed into graphs and then to geometric objects using PyTorch, as depicted in Figure 5. After transforming to geometric objects, the torch geometric labelled dataset is divided into a train dataset for training the GATs and a test dataset for testing the model using a fixed random seed of 42. Finally, an optimized GAT model is obtained for predicting the efficacy of a new SMILE. The MMDEP-GAT framework used an attention-based message-passing architecture using AttentiveFP, a GAT variant. Each graph encodes 9 node features and 3 edge features. 6 message-passing layers with 2 attention time-steps were used to iteratively refine node embeddings. Each layer uses a LeakyReLU activation followed by 0.2 dropout, with a hidden dimension of 70. The output layer produces four classes, representing the predicted efficacy of SMILES across EGFR, ALK, KRAS, and PDGFR targets. Training and testing were done for 25 epochs with a batch size of 32, using the Adam optimizer with a learning rate of 0.001 and a weight decay of 1×10^{-4}, with the cross-entropy loss function for multi-class classification. Finally, an optimized GAT model is obtained for predicting the efficacy of a new SMILE. Figure 6 depicts the overview of the dividing torch geometric labelled dataset, and the process of training and testing the GAT model.

After the computation of Phase III, the optimized GAT model is obtained to predict the efficacy of a new SMILE.

4. Results and Discussion

4.1. Phase-I

JS was computed between 1.048 million drugs and each of the 15 reference drugs. The summary statistics of JS w.r.t. the EGFR-targeting reference drugs coded as E1, E2, and E3 are represented in Table 2 and also depicted in Figure 7.

Table 2. Summary statistics of JS between 1.048 million drugs and EGFR reference SMILEs (E1, E2, and E3)

	Min	25% (Q1)	50% (Median)	75% (Q3)	Max	Mean	Std Dev
E1_T	0.000	0.044	0.078	0.118	0.351	0.084	0.051
E2_T	0.000	0.048	0.086	0.129	0.442	0.093	0.056
E3_T	**0.000**	**0.050**	**0.098**	**0.157**	**0.463**	**0.110**	**0.074**

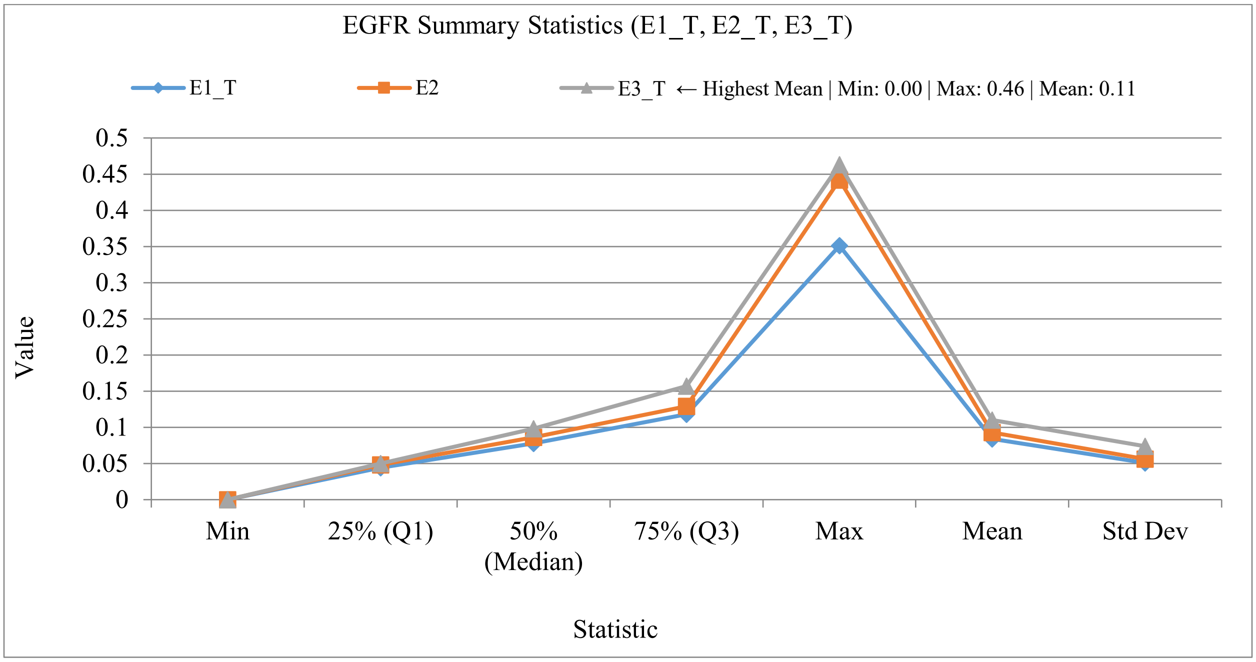


Fig. 7 Line chart representing summary statistics of JS between E1, E2, E3, and 1.048 million drugs

From Figure 7, it is observed that E3 (Osimertinib) shows the highest maximum (0.46) and average (0.11) similarity among the three drugs. This indicates that E3 shares stronger structural features with a major portion of the dataset.

The summary statistics of JS w.r.t. the ALK-targeting reference drugs coded as A1, A2, A3, A4, and A5 are represented in Table 3 and also depicted in Figure. 8.

From Figure 8, it is observed that A5 (Ensartinib) shows the highest maximum and average similarity among the five drugs. This indicates that A5 shares stronger structural features with a major portion of the dataset. The summary statistics of JS w.r.t. the KRAS-targeting reference drugs coded as K1 and K2 are represented in Table 4 and also depicted in Figure 9. From Figure 9, it is observed that both K1 and K2 have shown the same maximum and average similarity. This indicates that both K1 and K2 share stronger structural features with a major portion of the dataset.

The summary statistics of JS w.r.t. the PDGFR-targeting reference drugs coded as P1, P2, P3, P4, and P5 are represented in Table 5 and also depicted in Figure 10. From Figure 10, it is observed that both P2 and P3 have shown the nearest maximum (0.39 and 0.34, respectively) and the same average (0.10) similarity. This indicates that both P2 and P3 share stronger structural features with a major portion of the dataset.

Table 3. Summary statistics of JS between 1.048 million drugs and ALK reference SMILEs (A1, A2, A3, A4, and A5)

	Min	25% (Q1)	50% (Median)	75% (Q3)	Max	Mean	Std Dev
A1_T	0.000	0.041	0.073	0.112	0.354	0.080	0.049
A2_T	0.000	0.045	0.075	0.107	0.313	0.078	0.043
A3_T	0.000	0.030	0.054	0.083	0.302	0.060	0.038
A4_T	0.000	0.045	0.081	0.124	0.347	0.088	0.055
A5_T	**0.000**	**0.046**	**0.083**	**0.127**	**0.388**	**0.091**	**0.057**

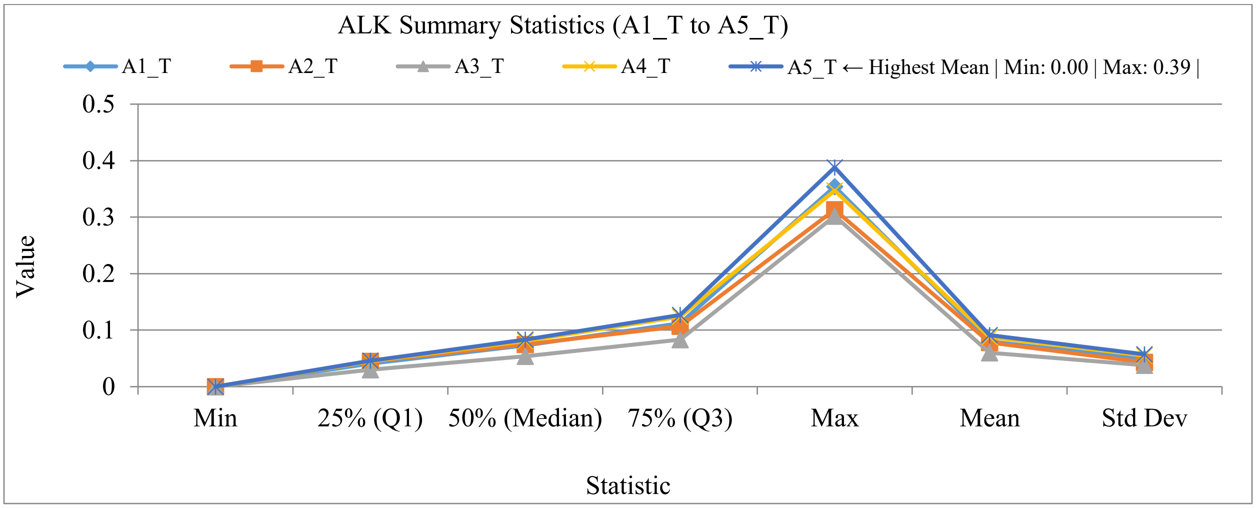


Fig. 8 Line chart representing summary statistics of JS between A1, A2, A3, A4, A5, and 1.048 million drugs

Table 4. Summary statistics of JS between 1.048 million drugs and KRAS reference SMILEs (K1 and K2)

	Min	25% (Q1)	50% (Median)	75% (Q3)	Max	Mean	Std Dev
K1_T	0.000	0.041	0.072	0.109	0.393	0.078	0.048
K2_T	**0.000**	**0.041**	**0.072**	**0.109**	**0.393**	**0.078**	**0.048**

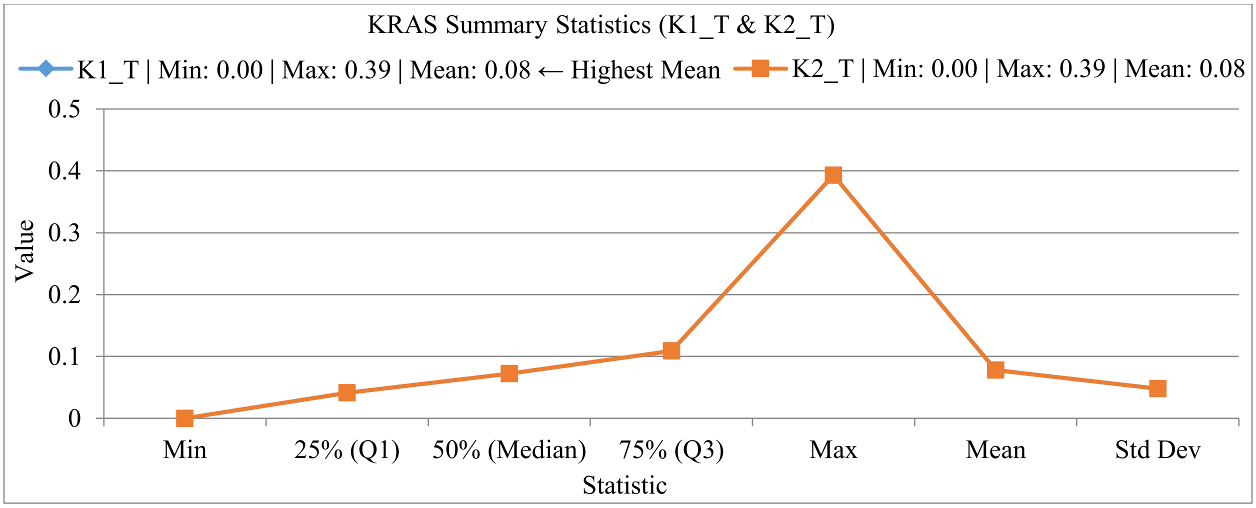


Fig. 9 Line chart representing summary statistics of JS between K1, K2, and 1.048 million drugs

Table 5. Summary statistics of JS between 1.048 million drugs and PDGFR reference SMILEs (P1, P2, P3, P4, and P5)

	Min	25% (Q1)	50% (Median)	75% (Q3)	Max	Mean	Std Dev
P1_T	0.00	0.04	0.07	0.11	0.39	0.08	0.05
P2_T	**0.00**	**0.05**	**0.09**	**0.14**	**0.39**	**0.10**	**0.06**
P3_T	**0.00**	**0.05**	**0.09**	**0.13**	**0.34**	**0.10**	**0.05**
P4_T	0.00	0.04	0.06	0.09	0.81	0.07	0.04
P5_T	0.00	0.04	0.08	0.12	0.42	0.09	0.05

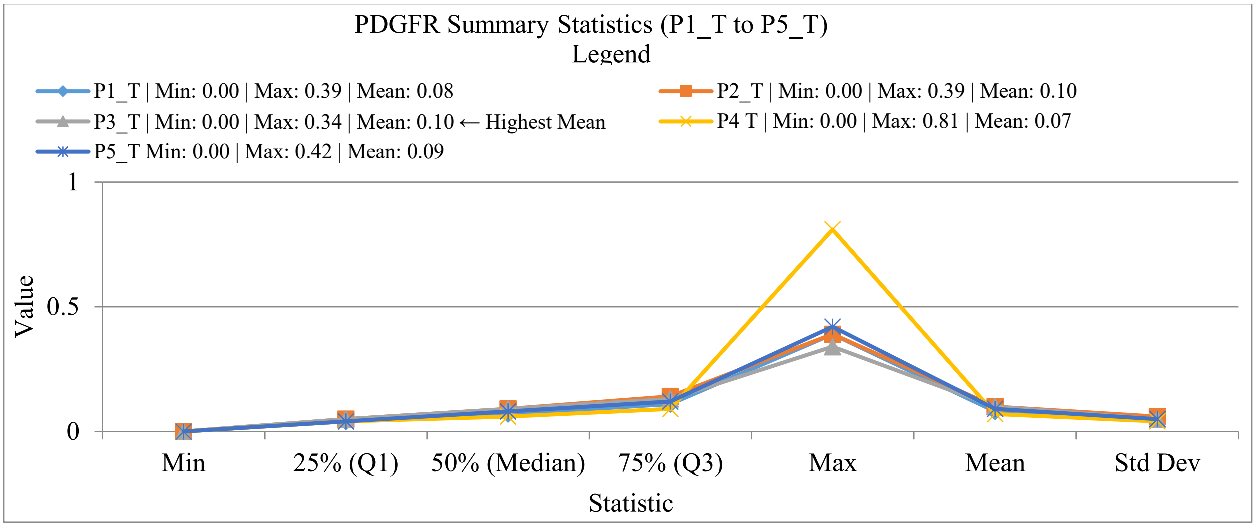
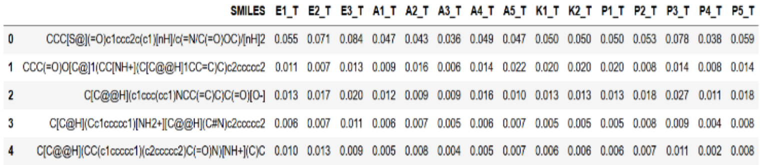


Fig. 10 Line chart representing summary statistics of JS between P1, P2, P3, P4, P5, and 1.048 million drugs



	SMILES	E1_T	E2_T	E3_T	A1_T	A2_T	A3_T	A4_T	A5_T	K1_T	K2_T	P1_T	P2_T	P3_T	P4_T	P5_T
0	CCC[S@](=O)c1ccc2c(c1)[nH]/c(=N/C(=O)OC)/[nH]2	0.055	0.071	0.084	0.047	0.043	0.036	0.049	0.047	0.050	0.050	0.050	0.053	0.078	0.038	0.059
1	CCC(=O)O[C@]1(CC[NH+](C[C@@H]1CC=C)C)c2ccccc2	0.011	0.007	0.013	0.009	0.016	0.006	0.014	0.022	0.020	0.020	0.020	0.008	0.014	0.008	0.014
2	C[C@@H](c1ccc(cc1)NCC(=C)C)C(=O)[O-]	0.013	0.017	0.020	0.012	0.009	0.009	0.016	0.010	0.013	0.013	0.013	0.018	0.027	0.011	0.018
3	C[C@H](Cc1ccccc1)[NH2+][C@@H](C#N)c2ccccc2	0.006	0.007	0.011	0.006	0.007	0.005	0.006	0.007	0.005	0.005	0.005	0.008	0.009	0.004	0.008
4	C[C@@H](CC(c1ccccc1)(c2ccccc2)C(=O)N)[NH+](C)C	0.010	0.013	0.009	0.005	0.008	0.004	0.005	0.007	0.006	0.006	0.006	0.007	0.011	0.002	0.008

Fig. 11 Overview of the dataset obtained after phase I

The computation of JS between 1.048 million drugs and each of the 15 reference drugs generates an unlabeled dataset as depicted in Figure 11.

4.2. Phase-II

The elbow method is applied to an unlabeled dataset obtained in Phase I, and Figure 12 represents the graph of the elbow method.

From Figure 12, it is observed that there is a significant difference in WCSS among the first 4 clusters. But from the 5th cluster onwards, the change has become minimal. Hence, the optimal number of clusters is considered to be 4.

Now, by using K values of 4, clustering is done. To visualize these clusters clearly and effectively, Principal Component Analysis (PCA) is used, which transforms high-dimensional data into a two-dimensional space while preserving as much variance as possible. The visualization of clusters using PCA is depicted in Figure 13.The total number of SMILE entries in each cluster is represented in Table 6 and depicted in Figure 14.

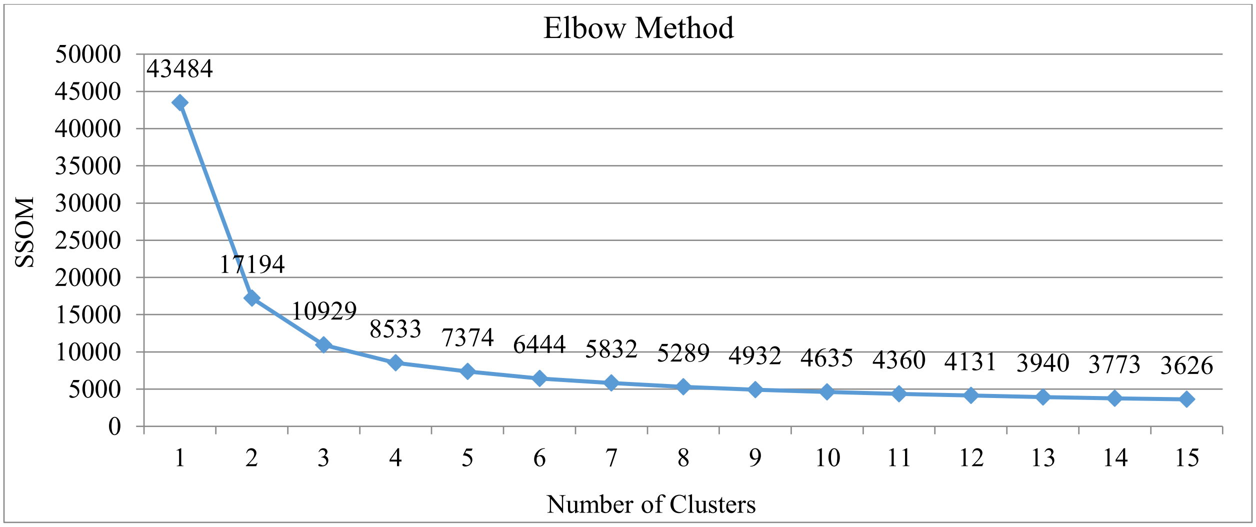


Fig. 12 Elbow method graph - determining the number of clusters (K)

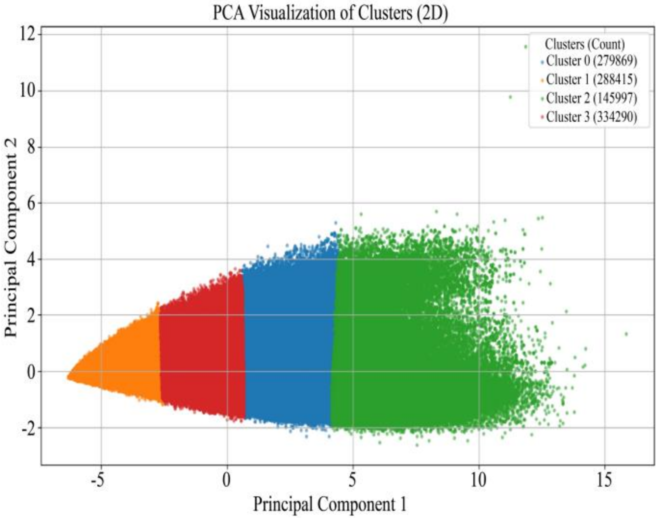


Fig. 13 PCA visualization of clusters

Table 6. Cluster-wise count of SMILES

S. No.	Cluster Number	SMILE Count
1	Cluster 0	279869
2	Cluster 1	288415
3	Cluster 2	145997
4	Cluster 3	334290

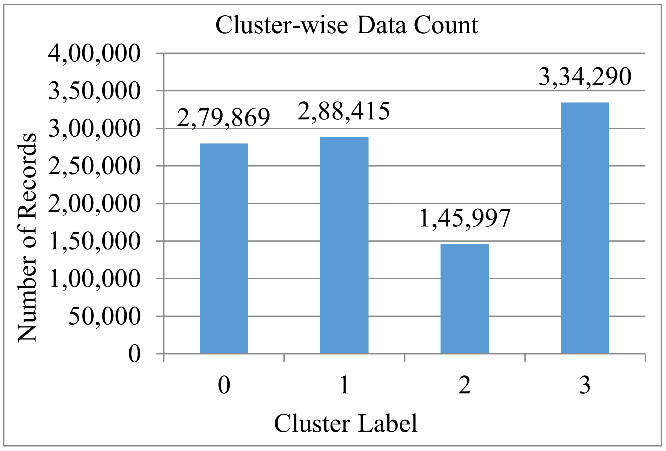
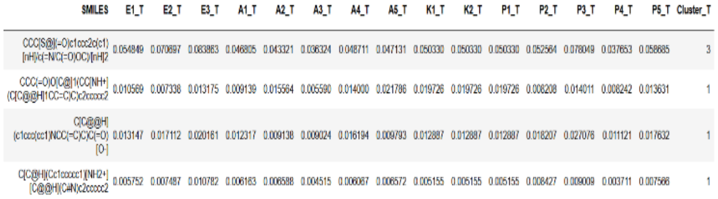


Fig. 14 Bar graph representing SMILE count of each and every cluster

From Figure 14, it is observed that cluster number 3 has the highest number of SMILES entries among all the clusters. After clustering, an unlabeled dataset is converted to a labeled dataset with an added column for specifying the cluster to which SMILE belongs. Now, the labeled dataset has 17 columns, with 1 column as a SMILE string, 15 columns as JS values, and 1 column indicating cluster number, as depicted in Figure 15.



SMILES	E1_T	E2_T	E3_T	A1_T	A2_T	A3_T	A4_T	A5_T	K1_T	K2_T	P1_T	P2_T	P3_T	P4_T	P5_T	Cluster_T
CCC[S@](=O)c1ccc2c(c1)[nH]/c(=N/C(=O)OC)/[nH]2	0.054849	0.070897	0.083863	0.046805	0.043321	0.036324	0.048711	0.047131	0.050330	0.050330	0.050330	0.052564	0.078049	0.037653	0.058685	3
CCC(=O)O[C@]1(CC[NH+](C[C@@H]1CC=C)C)c2ccccc2	0.010569	0.007338	0.013175	0.009139	0.015564	0.005590	0.014000	0.021786	0.019726	0.019726	0.019726	0.008208	0.014011	0.008242	0.013631	1
C[C@@H](c1ccc(cc1)NCC(=C)C)C(=O)[O-]	0.013147	0.017112	0.020161	0.012317	0.009138	0.009024	0.016194	0.009793	0.012887	0.012887	0.012887	0.018207	0.027076	0.011121	0.017632	1
C[C@H](Cc1ccccc1)[NH2+][C@@H](C#N)c2ccccc2	0.005752	0.007487	0.010782	0.006163	0.006588	0.004515	0.006067	0.006572	0.005155	0.005155	0.005155	0.008427	0.009009	0.003711	0.007566	1

Fig. 15 Overview of the labeled dataset with 17 columns obtained after clustering

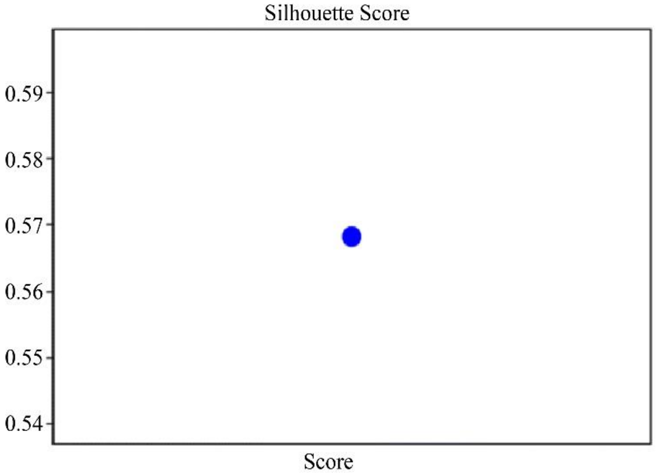


Fig. 16 Silhouette score representation for K-means clustering

Now, Silhouette Score is applied to evaluate the quality of clustering done by K-means. It is computed as a score of 0.56, as depicted in Figure 16, which specifies that the clustering done by K-means is efficient.

Now, to evaluate the effectiveness of every cluster, two metrics have been considered from clustering profiling

1. min-max analysis, i.e., analyzing the minimum and maximum values of 15 JS values in every cluster.
2. Mean analysis, i.e., analyzing the cluster-wise mean.

Cluster-wise min-max analyses are depicted in Figures 17(a) to 17(d).

Figure 17(a) depicts the min-max analysis of 15 JS in cluster 0 that contains 279869 SMILES, from which it is observed that in cluster 0, the JS's maximum value ranges from 0.21 to 0.32, its minimum from 0.03 to 0.06, and its average value is 0.11.

Figure 17(b) depicts the min-max analysis of 15 JS in cluster 1 that contains 288415 SMILES, from which it is observed that in cluster 1, the JS's maximum value ranges from 0.08 to 0.11, its minimum from 0.00 to 0.00, and its average value is 0.03. Figure 17(c) depicts the min-max analysis of 15 JS in cluster 2 that contains 145997 SMILES, from which it is observed that in cluster 2, the JS's maximum value ranges from 0.30 to 0.81, its minimum from 0.06 to 0.10, and its average value is 0.17. Figure 17(d) depicts the min-max analysis of 15 JS in cluster 3, which consists of 334290 SMILES with JS's maximum and minimum values ranging from 0.14 to 0.24 and 0.01 to 0.03, with an average of 0.07.

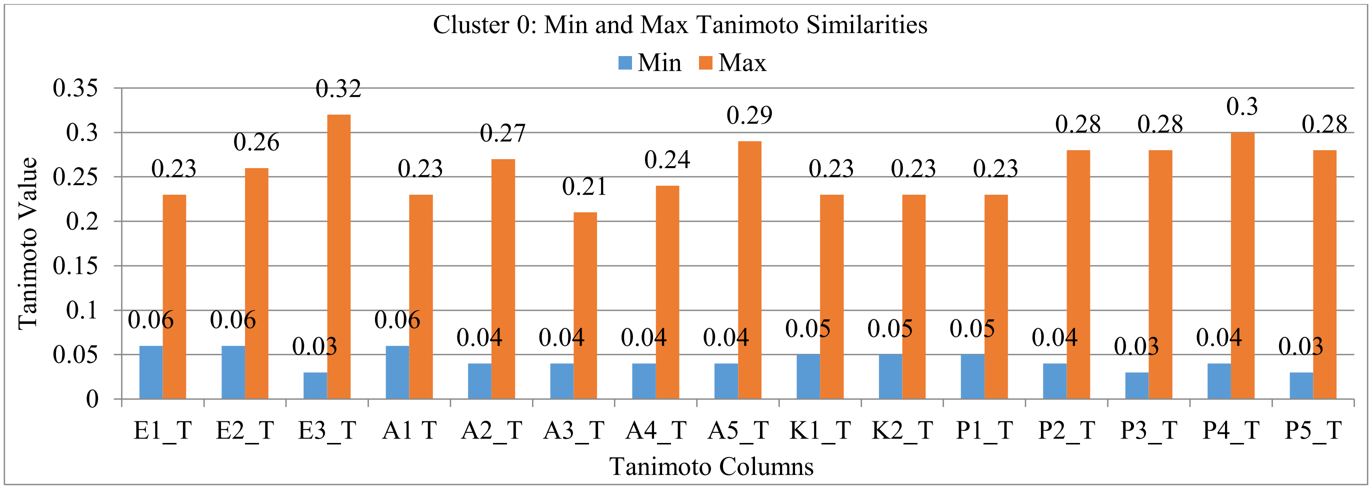


Fig. 17(a) Min-max analysis of 15 JS values in cluster 0

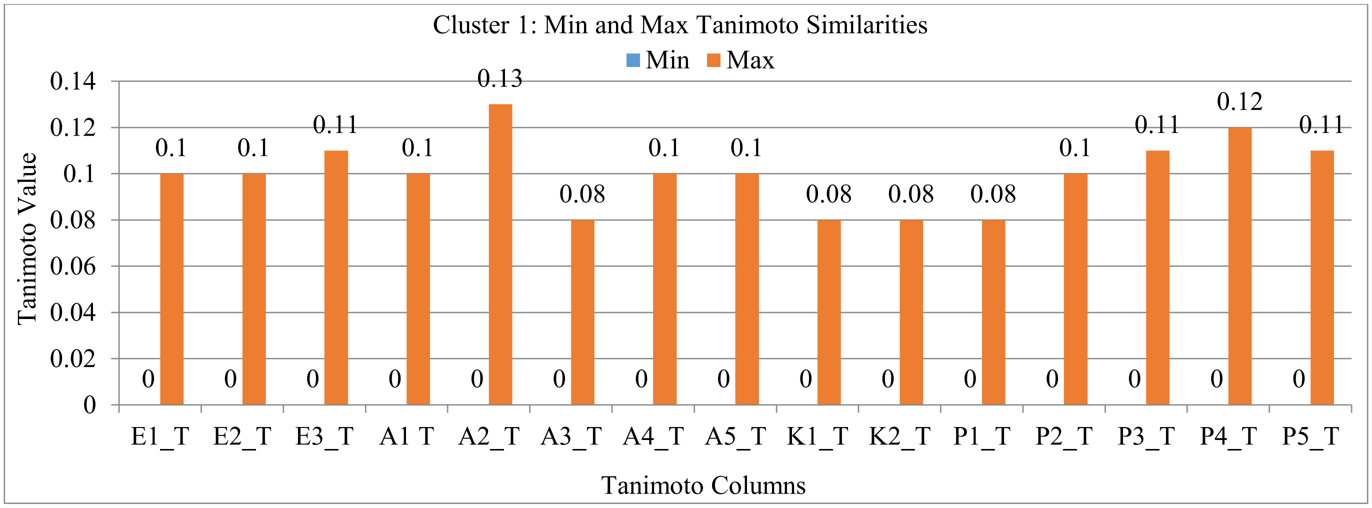


Fig. 17(b) Min-max analysis of 15 JS values in cluster 1

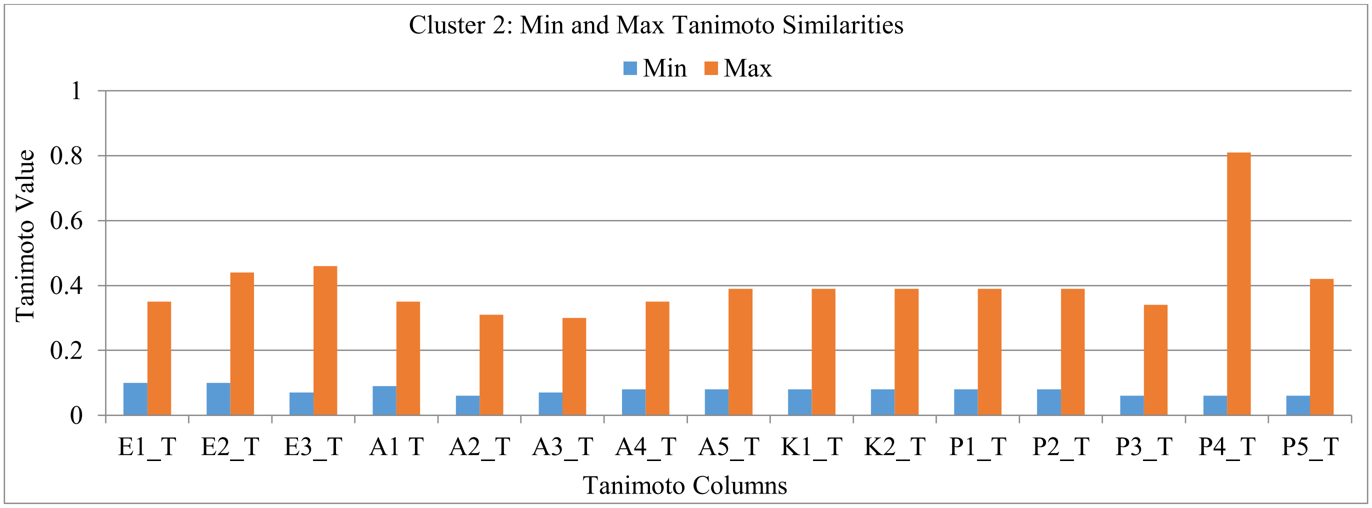


Fig. 17(c) Min-max analysis of 15 JS values in cluster 2

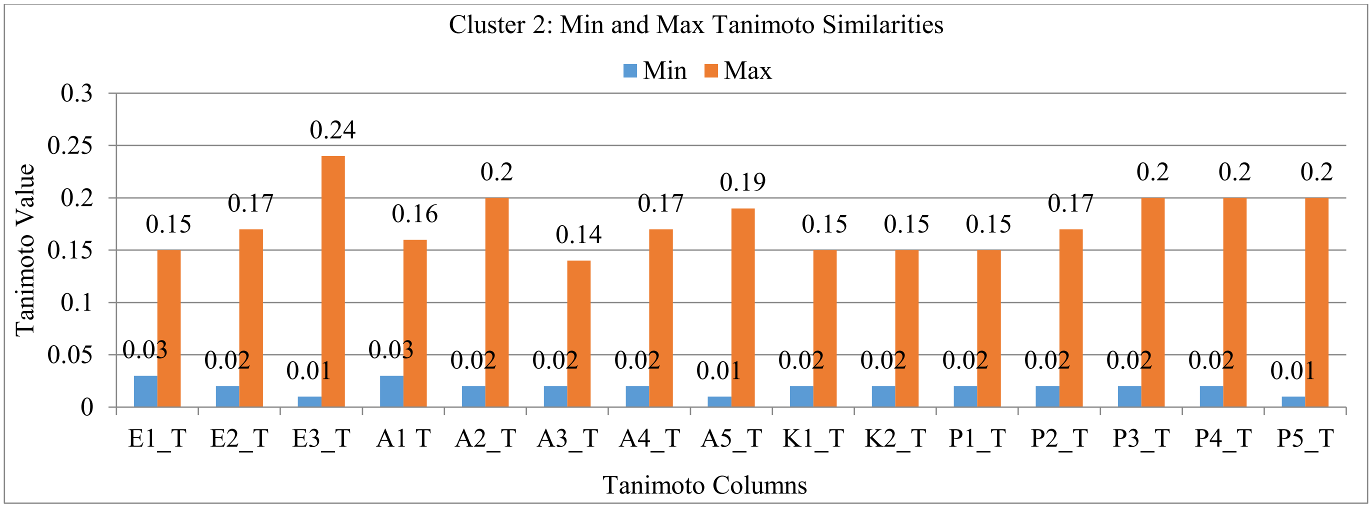


Fig. 17(d) Min-max analysis of 15 JS values in cluster 3

Table 7. Cluster-wise min-max and average JS analysis

S. No.	Cluster Number	SMILE Count	Range of Maximum JS	Range of Minimum JS	Average JS
1	Cluster 0	279869	0.21 to 0.32	0.03 to 0.06	0.11
2	Cluster 1	288415	0.08 to 0.11	0.00 to 0.00	0.03
3	**Cluster 2**	**145997**	**0.30 to 0.81**	**0.06 to 0.10**	**0.17**
4	Cluster 3	334290	0.14 to 0.24	0.01 to 0.03	0.07

Min-max and average analyses of JS in each cluster, along with the SMILES count, are represented in Table 7. From the analysis of JS represented in Table 7, the descending order of clusters with respect to average JS is cluster 2, cluster 0, cluster 3, and cluster 1. Cluster 2 shows the highest average JS of 0.17, indicating strong structural similarity; therefore, it can be labeled as "very high."

Next, Cluster 0 has the next highest average JS of 0.11; therefore, it can be labelled as "high." In a similar way, Cluster 3 with a moderate average JS of 0.07 is labeled as "moderate," while Cluster 1 with the lowest average JS of 0.03, indicating weak similarity, is labeled as "low." Figure 18 depicts the bar graph indicatig the cluster-wise count, mean values, and labeling of each cluster.

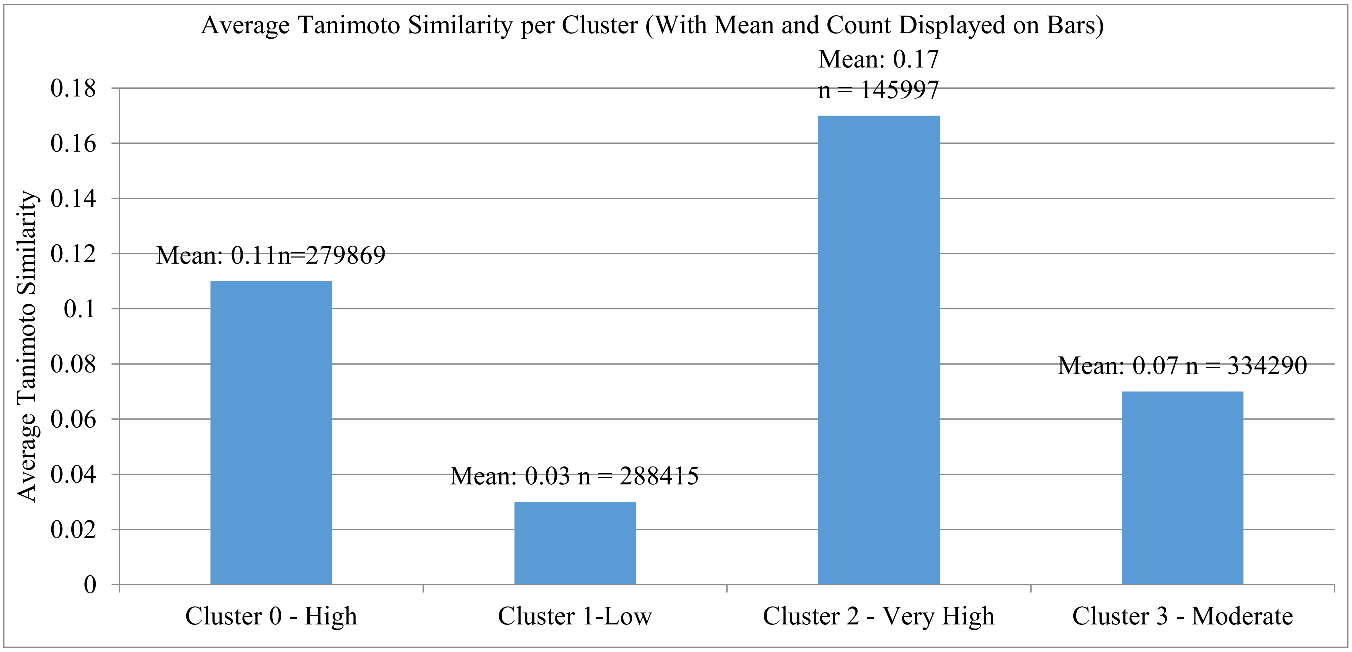
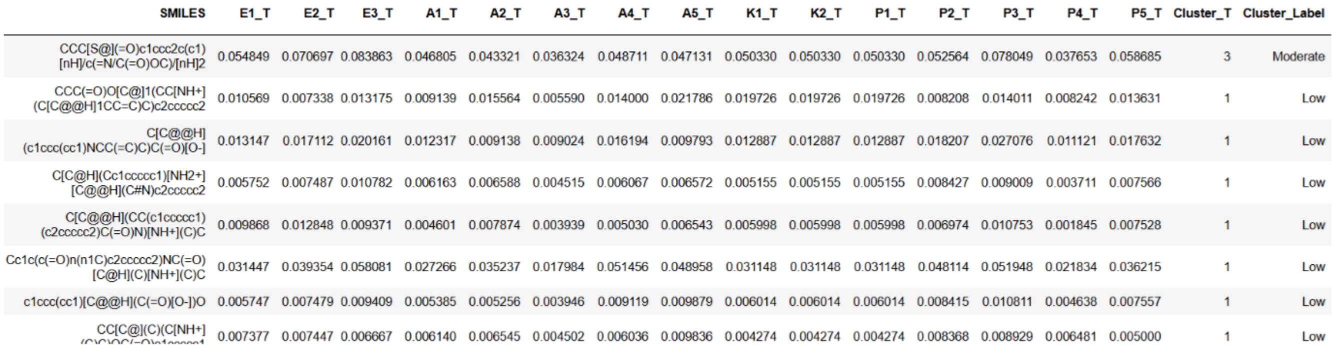


Fig. 18 Bar graph representing cluster profiling with mean and count for each labeled cluster



SMILES	E1_T	E2_T	E3_T	A1_T	A2_T	A3_T	A4_T	A5_T	K1_T	K2_T	P1_T	P2_T	P3_T	P4_T	P5_T	Cluster_T	Cluster_Label
CCC[S@](=O)c1ccc2c(c1)[nH]/c(=N/C(=O)OC)/[nH]2	0.054849	0.070697	0.083863	0.046805	0.043321	0.036324	0.048711	0.047131	0.050330	0.050330	0.050330	0.052564	0.078049	0.037653	0.058685	3	Moderate
CCC(=O)O[C@]1(CC[NH+](C[C@@H]1CC=C)C)c2ccccc2	0.010569	0.007338	0.013175	0.009139	0.015564	0.005590	0.014000	0.021786	0.019726	0.019726	0.019726	0.008208	0.014011	0.008242	0.013631	1	Low
C[C@@H](c1ccc(cc1)NCC(=C)C)C(=O)[O-]	0.013147	0.017112	0.020161	0.012317	0.009138	0.009024	0.016194	0.009793	0.012887	0.012887	0.012887	0.018207	0.027076	0.011121	0.017632	1	Low
C[C@H](Cc1ccccc1)[NH2+][C@@H](C#N)c2ccccc2	0.005752	0.007487	0.010782	0.006163	0.006588	0.004515	0.006067	0.006572	0.005155	0.005155	0.005155	0.008427	0.009009	0.003711	0.007566	1	Low
C[C@@H](CC(c1ccccc1)(c2ccccc2)C(=O)N)[NH+](C)C	0.009868	0.012848	0.009371	0.004601	0.007874	0.003939	0.005030	0.006543	0.005998	0.005998	0.005998	0.006974	0.010753	0.001845	0.007528	1	Low
Cc1c(c(=O)n(n1C)c2ccccc2)NC(=O)[C@H](C)[NH+](C)C	0.031447	0.039354	0.058081	0.027266	0.035237	0.017984	0.051456	0.048958	0.031148	0.031148	0.031148	0.048114	0.051948	0.021834	0.036215	1	Low
c1ccc(cc1)[C@@H](C(=O)[O-])O	0.005747	0.007479	0.009409	0.005385	0.005256	0.003946	0.009119	0.009879	0.006014	0.006014	0.006014	0.008415	0.010811	0.004638	0.007557	1	Low
CC[C@](C)(C[NH+]	0.007377	0.007447	0.006667	0.006140	0.006545	0.004502	0.006036	0.009836	0.004274	0.004274	0.004274	0.008368	0.008929	0.006481	0.005000	1	Low

Fig. 19 Overview of the labeled dataset with 18 columns obtained after cluster profiling

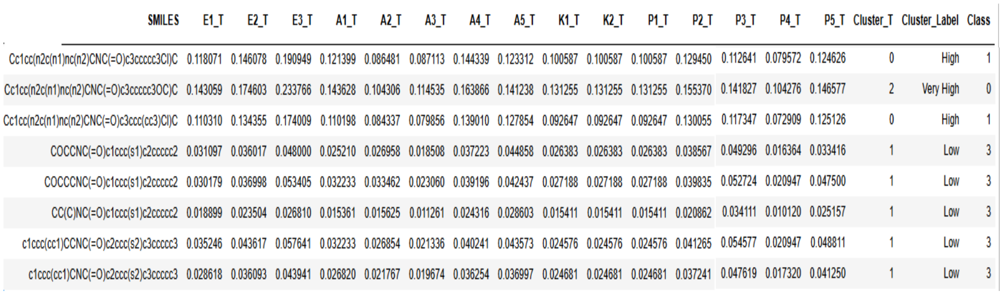
After clustering profiling, a labeled dataset has been added with one more column for specifying the cluster label that specifies its quality. Now, the labeled dataset has 18 columns, with 1 column as a SMILE string, 15 columns as JS values, 1 indicating cluster number, and 1 indicating cluster label, as depicted in Figure 19.

Now, the dataset has been added with one more column specifying the outcome variable by assigning the class value for a Cluster_Label as specified in Table 8.

After assigning class values to each Cluster_Label, a labeled dataset has been added with one more column for specifying the class to which SMILE belongs, which is the outcome variable. Now, the labeled dataset has 19 columns, with 1 column as a SMILE string, 15 columns as JS values, 1 indicating cluster number, 1 indicating cluster label, and 1 indicating class, as depicted in Figure 20.

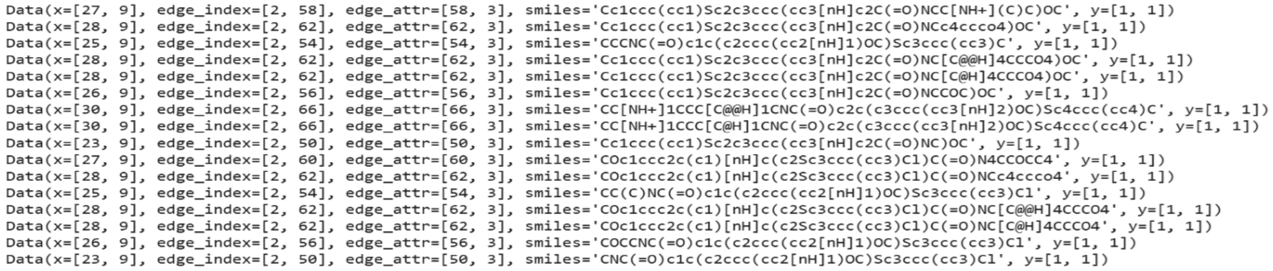
Table 8. Cluster-labels with their assigned class values based on analysis of JS

S. No.	Cluster_Label	Cluster Number	Class Value Assigned
1	**Very High**	**Cluster 2**	**0**
2	High	Cluster 0	1
3	Moderate	Cluster 3	2
4	Low	Cluster 1	3



SMILES	E1_T	E2_T	E3_T	A1_T	A2_T	A3_T	A4_T	A5_T	K1_T	K2_T	P1_T	P2_T	P3_T	P4_T	P5_T	Cluster_T	Cluster_Label	Class
Cc1cc(n2c(n1)nc(n2)CNC(=O)c3ccccc3Cl)C	0.118071	0.146078	0.190949	0.121399	0.086481	0.087113	0.144339	0.123312	0.100587	0.100587	0.100587	0.129450	0.112641	0.079572	0.124626	0	High	1
Cc1cc(n2c(n1)nc(n2)CNC(=O)c3ccccc3OC)C	0.143059	0.174603	0.233766	0.143628	0.104306	0.114535	0.163866	0.141238	0.131255	0.131255	0.131255	0.155370	0.141827	0.104276	0.146577	2	Very High	0
Cc1cc(n2c(n1)nc(n2)CNC(=O)c3ccc(cc3)Cl)C	0.110310	0.134355	0.174009	0.110198	0.084337	0.079856	0.139010	0.127854	0.092647	0.092647	0.092647	0.130055	0.117347	0.072909	0.125126	0	High	1
COCCNC(=O)c1ccc(s1)c2ccccc2	0.031097	0.036017	0.048000	0.025210	0.026958	0.018508	0.037223	0.044858	0.026383	0.026383	0.026383	0.038567	0.049296	0.016364	0.033416	1	Low	3
COCCCNC(=O)c1ccc(s1)c2ccccc2	0.030179	0.036998	0.053405	0.032233	0.033462	0.023060	0.039196	0.042437	0.027188	0.027188	0.027188	0.039835	0.052724	0.020947	0.047500	1	Low	3
CC(C)NC(=O)c1ccc(s1)c2ccccc2	0.018899	0.023504	0.026810	0.015361	0.015625	0.011261	0.024316	0.028603	0.015411	0.015411	0.015411	0.020862	0.034111	0.010120	0.025157	1	Low	3
c1ccc(cc1)CCNC(=O)c2ccc(s2)c3ccccc3	0.035246	0.043617	0.057641	0.032233	0.026854	0.021336	0.040241	0.043573	0.024576	0.024576	0.024576	0.041265	0.054577	0.020947	0.048811	1	Low	3
c1ccc(cc1)CNC(=O)c2ccc(s2)c3ccccc3	0.028618	0.036093	0.043941	0.026820	0.021767	0.019674	0.036254	0.036997	0.024681	0.024681	0.024681	0.037241	0.047619	0.017320	0.041250	1	Low	3

Fig. 20 Overview of the labeled dataset with 19 columns obtained after assigning the class value



```
Data(x=[27, 9], edge_index=[2, 58], edge_attr=[58, 3], smiles='Cc1ccc(cc1)Sc2c3ccc(cc3[nH]c2C(=O)NCC[NH+](C)C)OC', y=[1, 1])
Data(x=[28, 9], edge_index=[2, 62], edge_attr=[62, 3], smiles='Cc1ccc(cc1)Sc2c3ccc(cc3[nH]c2C(=O)NCc4ccco4)OC', y=[1, 1])
Data(x=[25, 9], edge_index=[2, 54], edge_attr=[54, 3], smiles='CCCNC(=O)c1c(c2ccc(cc2[nH]1)OC)Sc3ccc(cc3)C', y=[1, 1])
Data(x=[28, 9], edge_index=[2, 62], edge_attr=[62, 3], smiles='Cc1ccc(cc1)Sc2c3ccc(cc3[nH]c2C(=O)NC[C@@H]4CCCO4)OC', y=[1, 1])
Data(x=[28, 9], edge_index=[2, 62], edge_attr=[62, 3], smiles='Cc1ccc(cc1)Sc2c3ccc(cc3[nH]c2C(=O)NC[C@H]4CCCO4)OC', y=[1, 1])
Data(x=[26, 9], edge_index=[2, 56], edge_attr=[56, 3], smiles='Cc1ccc(cc1)Sc2c3ccc(cc3[nH]c2C(=O)NCCOC)OC', y=[1, 1])
Data(x=[30, 9], edge_index=[2, 66], edge_attr=[66, 3], smiles='CC[NH+]1CCC[C@@H]1CNC(=O)c2c(c3ccc(cc3[nH]2)OC)Sc4ccc(cc4)C', y=[1, 1])
Data(x=[30, 9], edge_index=[2, 66], edge_attr=[66, 3], smiles='CC[NH+]1CCC[C@H]1CNC(=O)c2c(c3ccc(cc3[nH]2)OC)Sc4ccc(cc4)C', y=[1, 1])
Data(x=[23, 9], edge_index=[2, 50], edge_attr=[50, 3], smiles='Cc1ccc(cc1)Sc2c3ccc(cc3[nH]c2C(=O)NC)OC', y=[1, 1])
Data(x=[27, 9], edge_index=[2, 60], edge_attr=[60, 3], smiles='COc1ccc2c(c1)[nH]c(c2Sc3ccc(cc3)Cl)C(=O)N4CCOCC4', y=[1, 1])
Data(x=[28, 9], edge_index=[2, 62], edge_attr=[62, 3], smiles='COc1ccc2c(c1)[nH]c(c2Sc3ccc(cc3)Cl)C(=O)NCc4ccco4', y=[1, 1])
Data(x=[25, 9], edge_index=[2, 54], edge_attr=[54, 3], smiles='CC(C)NC(=O)c1c(c2ccc(cc2[nH]1)OC)Sc3ccc(cc3)Cl', y=[1, 1])
Data(x=[28, 9], edge_index=[2, 62], edge_attr=[62, 3], smiles='COc1ccc2c(c1)[nH]c(c2Sc3ccc(cc3)Cl)C(=O)NC[C@@H]4CCCO4', y=[1, 1])
Data(x=[28, 9], edge_index=[2, 62], edge_attr=[62, 3], smiles='COc1ccc2c(c1)[nH]c(c2Sc3ccc(cc3)Cl)C(=O)NC[C@H]4CCCO4', y=[1, 1])
Data(x=[26, 9], edge_index=[2, 56], edge_attr=[56, 3], smiles='COCCNC(=O)c1c(c2ccc(cc2[nH]1)OC)Sc3ccc(cc3)Cl', y=[1, 1])
Data(x=[23, 9], edge_index=[2, 50], edge_attr=[50, 3], smiles='CNC(=O)c1c(c2ccc(cc2[nH]1)OC)Sc3ccc(cc3)Cl', y=[1, 1])
```

Fig. 21 Overview of the torch geometric objects for 1.048 million SMILES

4.3. Phase-III

To make the data ready for building a GAT model, each SMILE string of 1.048 million drug-like molecules is converted into the form of a graph by using the from_smiles() function from the Torch Geometric library and then converted into a Torch Geometric object by using torch.tensor().

Figure 21 depicts the overview of the transformed torch geometric object data for 1.048 million SMILES.

Now, the above transformed dataset of torch geometric objects is divided into a 75% train dataset and a 25% test dataset.

The GAT algorithm is trained on the training dataset by using the Python torch geometric library. After training, the GAT Model is evaluated on the test dataset over 25 epochs. In all the epochs, the training and testing accuracy are increased while the training and testing loss are decreased. Figure 22 depicts the graphs showing training, testing accuracy, and loss over 25 epochs, and Table 9 summarizes the performance.

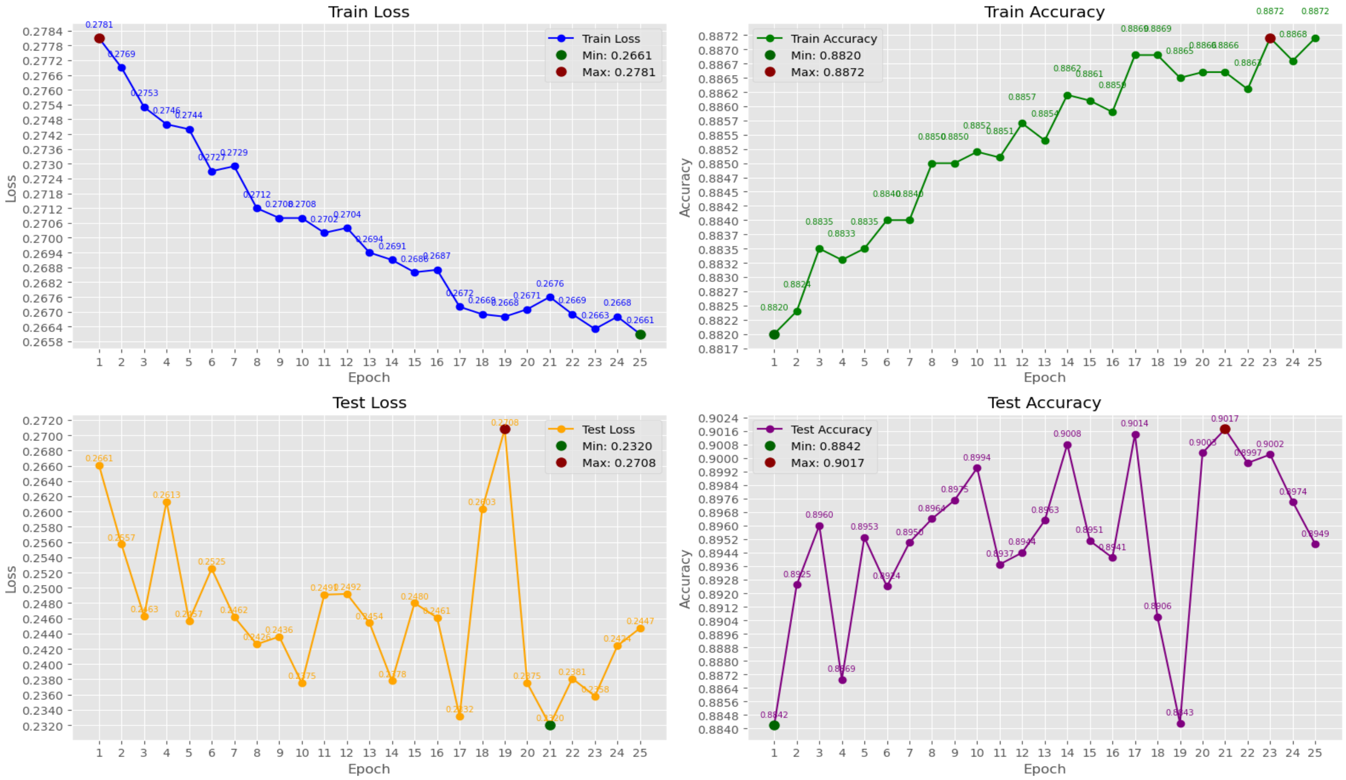


Fig. 22 Accuracy and loss obtained during training and testing over 25 epochs

Table 9. Training and testing performance over 25 epochs

	No. of Epochs	Range of Accuracy	Range of Loss
Training	25	88.20% to 88.72%	27.81% to 26.61%
Testing	25	88.42% to 90.17%	27.08% to 23.20%

From the performance analysis of training and testing of GATs over 25 epochs, represented in Figure 22 and Table 9, it is observed that,

- In training, accuracy has increased steadily from 88.20% to 88.72%, and the loss has decreased consistently from 27.81% to 26.61%; this indicates the effectiveness and consistency of the model in learning patterns from the training data.
- In testing, accuracy and the loss have sudden but little fluctuations between 88.42% and 90.17% and 27.08% and 23.20%, respectively; this indicates the efficiency of the model in processing new data.
- There is a very small difference in the training and the testing accuracy; this indicates that there is no overfitting.

Overall, the GAT model shows a generalized performance.

The confusion matrix of testing, which had 262143 torch geometric objects, is depicted in Figure 23.

From the confusion matrix depicted in Figure 23,

- it is observed that the GAT model has balanced performance across all the classes, as it shows high true positives.
- Class-wise performance metrics, such as precision, recall, and F1-score, are computed and are represented in Table 10 and also depicted in Figure 24.
- Overall performance metrics, such as accuracy, macro precision, macro recall, and macro F1-score, are computed and are represented in Table 11 and also depicted in Figure 25.

Class-wise and overall performance metrics of the GAT model, represented in Figures 24 and 25, show that the model has shown a strong, consistent classification with an overall high accuracy of 89.5%. This demonstrates the GAT model's robustness in predicting drug efficacy.

Now, when a SMILE is given as an input to the GAT Model, it predicts the output of that SMILE as either 0 or 1 or 2 or 3, as depicted in Figure 26. Based on the output of the model, one can decide to proceed further or not.

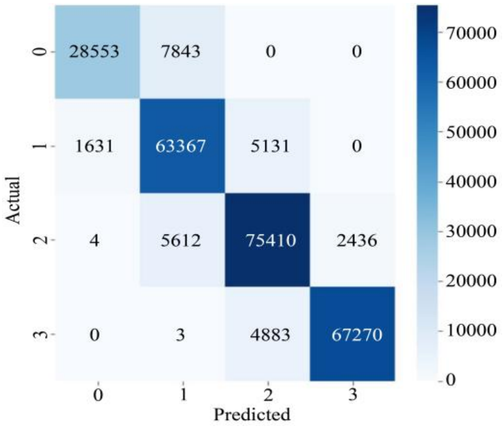


Fig. 23 GAT model's confusion matrix of testing after 25 epochs

Table 10. GAT model's class-wise performance metrics of testing over 25 epochs

Class	Precision	Recall	F1-Score
0	94.6%	78.4%	85.8%
1	82.5%	90.4%	86.2%
2	88.3%	90.3%	89.3%
3	96.5%	93.2%	94.8%

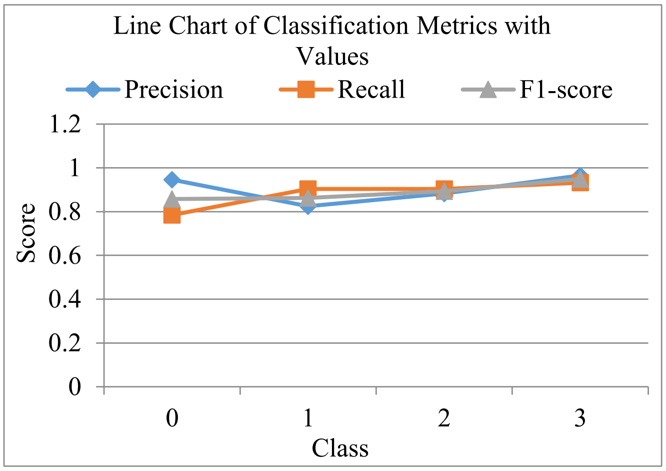


Fig. 24 Line chart representing the GAT model's class-wise performance of testing

Table 11. GAT model's overall performance metrics of testing over 25 epochs

No. of Epochs	Accuracy	Precision	Recall	F1-Score
25	89.5%	90.5%	88.1%	89.2%

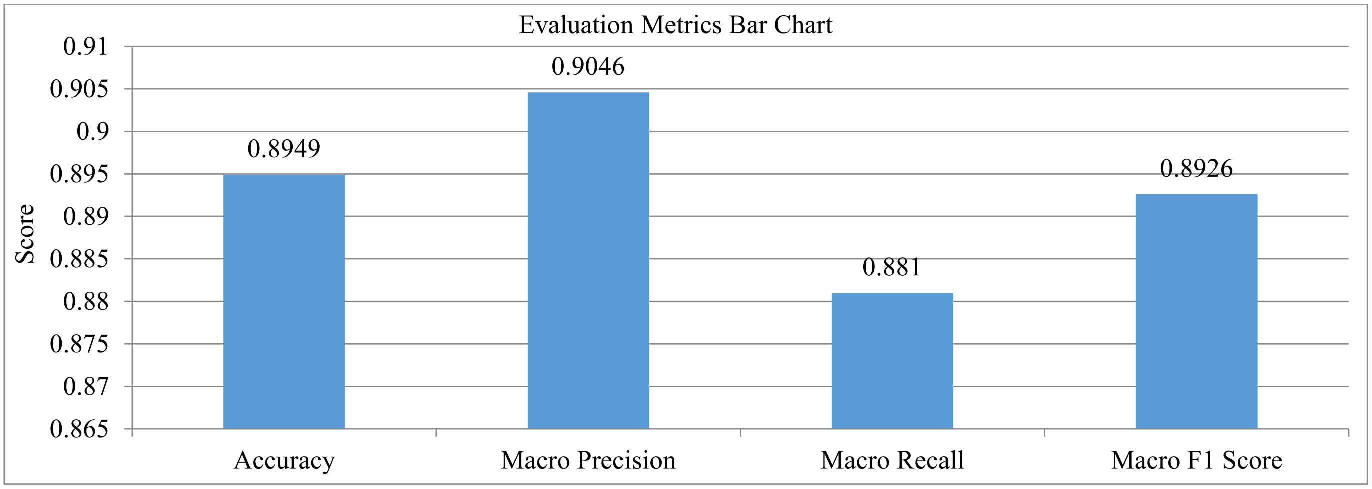
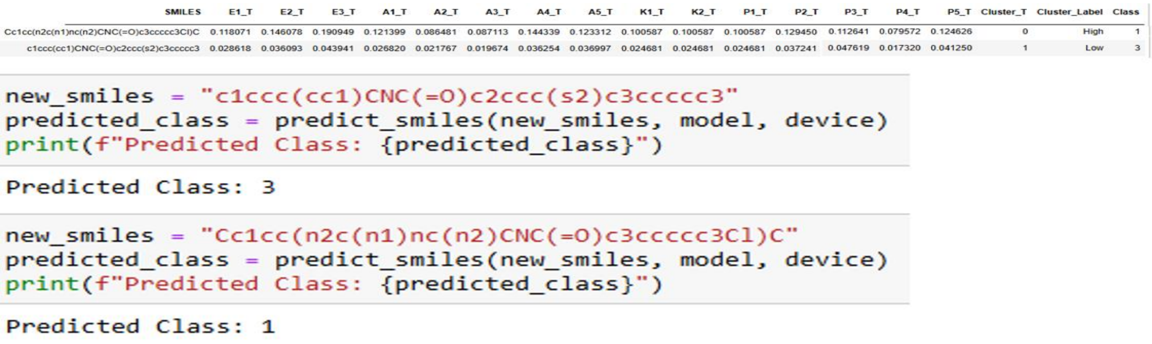


Fig. 25 Bar chart representing the GAT model's performance in testing



SMILES	E1_T	E2_T	E3_T	A1_T	A2_T	A3_T	A4_T	A5_T	K1_T	K2_T	P1_T	P2_T	P3_T	P4_T	P5_T	Cluster_T	Cluster_Label	Class
Cc1cc(n2c(n1)nc(n2)CNC(=O)c3ccccc3Cl)C	0.118071	0.146078	0.190949	0.121399	0.086481	0.087113	0.144339	0.123312	0.100587	0.100587	0.100587	0.129450	0.112641	0.079572	0.124626	0	High	1
c1ccc(cc1)CNC(=O)c2ccc(s2)c3ccccc3	0.028618	0.036093	0.043941	0.026820	0.021767	0.019674	0.036254	0.036997	0.024681	0.024681	0.024681	0.037241	0.047619	0.017320	0.041250	1	Low	3

```
new_smiles = "c1ccc(cc1)CNC(=O)c2ccc(s2)c3ccccc3"
predicted_class = predict_smiles(new_smiles, model, device)
print(f"Predicted Class: {predicted_class}")

Predicted Class: 3
```

```
new_smiles = "Cc1cc(n2c(n1)nc(n2)CNC(=O)c3ccccc3Cl)C"
predicted_class = predict_smiles(new_smiles, model, device)
print(f"Predicted Class: {predicted_class}")

Predicted Class: 1
```

Fig. 26 Efficacy prediction for a new SMILE using GAT model

Figure 26 demonstrates that the GAT model is capable of accurately predicting the efficacy class for a new SMILES representation.

5. Conclusion

NSCLC represents 85% of LC cases, is the deadliest disease with a higher occurrence in Asia, and is also the second most common cancer in India. NSCLC occurs due to various alterations, and it is difficult to find the exact alteration causing it. Even most of the drugs become resistant after being treated for some time. Hence, there will always be a huge need to develop a new drug, especially a multitargeted drug, but the traditional process of developing a drug is expensive and takes several years, and most of the drugs fail at clinical trials. To address this, MMDEP-GAT (Multifaceted and Multitargeted Drug Efficacy Prediction Leveraging Graph Attention Networks) is proposed. MMDEP-GAT works in 3 phases: the first phase is to obtain an unlabeled dataset using Jaccard similarities of 1.048 million SMILES with reference to 15 FDA-approved drugs; the second phase is to convert the unlabeled dataset into a labeled dataset using K-means cluster profiling; and the third phase is to train GAT algorithm and it is evaluated on test dataset over 25 epochs to predict efficacy of new SMILE/Drug. The GAT Model showed an accuracy of 89.5% and it can be used at the initial step in designing a new drug for NSCLC before entering into the virtual screening, docking, and simulation step.

5.1. Future Scope

The current framework predicted the efficacy of new SMILES for identified key targets such as PDGFR, ALK, KRAS, and EGFR that cause NSCLC. Extending this method to another type of cancer increases the chance of identifying a new drug/SMILE very fast. Usage of these models in the laboratory will enhance the process of identifying a lead compound with reduced time and cost.

References

[1] Mahmood Araghi et al., “Recent Advances in Non-Small Cell Lung Cancer Targeted Therapy; An Update Review,” *Cancer Cell International*, vol. 23, no. 1, pp. 1-25, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[2] Manas Gunani et al., “Spotlight on Lung Cancer Disparities in India,” *JCO Global Oncology*, vol. 11, no. 11, pp. 1-7, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[3] Freddie Bray et al., “Global Cancer Statistics 2022: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries,” *CA: A Cancer Journal for Clinicians*, vol. 74, no. 3, pp. 229-263, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[4] Jialin Zhou et al., “Global Burden of Lung Cancer in 2022 and Projections to 2050: Incidence and Mortality Estimates from GLOBOCAN,” *Cancer Epidemiology*, vol. 93, pp. 1-11, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[5] Jinto Edakkalathoor George et al., “Global Trends in Lung Cancer Incidence and Mortality by Age, Gender and Morphology and Forecast: A Bootstrap-Based Analysis,” *Lung Cancer*, vol. 205, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[6] Pankaj Garg et al., “Advances in Non-Small Cell Lung Cancer: Current Insights and Future Directions,” *Journal of Clinical Medicine*, vol. 13, no. 14, pp. 1-24, 2024.[CrossRef] [Google Scholar] [Publisher Link]

[7] David C. Planchard et al., “Metastatic Non-Small Cell Lung Cancer: ESMO Clinical Practice Guidelines for Diagnosis, Treatment and Follow-Up,” *Annals of Oncology*, vol. 29, no. 4, pp. iv192-iv237, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[8] Yaser Alduais et al., “Non-Small Cell Lung Cancer (NSCLC): A Review of Risk Factors, Diagnosis, and Treatment,” *Medicine*, vol. 102, no. 8, pp. 1-5, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[9] Mathieu Chevallier et al., “Oncogenic Driver Mutations in Non-Small Cell Lung Cancer: Past, Present And Future,” *World Journal of Clinical Oncology*, vol. 12, no. 4, pp. 217-237, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[10] Alex Friedlaender et al., “Oncogenic Alterations in Advanced NSCLC: A Molecular Super-Highway,” *Biomarker Research*, vol. 12, no. 1, pp. 1-30, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[11] Wu Quanyang et al., “Artificial Intelligence in Lung Cancer Screening: Detection, Classification, Prediction, and Prognosis,” *Cancer Medicine*, vol. 13, no. 7, pp. 1-19, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[12] Gabriel Dernbach et al., “Dissecting AI-Based Mutation Prediction in Lung Adenocarcinoma: A Comprehensive Real-World Study,” *European Journal of Cancer*, vol. 211, pp. 1-9, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[13] Hind M AlOsaimi et al., “AI Models for the Identification of Prognostic and Predictive Biomarkers in Lung Cancer: A Systematic Review and Meta-Analysis,” *Frontiers in Oncology*, vol. 15, pp. 1-14, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[14] Tingshan He et al., “Artificial Intelligence Predictive System of Individual Survival Rate for Lung Adenocarcinoma,” *Computational and Structural Biotechnology Journal*, vol. 20, pp. 2352-2359, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[15] Mohammed Kanan et al., “AI-Driven Models for Diagnosing and Predicting Outcomes in Lung Cancer: A Systematic Review and Meta-Analysis,” *Cancers*, vol. 16, no. 3, pp. 1-18, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[16] Juan Carlos Restrepo et al., “Advances in Genomic Data and Biomarkers: Revolutionizing NSCLC Diagnosis and Treatment,” *Cancers*, vol. 15, no. 13, pp. 1-30, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[17] Dolly A. Parasrampuria, Leslie Z. Benet, and Amarnath Sharma, “Why Drugs Fail in Late Stages of Development: Case Study Analyses from the Last Decade and Recommendations,” *The AAPS Journal*, vol. 20, no. 3, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[18] Petar Veličković et al., “Graph Attention Networks,” *arXiv Preprint*, pp. 1-12, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[19] Bharti Khemani et al., “A Review of Graph Neural Networks: Concepts, Architectures, Techniques, Challenges, Datasets, Applications, and Future Directions,” *Journal of Big Data*, vol. 11, no. 1, pp. 1-43, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[20] Antonio Lavecchia, “Advancing Drug Discovery with Deep Attention Neural Networks,” *Drug Discovery Today*, vol. 29, no. 8, pp. 1-17, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[21] Monal Yuwanati et al., “Graph Attention Networks for Predicting Drug-Gene Association of Glucocorticoid in Oral Squamous Cell Carcinoma: A Comparison with Graphsage,” *PloS one*, vol. 20, no. 7, pp. 1-11, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[22] Matthew P. Smeltzer et al., “The International Association for the Study of Lung Cancer Global Survey on Molecular Testing in Lung Cancer,” *Journal of Thoracic Oncology*, vol. 15, no. 9, pp. 1434-1448, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[23] Xin Xie et al., “Recent Advances in Targeting the “Undruggable” Proteins: from Drug Discovery to Clinical Trials,” *Signal Transduction and Targeted Therapy*, vol. 8, no. 1, pp. 1-71, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[24] Ashwini Kumar, and Awanish Kumar, “Non-Small-Cell Lung Cancer-Associated Gene Mutations and Inhibitors,” *Advances in Cancer Biology-Metastasis*, vol. 6, pp. 1-5, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[25] Dorine de Jong et al., “Novel Targets, Novel Treatments: The Changing Landscape of Non-Small Cell Lung Cancer,” *Cancers*, vol. 15, no. 10, pp. 1-21, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[26] Ha Yeong Choi, and Ji-Eun Chang, “Targeted Therapy for Cancers: From Ongoing Clinical Trials to FDA-Approved Drugs,” *International Journal of Molecular Sciences*, vol. 24, no. 17, pp. 1-27, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[27] Shivani Sharma et al., “Comprehensive Genomic Profiling of Indian Patients with Lung Cancer,” *JCO Global Oncology*, vol. 11, no. 11, pp. 1-13, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[28] Shigeto Nishikawa, and Tomoo Iwakuma, “Drugs Targeting P53 Mutations with FDA Approval and in Clinical Trials,” *Cancers*, vol. 15, no. 2, pp. 1-21, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[29] T Negri et al., “Evidence for PDGFRA, PDGFRB and KIT Deregulation in an NSCLC Patient,” *British Journal of Cancer*, vol. 96, no. 1, pp. 180-181, 2007. [CrossRef] [Google Scholar] [Publisher Link]

[30] Ammad Ahmad Farooqi, and Zahid H. Siddik, “Platelet-Derived Growth Factor (PDGF) Signalling in Cancer: Rapidly Emerging Signalling Landscape,” *Cell Biochemistry and Function*, vol. 33, no. 5, pp. 257-265, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[31] A. Nigel Brooks, Elaine Kilgour, and Paul D. Smith, “Molecular Pathways: Fibroblast Growth Factor Signaling: A New Therapeutic Opportunity in Cancer,” *Clinical Cancer Research*, vol. 18, no. 7, pp. 1855-1862, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[32] Yang Yang et al., “Inhibition of PDGFR by CP-673451 Induces Apoptosis and Increases Cisplatin Cytotoxicity in NSCLC Cells Via Inhibiting the Nrf2-Mediated Defense Mechanism,” *Toxicology Letters*, vol. 295, pp. 88-98, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[33] Kiyoshi Okamoto et al., “Antitumor Activities of the Targeted Multi-Tyrosine Kinase Inhibitor Lenvatinib (E7080) Against RET Gene Fusion-Driven Tumor Models,” *Cancer Letters*, vol. 340, no. 1, pp. 97-103, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[34] G. Minniti et al., “Chemotherapy for Glioblastoma: Current Treatment and Future Perspectives for Cytotoxic and Targeted Agents,” *Anticancer Research*, vol. 29, no. 12, pp. 5171-5184, 2009. [CrossRef] [Google Scholar] [Publisher Link]

[35] Jens Gille, “Antiangiogenic Cancer Therapies Get their Act Together: Current Developments and Future Prospects of Growth Factor-and Growth Factor Receptor-Targeted Approaches,” *Experimental Dermatology*, vol. 15, no. 3, pp. 175-186, 2006. [CrossRef] [Google Scholar] [Publisher Link]

[36] Edward S Kim et al., “EGFR Tyrosine Kinase Inhibitors for EGFR Mutation-Positive Non-Small-Cell Lung Cancer: Outcomes in Asian Populations,” *Future Oncology*, vol. 17, no. 18, pp. 2395-2408, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[37] Victor Zia et al., “Advancements of ALK Inhibition of Non-Small Cell Lung Cancer: A Literature Review,” *Translational Lung Cancer Research*, vol. 12, no. 7, pp. 1563-1574, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[38] Vivek Yadav et al., “Advancing Lung Cancer Treatment through ALK Receptor-Targeted Drug Metabolism and Pharmacokinetics,” *Computational Methods for Rational Drug Design*, pp. 477-491, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[39] Tzu-Rong Peng et al., “Comparative Efficacy of Adagrasib and Sotorasib in KRAS G12C-Mutant NSCLC: Insights from Pivotal Trials,” *Cancers*, vol. 16, no. 21, pp. 1-13, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[40] Jian Li et al., “Efficacy and Safety of Avapritinib in Treating Unresectable or Metastatic GIST: A Phase I/II, Open-Label, Multicenter Study,” *The Oncologist*, vol. 28, no. 2, pp. 187-e114, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[41] A. Teuber et al., “Avapritinib-Based SAR Studies Unveil a Binding Pocket in KIT and PDGFRA,” *Nature Communications*, vol. 15, no. 1, pp. 1-11, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[42] Lian Yu et al., “Multi-Target Angiogenesis Inhibitor Combined with PD-1 Inhibitors May Benefit Advanced Non-Small Cell Lung Cancer Patients in Late Line After Failure Of EGFR-TKI Therapy,” *International Journal of Cancer*, vol. 153, no. 3, pp. 635-643, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[43] Peiliang Wang et al., "Efficacy and Safety of Anti-PD-1 Plus Anlotinib in Patients with Advanced Non-Small-Cell Lung Cancer after Previous Systemic Treatment Failure-A Retrospective Study," *Frontiers in Oncology*, vol. 11, pp. 1-8, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[44] Luisa Carbognin et al., "Platelet-Derived Growth Factor Receptor Inhibitors for Non-Small Cell Lung Cancer: is the Odyssey Over?," *Expert Opinion on Investigational Drugs*, vol. 25, no. 6, pp. 635-638, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[45] Ping Wang et al., "Crenolanib, a PDGFR Inhibitor, Suppresses Lung Cancer cell Proliferation and Inhibits Tumor Growth in Vivo," *OncoTargets and Therapy*, vol. 7, pp. 1761-1768, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[46] G. Vlahovic et al., "Treatment with Imatinib Improves Drug Delivery and Efficacy in NSCLC Xenografts," *British Journal of Cancer*, vol. 97, no. 6, pp. 735-740, 2007. [CrossRef] [Google Scholar] [Publisher Link]

[47] Jingwei Li et al., "Artificial Intelligence-Assisted Decision Making for Prognosis and Drug Efficacy Prediction in Lung Cancer Patients: A Narrative Review," *Journal of Thoracic Disease*, vol. 13, no. 12, pp. 7021-7033, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[48] Jaryd R. Christie et al., "Artificial Intelligence in Lung Cancer: Bridging the gap Between Computational Power and Clinical Decision-Making," *Canadian Association of Radiologists Journal*, vol. 72, no. 1, pp. 86-97, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[49] Lulu Wang, "Deep Learning Techniques to Diagnose Lung Cancer," *Cancers*, vol. 14, no. 22, pp. 1-24, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[50] Shalini Wankhade, and S. Vigneshwari, "A Novel Hybrid Deep Learning Method for Early Detection of Lung Cancer using Neural Networks," *Healthcare Analytics*, vol. 3, pp. 1-13, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[51] Fahsai Nakarin et al., "Assisting Multitargeted Ligand Affinity Prediction of Receptor Tyrosine Kinases Associated Nonsmall Cell Lung Cancer Treatment with Multitasking Principal Neighborhood Aggregation," *Molecules*, vol. 27, no. 4, pp. 1-18, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[52] Grigoriy Gogoshin, and Andrei S. Rodin, "Graph Neural Networks in Cancer and Oncology Research: Emerging and Future Trends," *Cancers*, vol. 15, no. 24, pp. 1-19, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[53] Shuke Zhang et al., "SS-GNN: A Simple-Structured Graph Neural Network for Affinity Prediction," *ACS Omega*, vol. 8, no. 25, pp. 22496-22507, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[54] Shudong Wang et al., "MSGNN-DTA: Multi-Scale Topological Feature Fusion based on Graph Neural Networks for Drug-Target Binding Affinity Prediction," *International Journal of Molecular Sciences*, vol. 24, no. 9, pp. 1-17, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[55] Jing Chen, Xiaolin Yang, and Haoyu Wu, "A Multibranch Neural Network For Drug-Target Affinity Prediction Using Similarity Information," *ACS Omega*, vol. 9, no. 33, pp. 35978-35989, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[56] Conghao Wang, Gaurav Asok Kumar, and Jagath C. Rajapakse, "Drug Discovery and Mechanism Prediction with Explainable Graph Neural Networks," *Scientific Reports*, vol. 15, no. 1, pp. 1-14, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[57] Ingo Muegge, and Prasenjit Mukherjee, "An Overview of Molecular Fingerprint Similarity Search in Virtual Screening," *Expert Opinion on Drug Discovery*, vol. 11, no. 2, pp. 137-148, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[58] Malcolm J. McGregor, and Steven M. Muskal, "Pharmacophore Fingerprinting. 1. Application to QSAR and Focused Library Design," *Journal of Chemical Information and Computer Sciences*, vol. 39, no. 3, pp. 569-574, 1999. [CrossRef] [Google Scholar] [Publisher Link]

[59] Wenming Yang, Jiali Zou, and Le Yin, "Compound Property Prediction based on Multiple Different Molecular Features and Ensemble Learning," *CCKS 2022 - Evaluation Track*: *7th China Conference on Knowledge Graph and Semantic Computing Evaluations, CCKS 2022*, Qinhuangdao, China, pp. 57-69, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[60] Peter Willett, "Similarity Methods in Chemoinformatics," *Annual Review of Information Science and Technology*, vol. 43, no. 1, pp. 1-117, 2009. [CrossRef] [Google Scholar] [Publisher Link]

[61] Greg Landrum, "RDKit: A Software Suite for Cheminformatics, Computational Chemistry, and Predictive Modeling," *Greg Landrum*, vol. 8, no. 31.10, pp. 1-31, 2013. [Google Scholar]

[62] Peter Willett, "The Calculation of Molecular Structural Similarity: Principles and Practice," *Molecular Informatics*, vol. 33, no. 6-7, pp. 403-413, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[63] Dávid Bajusz, Anita Rácz, and Károly Héberger, "Why is Tanimoto Index an Appropriate Choice for Fingerprint-Based Similarity Calculations?," *Journal of Cheminformatics*, vol. 7, no. 1, pp. 1-13, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[64] John David MacCuish, and Norah E. MacCuish, *Clustering in Bioinformatics and Drug Discovery*, CRC Press, 1st ed., 2010. [CrossRef] [Google Scholar] [Publisher Link]

[65] Ahmed Alsayat, and Hoda El-Sayed, "Efficient Genetic K-Means Clustering for Health Care Knowledge Discovery," *2016 IEEE 14th International Conference on Software Engineering Research, Management and Applications (SERA)*, Towson, MD, USA, pp. 45-52, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[66] Souad Moufok, Anas Mouattah, and Khalid Hachemi, "K-Means and DBSCAN for Look-Alike Sound-Alike Medicines Issue," *International Journal of Data Mining, Modelling and Management*, vol. 16, no. 1, pp. 49-65, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[67] Rachid Sammouda, and Ali El-Zaart, "An Optimized Approach for Prostate Image Segmentation using K-Means Clustering Algorithm with Elbow Method," *Computational Intelligence and Neuroscience*, vol. 2021, no. 1, pp. 1-13, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[68] Ketan Rajshekhar Shahapure, and Charles Nicholas, "Cluster Quality Analysis using Silhouette Score," *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, Sydney, NSW, Australia, pp. 747-748, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[69] F. Mohamed Ilyas, and S. Silvia Priscila, "An Optimized Clustering Quality Analysis in K-Means Cluster using Silhouette Scores," *Explainable AI Applications for Human Behavior Analysis*, pp. 49-63, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[70] Parimala Palli, Satyasis Mishra, and P. Srinivasa Rao, "Inferring Compound Similarity: A Clustering Approach in Drug Discovery," *2024 1st International Conference on Cognitive, Green and Ubiquitous Computing (IC-CGU)*, Bhubaneswar, India, pp. 1-6, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[71] Kamel Mansouri et al., "Unlocking the Potential of Clustering and Classification Approaches: Navigating Supervised and Unsupervised Chemical Similarity," *Environmental Health Perspectives*, vol. 132, no. 8, pp. 1-12, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[72] Zonghan Wu et al., "A Comprehensive Survey on Graph Neural Networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 1, pp. 4-24, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[73] Foster Provost, Tom Fawcett, and Benjamin Lange, *Data Science for Business: What you need to know about Data Mining and Data-Analytic Thinking*, O'Reilly Media, 2013. [Google Scholar] [Publisher Link]

[74] Mohammad Hossin, and Md Nasir Sulaiman, "A Review on Evaluation Metrics for Data Classification Evaluations," *International Journal of Data Mining and Knowledge Management Process*, vol. 5 no. 2, pp. 1-11, 2015. [CrossRef] [Google Scholar] [Publisher Link]