IJHSR
ijhighschoolresearch.org
RESEARCH ARTICLE

Artificial Intelligence-Based Identification of Metastatic Targets and Natural Inhibitors in Non-Small Cell Lung Cancer

Prachet Patakula

Manthan School, Hyderabad, Telangana, 502032, India; patakula.prachet@gmail.com
Mentor: Nirupma Singh

ABSTRACT: Non-small cell lung cancer (NSCLC) is primarily lethal due to its high metastatic potential. In this study, we hypothesize to identify effective therapeutic targets and natural compounds to inhibit metastasis in NSCLC through an integrated computational framework that uses differential expression (DE) analysis, gene-interaction network analysis using mathematical modelling, virtual screening of natural compounds, and artificial intelligence. TCGA (The Cancer Genome Atlas) and GEO (Gene Expression Omnibus) were used to extract gene expression and clinical data. DE analysis was performed, and significant (p-value < 0.05) genes were identified. After validation for metastatic outcome, a final gene list of 850 genes was retained, which was analyzed using Reactome to map important pathways. STRING and Cytoscape were used for network construction and analysis. The top 10 hub genes were identified based on computed topological parameters. The genes were modelled with nonlinear ordinary differential equations (ODEs) to construct a dynamic network. The intersection of the top genes of static and dynamic networks presented 9 common genes. From the top genes, *BUB1* was shortlisted as the most important and relatively lesser-explored therapeutic target in NSCLC metastasis literature, based on a review of prior studies. Virtual Screening of natural compounds was run for *BUB1*, followed by building an artificial intelligence (AI) model for prioritization and feature-based interpretation of high-activity candidates. Thus, this study integrates gene expression data, network biology, dynamic systems modeling, and AI-augmented virtual screening to uncover regulatory hubs and therapeutic leads in NSCLC metastasis.
KEYWORDS: NSCLC, Bioinformatics, Cancer Metastasis, Natural Compounds, Protein-Protein Interactions.

Introduction

Lung cancer is one of the most common and leading causes of cancer deaths worldwide, with NSCLC (non-small cell lung cancer) being the most prevalent subtype.[1] Metastasis, the primary contributor to lung cancer mortality due to its ability to spread the disease systemically, has become a major focus of interest, particularly in the application of bioinformatics and artificial intelligence (AI).[2] From 2014 to 2018, the annual decline in lung cancer mortality in the U.S. accelerated to 5.5% in men and 4.4% in women, driven largely by advances in treatment for non-small cell lung cancer (NSCLC), whose two-year relative survival rate improved from 34% (2009–2010) to 42% (2015–2016). Despite these gains, NSCLC still accounts for the majority of lung cancer-related deaths.[3,4] The Indian subcontinent has a lower incidence rate of about 6.6 per 100,000 between 1990 and 2016 (10.3 to 11.2 in men and from 2.6 to 4.5 in women).[5] Lung cancer metastasis involves complex molecular signaling, particularly through protein–protein interactions that drive tumor spread.[6] Understanding these interactions is critical for identifying novel therapeutic targets and early biomarkers. Natural compounds have shown promise in inhibiting metastatic proteins but require computational modelling for effective screening.[2,7]

Although previous studies have mapped critical oncogenic pathways, predictive models integrating AI and PPI (protein–protein interaction) data remain limited. While traditional methods struggle to handle the high complexity and diversity of cancer metastasis, AI can significantly enhance the prediction of key metastatic drivers and their inhibition through *in silico* screening.[8] This integration can significantly bolster precision medicine efforts, with notable benefits for personalized cancer therapy. Major initiatives such as the TCGA (The Cancer Genome Atlas) and the Human Protein Atlas are attempting to investigate and better understand the molecular-level interactions behind metastasis.[9] Currently, the USA is the global leader in the development of AI-based mathematical modelling (use of algorithms to simulate biological behaviors and predict outcomes in complex systems) and drug discovery platforms to identify novel methods to target metastasis.[3]

Seven national bodies like ICMR, DBT, and CSIR are funding projects integrating bioinformatics and network biology (a systems biology approach to study interactions between biological entities) to combat cancer.[10] Furthermore, India has rich biodiversity, leading to increased interest in natural compound inhibition (the process of blocking disease-related proteins using naturally derived chemical compounds) as an anti-cancer drug.[11] AI and ML (machine learning) tools can swiftly analyze vast biological and medical datasets to predict protein functions, interactions, identify drug targets, and simulate molecular interactions.[8,12] In cancer research, AI models can detect hidden patterns in multi-omics data to forecast metastatic potential.[13] In the U.S. and India, AI is becoming central in drug repurposing, natural compound screening, and oncology diagnostics, with ML algorithms such as random forests, support vector machines, and deep learning models being used for PPI prediction.[14]

DOI: 10.36838/IJHSR88.90

Natural compound studies often rely on wet-lab approaches, which are time and resource-consuming, with computational approaches being generally underutilized.[15] Collaborations between AI developers and life scientists are still limited, and most research focuses on individual omics layers, rarely integrating multi-omics and network data.[16] Thus, it was hypothesized that AI-driven mathematical modelling, when integrated with network biology, can accurately predict PPIs driving NSCLC metastasis. Additionally, the computational screening of natural compound libraries can identify potential natural compound metastasis inhibitors, aiding drug development. The main objective of this paper is to contribute to the early prediction and prevention of metastasis in NSCLC patients, enable faster drug discovery by identifying bioactive natural compounds through AI models, and support the usage of indigenous medicinal compounds for cancer treatment, while contributing globally to the evolving AI-oncology intersection.

Methods

Data Collection and Pre-processing:

A multi-stage data collection and filtering process was used to identify robust gene candidates involved in NSCLC, integrating transcriptomic datasets from multiple sources and taking rigorous pre-processing steps. Differentially expressed genes were initially obtained from a TCGA-based dataset available on the Lung Cancer Explorer (LCE) portal. Both of the subtypes of LUAD (lung adenocarcinoma) and LUSC (lung squamous cell carcinoma) were taken into account during the analysis to ensure comprehensive subtype coverage. Non-coding and microRNA entries were filtered from the original dataset, with only protein-coding transcripts of likely functional significance retained. An adjusted p-value threshold of < 0.05 was chosen to pick up statistically significant genes, with a standardized mean difference (SMD) threshold of > 0.5 chosen to secure practical biological significance. SMD is a measure of effect size that expresses the difference between two group means relative to the variability (standard deviation) of the data, allowing comparison across studies with different measurement scales. Additionally, only genes were retained that were directionally consistent in both LUAD and LUSC, displaying uniform upregulation or uniform downregulation. To complement this data, expression and clinical metadata for LUAD and LUSC were retrieved from The Cancer Genome Atlas (TCGA) via the cBioportal platform. Only diploid samples, defined as exhibiting a regular genomic copy number without significant amplification or deletion, were selected for analysis. The dataset included z-score normalized mRNA expression data for both LUAD and LUSC diploid samples. Only those showing consistent upregulation or downregulation in both LUAD and LUSC subtypes were kept.

To further validate and emphasize relevance to cancer metastasis, three additional datasets were selected from the Gene Expression Omnibus (GEO): GSE161116, GSE166720, and GSE263726. These datasets were chosen based on their association with metastatic character in NSCLC. Specifically, they were prioritized because they included clear metastatic versus non-metastatic annotations (or metastatic stage grouping) and sufficient sample sizes to support reliable differential expression analysis. The DEG (differentially expressed gene(s)) lists for each of these datasets were generated using GEO2R, an in-built tool within GEO that implements standard statistical procedures to identify differentially expressed genes between two or more experimental groups. Default thresholds applied by GEO2R were used for the analysis, including a p-value cut-off of less than 0.05. The DEGs identified from each of the three studies were combined into a unified GEO-sourced DEG list. The union of the GEO-derived DEGs was intersected with the curated LCE–cBioPortal gene list to produce a final gene list.

Protein-Protein Interaction (PPI) Network Construction and Pathway Analysis:

To evaluate the structural relationships between metastasis-associated genes in NSCLC, a PPI network was constructed. The final list of 850 DEGs was queried against the STRING database using a custom Python script that implemented batch submission of gene symbols. The script parsed STRING's API and filtered results to include only interactions with a combined score ≥ 0.7 and limited to Homo sapiens. The resulting interaction table (in TSV 'tab-separated values' format) included columns for interacting protein pairs, interaction scores, and evidence types. This interaction matrix was imported into the Cytoscape software. Within Cytoscape, the Network Analyzer plugin computed multiple topological parameters. To gain functional insight into the metastasis-associated genes identified through expression integration and network modeling, pathway enrichment analysis was conducted using the Reactome Pathway Database. Statistical significance was determined using hypergeometric testing with Benjamini–Hochberg correction to control the false discovery rate (FDR).

Dynamic Modelling of PPI Networks:

While centrality highlights structurally important hubs, dynamic modeling evaluates how these genes behave over time under regulatory interactions, helping distinguish consistently influential drivers from nodes that are merely well connected. To simulate dynamic regulatory behavior within this network, a non-linear ODE model was implemented to capture the time evolution of protein concentrations across all the genes. A non-linear ordinary differential equation is an ODE in which the unknown function or its derivatives appear with powers, products, or other non-linear functions, so the equation cannot be written as a sum of terms where the function and its derivatives are only to the first power multiplied by functions of the independent variable. Each gene was modelled by the nonlinear equation $dx/dt = \alpha - \beta x + A \cdot x^2$. Both α and β were fixed at 1 across all genes, simplifying the dynamics while retaining general biochemical interpretability. Initial activity levels for each gene were randomly initialized within the range of 0.01 to 0.1 to mimic low resting expression. Simulations were conducted over 100 discrete time steps using the LSODA algorithm through SciPy's solve_ivp function, which adaptively

handles stiff and non-stiff systems. All simulation parameters, including step size and tolerances, were held constant across the entire gene set. The model produced a time-series matrix of gene expression trajectories. All the genes were retained throughout the simulation, and their dynamic profiles were ranked based on convergence behavior.

Virtual Screening of BUB1:

Molecular docking was performed to explore therapeutic inhibition of the key metastasis-associated gene *BUB1*. An *in silico* pipeline based on PyRx, AutoDock Tools, and Open Babel was used for the analysis. The 3D crystal structure of BUB1 was obtained from the RCSB Protein Data Bank in PDB (Protein Data Bank) format. The structure was pre-processed before docking using AutoDock Tools (ADT). This included: the removal of all water molecules, the addition of polar hydrogens to support accurate hydrogen bonding interactions, and the assignment of Kollman charges to all atoms. The following structure was then converted into PDBQT (AutoDock Protein Data Bank format with partial charges and atom types) format. For ligand preparation, a library of natural compounds was sourced from PubChem in SDF (Structure Data File) format and imported into PyRx. Using the Open Babel module integrated within PyRx, the compounds were converted into 3D conformers and energy-minimized using the MMFF94 force field to optimize their geometry. These processed ligands were then exported into PDBQT format, which is required for docking. A blind docking approach was employed to scan the entire surface of the BUB1 protein for potential binding sites. Within PyRx, the docking grid box was configured to encompass the full macromolecule to ensure that no potential binding cavity was overlooked. The center and dimensions of the grid box were manually set based on visual inspection of the protein's structure. Docking grid parameters were optimized once for the BUB1 binding site and then reused unchanged across all ligands to ensure consistent sampling and reproducibility. Docking was conducted using AutoDock Vina through the Vina Wizard interface in PyRx. Flexible ligand docking was performed against the prepared BUB1 protein, with AutoDock Vina applying its scoring algorithm to estimate binding affinities. The results were reported as binding free energies (ΔG, in kcal/mol), where more negative values indicated stronger predicted binding interactions. Following molecular docking, the top five ligands ranked by binding affinity were selected for detailed interaction analysis using BIOVIA Discovery Studio Visualizer. Upon opening each complex in Discovery Studio, the protein and ligand chains were visually inspected. For each loaded complex, the built-in "Receptor-Ligand Interactions" protocol was executed.

Machine Learning Model:

A supervised machine learning approach was implemented to predict the binding potential of natural compounds targeting the BUB1 protein. These compounds had previously been screened in the molecular docking pipeline and were selected from a curated natural product library retrieved from PubChem. The physicochemical descriptors of the ligands, such as molecular weight, topological polar surface area (TPSA), XLogP, the number of hydrogen bond donors and acceptors, rotatable bonds, and several additional drug-likeness indicators, were used as the feature set. TPSA is the sum of the surface areas of all polar atoms (typically oxygen and nitrogen) and their attached hydrogens in a molecule's 2D structure, used as a predictor of permeability and drug absorption. All descriptor data were compiled into a single feature matrix and paired with binary classification labels based on binding energy values derived from AutoDock Vina simulations. Compounds showing binding energies of –7.0 kcal/mol or better were labelled as 'active' and assigned a value of 1. Those with less favorable binding affinities were labelled 0. This threshold was selected to distinguish strong binders from weaker candidates while keeping the classification criterion both interpretable and stringent. The dataset was split into training and testing sets using an 80:20 ratio, with stratified sampling to maintain class distribution. Model performance was evaluated on the held-out test set using accuracy and ROC-AUC (with precision, recall, and F1-score reported as supporting metrics) to verify that predictive performance generalized beyond the training data. To interpret the influence of individual features on model output, SHAP (SHapley Additive exPlanations) values were computed using TreeExplainer. The SHAP summary plot highlighted the key features. In addition to binary class predictions, the model also generated probability scores for each compound using the predict_proba function. These confidence scores indicated the model's certainty in classifying a ligand as active and were subsequently used to prioritize candidates for downstream consideration. A final ranked list containing compound identifiers, predicted probabilities, and classification labels was exported for use in compound selection workflows.

Results and Discussion

Identification of Differentially Expressed Genes Across NSCLC Subtypes:

From the TCGA datasets, genes were considered differentially expressed based on standardized expression differences computed across metastatic and non-metastatic samples. In lung adenocarcinoma (LUAD), a total of 11,692 genes were differentially expressed, of which 7,066 were upregulated, and 4,626 were downregulated. Similarly, analysis of the lung squamous cell carcinoma (LUSC) dataset identified 12,039 differentially expressed genes (DEGs), including 6,799 upregulated and 5,240 downregulated. The intersection yielded 6,032 genes that were consistently upregulated or downregulated across both NSCLC subtypes. After performing DE analysis for chosen IDs, GSE161116, GSE166720, and GSE263726, a separate number of DEGs were identified for each. These GSE IDs were selected based on the presence of well-annotated metastatic phenotypes and control groups. GSE161116 returned 352 DEGs, 342 of which were downregulated. GSE166720 and GSE263726 identified 13,374 and 7,819 DEGs, respectively, with distributions that included both upregulated and downregulated candidates (Figure 1).

DOI: 10.36838/IJHSR88.90

After consolidation and removal of duplicates, the GEO-derived DEG pool comprised 16,861 unique genes. To establish a final metastasis-associated signature, the 6,032 subtype-consistent TCGA genes were intersected with the GEO-derived DEG pool. This cross-validation step resulted in a core set of 858 genes that were reproducibly dysregulated in both large-scale RNA-seq and independent microarray datasets. These 858 genes represent highly statistically robust and biologically validated candidates associated with NSCLC metastasis.

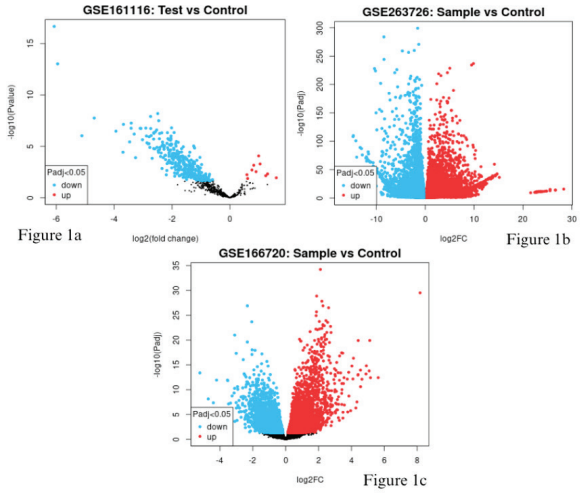

Figure 1: Differentially expressed genes in NSCLC metastasis studies from GEO (GSE161116, GSE166720, GSE263726). Volcano plots showing differential gene expression across the three GEO datasets. Each point represents a gene, plotted by $\log_2$ fold change (x-axis) and $-\log_{10}$ adjusted p-value (y-axis). Genes with statistically significant differential expression (adj p-value < 0.05) are colored: upregulated genes in red, downregulated genes in blue, and non-significant genes in black. These plots highlight the distribution and magnitude of transcriptional dysregulation between control and metastatic lung cancer samples in each study.

Network Construction and Hub Gene Identification:

The functional relationships among the 858 GEO-validated genes implicated in NSCLC metastasis were mapped in the form of a protein–protein interaction (PPI) network. The resulting network consisted of 858 nodes and 5,321 undirected edges, indicating a densely interconnected regulatory architecture (Figure 2). Topological features, including degree, betweenness, closeness, and clustering coefficient, were computed for each gene. Degree centrality was prioritized as the primary measure of node importance due to its intuitive representation of connectivity. Based on all these centrality metrics, the ten most highly connected hub genes were identified (Table 1). Among these, *BRCA1* showed the highest degree with 67 direct connections, marking it as a key regulatory node within the network.

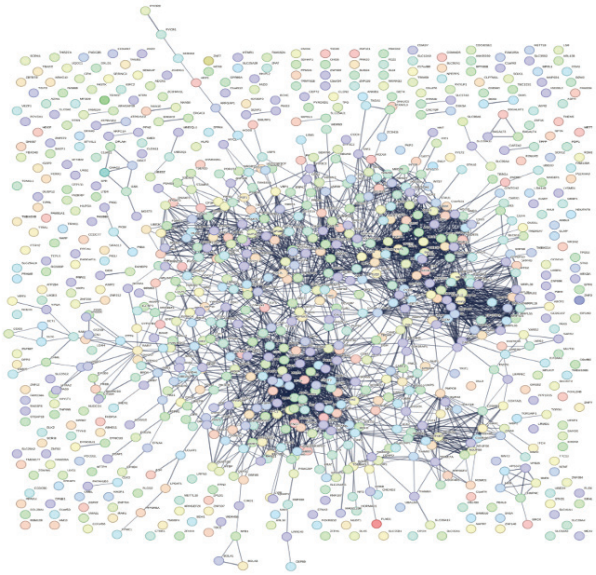

Figure 2: A Protein–Protein Interaction (PPI) Network of DEGs derived from TCGA (LUAD and LUSC) and GEO (GSE161116, GSE166720, GSE263726). The full PPI network was constructed using the STRING database for the 858 GEO-validated, TCGA-consistent differentially expressed genes (858 nodes; 5,321 edges). Round nodes represent proteins and edges indicate high-confidence interactions (combined score > 0.7).

Table 1: Summary of Top 10 Hub Genes Identified from Static Centrality Analysis of the NSCLC Metastasis Network. The table lists the top 10 hub genes ranked by degree centrality in the STRING-derived protein–protein interaction (PPI) network of 858 GEO-validated metastasis-associated genes.

S. No.	Gene	Degree	Betweenness Centrality	Closeness Centrality	Clustering Coefficient
1	*BRCA1*	67	0.1144026131	0.3372147182	0.1203075531
2	*EXO1*	57	0.01958042675	0.3070398643	0.2675438596
3	*POLR1B*	50	0.06076667154	0.3265674335	0.2563265306
4	*MRPS12*	48	0.01618940823	0.3116659492	0.4078014184
5	*NOP56*	47	0.02127416617	0.3097988875	0.3977798335
6	*FBL*	46	0.03830293692	0.3186619718	0.3806763285
7	*BYSL*	45	0.0215462708	0.2965997542	0.4141414141
8	*BUB1*	45	0.02350532586	0.3123382226	0.3696969697
9	*TOP2A*	45	0.01876130807	0.312203536	0.3808080808
10	*CDC45*	45	0.01509178348	0.3058724123	0.3707070707

Dynamic Modeling of Network Behavior:

To model the dynamic regulatory behavior of metastasis-associated genes in NSCLC, a nonlinear ordinary differential equation (ODE) simulation was employed using the STRING-derived protein–protein interaction (PPI) network. The validated list of 858 genes was mapped onto the network, and the corresponding adjacency matrix was extracted. Each node represented a protein and was modelled according to the equation $dx/dt = \alpha - \beta x + A \cdot x^2$, where x is the protein activity level, α is the synthesis rate, β is the degradation rate, and A is the weighted adjacency matrix. Initial conditions were randomly sampled between 0.01 and 0.1 to reflect low basal activity. Simulations were conducted for 50 time steps using the LSODA solver in SciPy's solve_ivp module. Nodes exhibiting oscillatory or null behavior were filtered out. A node was considered oscillatory if the time-series profile exhibited multiple

peaks or fluctuations without convergence, with a standard deviation > 0.1 over the final 50% of the simulation period. Final trajectories were compiled into a time-series CSV (comma-separated values) matrix, with temporal plots generated for select genes (Figure 3). All ten of the top-ranked genes from the static centrality analysis–*BRCA1, EXO1, TOP2A, NOP56, MRPS12, BUB1, CDC45, RRM2, NOP58,* and *FBL*–exhibited stable, non-zero equilibrium states. These profiles showed initial exponential increases in activity followed by smooth convergence to plateaus, typically within 10–15 time steps. To assess coherence between topological and kinetic importance, the top 10 dynamic nodes with the top 10 from static centrality measures were compared. Nine genes were shared: *BRCA1, BUB1, CDC45, EXO1, FBL, MRPS12, NOP56, NOP58,* and *TOP2A*. The only exception was BYSL, which ranked just outside the top 10 dynamically at position 11. When extended to the top 30 genes, 29 overlapped, demonstrating strong convergence between structural prominence and temporal persistence (Figure 4). This alignment supports the biological importance of the selected genes and affirms their prioritization for downstream experimental or therapeutic targeting.

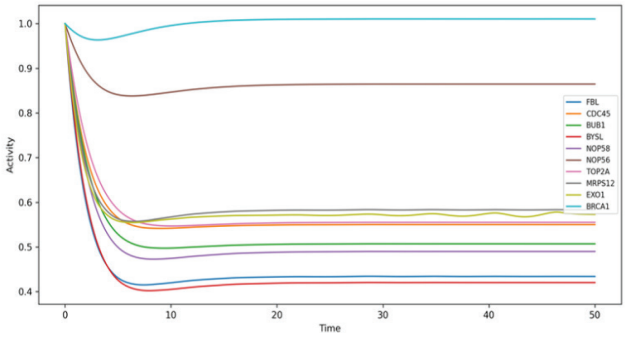

Figure 3: Dynamic Activity Profiles of Top 10 Hub Genes in Simulated PPI Network derived from TCGA (LUAD and LUSC) and GEO (GSE161116, GSE166720, GSE263726). Time-course trajectories depicting simulated activity dynamics of the top 10 hub genes (*BRCA1, EXO1, TOP2A, NOP56, MRPS12, BUB1, CDC45, RRM2, NOP58, FBL*) within the dynamic protein–protein interaction (PPI) network. Each curve represents the expression activity of a single gene over 50 simulation time steps based on a nonlinear ODE system. All trajectories converge to non-zero steady states.

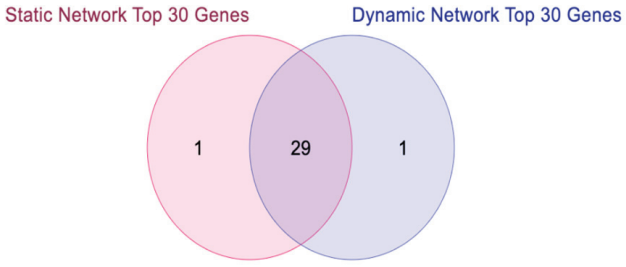

Figure 4: Comparison of Top-Ranked Genes Between Static and Dynamic Network Models derived from TCGA (LUAD and LUSC) and GEO (GSE161116, GSE166720, GSE263726). Venn diagram illustrating the overlap between the top 30 genes ranked by static degree centrality (left, red) and those prioritized through dynamic modeling of the same 858-gene protein–protein interaction network (right, blue). Of the 30 highest-ranked genes from each model, 29 were shared, demonstrating a high degree of agreement between structural centrality and dynamic stability-based prioritization. Only one gene was unique to each ranking list, validating the robustness of the consensus-derived gene set.

Pathway Enrichment Analysis:

After performing pathway analysis, a total of 107 pathways were significantly enriched (FDR < 0.05), with the majority clustering around cell cycle regulation, DNA replication, and checkpoint transitions. The top five enriched pathways identified were Mitotic G1 phase and G1/S transition, Cell Cycle, Mitotic, G1/S Transition, and Cell Cycle Checkpoints. These pathways collectively cover a substantial portion of the functional landscape defined by the 858-gene panel. Their prominence suggests a tightly regulated disruption of cell cycle progression and genomic maintenance in NSCLC metastasis. Many of the top-ranking hub genes from the network analysis, including *BRCA1, BUB1, TOP2A, CDC45*, and *EXO1*, were also present in these enriched pathways. Their recurrence across static centrality, dynamic simulation, and functional annotation further validates their mechanistic relevance.

Virtual Screening and Docking Analysis:

To identify candidate natural compounds capable of targeting key metastatic regulators in NSCLC, molecular docking was conducted for the BUB1 protein using a structure-based virtual screening approach. BUB1 was selected as the docking target due to its high topological centrality, temporal stability in dynamic modeling, and consistent enrichment in mitosis-related pathways. A total of 1,869 natural product-derived ligands were compiled from PubChem and prepared using Open Babel within the PyRx interface. AutoDock Vina returned binding free energy estimates (ΔG, in kcal/mol) for each ligand–protein complex. All results were aggregated and ranked by ascending ΔG values, and top compounds were identified (Table 2). The top binding affinities ranged from −11.7 to −9.6 kcal/mol, indicating strong interaction potential. The top-ranked compounds demonstrated consistent docking across multiple simulation runs and structural replicates. CID_135398501, the top hit, exhibited a predicted binding energy of −11.8 kcal/mol, placing it well within the range associated with high-affinity ligand binding. The majority of the top ligands contained heterocyclic cores and polar functional groups conducive to hydrogen bonding and hydrophobic pocket fitting. Structural analysis using PyMOL revealed favorable interactions within BUB1's ATP-binding domain and adjacent regulatory sites. Three of the five leading compounds –CID 135398501 (AKT inhibitor VIII), CID 11751922 (CDK inhibitor AT-7519), and CID 5154 (sanguinarine) are documented in PubChem as kinase inhibitors or bioactive antitumor agents, lending support to their pharmacophore relevance. The remaining hits show anticancer activity but lack specific kinase annotation. BIOVIA Discovery Studio provided a 2D interaction diagram for the top 5 ligands (Figure 5) showing the types of interactions.

DOI: 10.36838/IJHSR88.90

Table 2: Top Five Ligands Identified from Virtual Screening Against the BUB1 protein in NSCLC. List of top-performing ligands from structure-based virtual screening against the BUB1 protein using AutoDock Vina. Compounds were ranked based on binding affinity (ΔG, kcal/mol). Some of these compounds are known or suspected kinase inhibitors, suggesting therapeutic relevance in NSCLC metastasis inhibition.

S.No.	Rank	Compound Name / ID	Binding Energy (kcal/mol)
1	1	CID_135398501	−11.8
2	2	CID_5154	−11.7
3	3	CID_11751922	−11.5
4	4	CID_24882195	−11.5
5	5	CID_5281600	−11.4

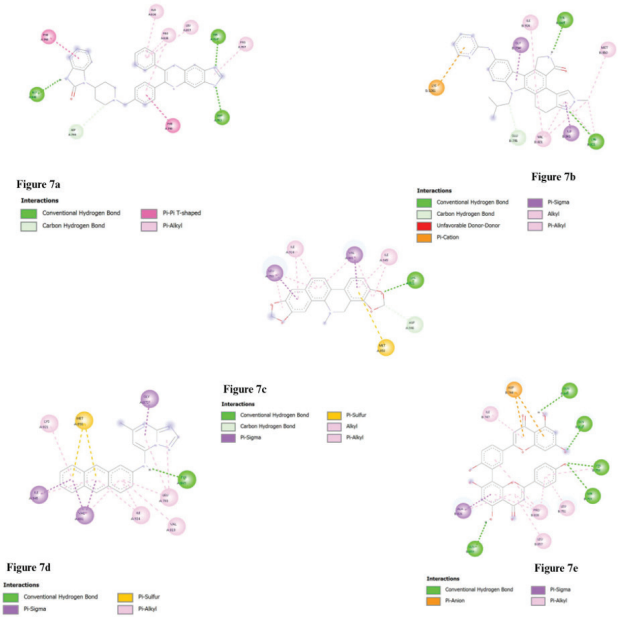

Figure 5: 2D representation of docking using Biovia Discovery Studio top poses after virtual screening. 2D interaction diagrams of the top five high-affinity ligands (pose 1) docked to the BUB1 protein (CID_135398501, CID_5154, CID_11751922, CID_24882195, CID_5281600). Visualizations were generated using BIOVIA Discovery Studio based on AutoDock Vina output poses.

Machine learning model:

To evaluate the drug-likeliness and binding potential of the screened natural compound library, a machine learning–based classification framework was employed to predict ligand activity against the BUB1 protein with two models, XGBoost and Random Forest. With both demonstrating acceptable predictive behavior, XGBoost was chosen for feature importance analysis and compound prioritization due to its interpretability via SHAP and efficiency in handling tabular data. Feature importance was explored using SHAP (SHapley Additive exPlanations), a technique that offers insight into how each input feature contributes to the final prediction. The SHAP summary plot showed that XLogP, topological polar surface area (TPSA), and the number of rotatable bonds had the greatest impact on classification outcomes. These descriptors, which reflect molecular hydrophobicity, polarity, and flexibility, are commonly linked to drug-like behavior. The model relied on these variables to distinguish between strong and weak binders. A global summary of these findings is presented in Figure 6, offering a clear visualization of overall feature relevance. Once the model was trained, it was used to assign binding activity probabilities to every compound in the dataset. Compounds predicted as active (based on classification probabilities) were flagged for prioritization. A total of 367 molecules met this criterion and were advanced for downstream consideration. The original training labels were based on docking scores below −7.0 kcal/mol, but the final model predictions relied solely on descriptor patterns. These selected compounds represented a wide range of chemical scaffolds and exhibited high confidence scores, making them strong candidates for redocking or experimental validation.

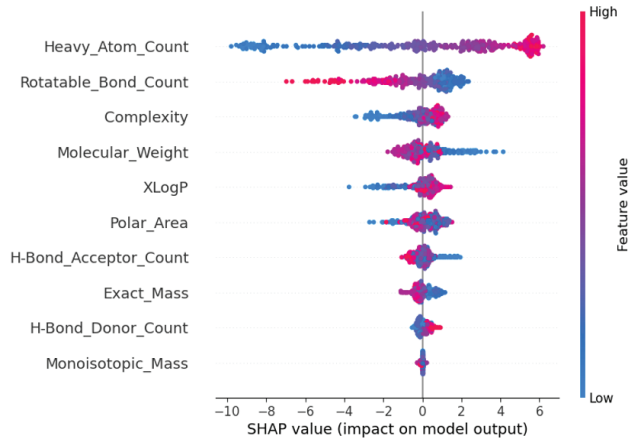

Figure 6: SHAP Summary Plot of Feature Importance in the XGBoost Model trained on the docked natural compound set (n = 1,869 ligands; "active" label defined as AutoDock Vina binding energy ≤ −7.0 kcal/mol). SHAP (SHapley Additive exPlanations) values summarize the contribution of each molecular descriptor to the classification of compounds as active or inactive against BUB1. Each dot represents a SHAP value for a given compound and feature. The horizontal axis indicates the impact on the model output, while color represents the feature value (red = high, blue = low).

This study presents a robust computational framework with multiple layers for identifying metastasis-associated gene signatures in non-small cell lung cancer (NSCLC). Merging transcriptomic data from both The Cancer Genome Atlas (TCGA) datasets and Gene Expression Omnibus (GEO), followed by protein–protein interaction (PPI) network construction, dynamical modeling, and virtual compound screening, enabled the identification of a high-confidence list of 858 genes with metastasis signature. Unlike previous studies, which rely on single-cohort or platform-specific analyses, this combined approach takes advantage of multi-cohort and cross-platform validation. This method arguably strengthens both biological relevance and further utility. The transcriptomic integration phase served to refine the candidate gene list by ensuring consistency across NSCLC subtypes and platforms. The initial intersection of LUAD and LUSC transcriptomic datasets excluded subtype-specific variability, reinforcing the generalizability of the metastasis-associated signature across NSCLC. This strategy was further strengthened through validation using GEO-derived gene expression data. By lever-

aging datasets across distinct platforms and clinical contexts, the resulting gene set exhibits strong cross-platform robustness. Such integrative approaches are increasingly recognized for their biomarker discovery power. For example, a research group constructed a multi-gene LUAD prognostic model by integrating TCGA and GEO cohorts, highlighting how transcriptomic convergence improves generalizability and clinical utility.[17] While our approach shares this integrative nature, it differs by highlighting metastasis-specific signatures.

The PPI network of the 858-gene set revealed a scale-free topology enriched in cell cycle and DNA repair regulators. Among these, some genes (e.g., *BRCA1, BUB1*, and *TOP2A*) stood out with high degree centrality and were further validated by being prioritized through nonlinear ordinary differential equation (ODE) modeling. These genes maintained stable, non-zero equilibrium states over simulated timesteps, suggesting that they play a persistent regulatory role in maintaining network homeostasis under dynamic conditions relevant to metastasis. This points toward a form of kinetic resilience (the ability of a node to retain regulatory influence over time despite perturbations) that may explain their influence in NSCLC progression. This interpretation aligns with Ma *et al*., who showed *BRCA1*'s involvement in homologous recombination repair (HRR) and its predictive value in NSCLC prognosis.[18] Additionally, systems-level network analysis was conducted by another study, where they emphasized how central mitotic regulators, like *BUB1* and *TOP2A*, often correspond to key phenotypic drivers in cancer.[19] These studies validate our reliance on static centrality as a biologically meaningful criterion.

To contextualize these hubs functionally, enrichment analysis using Reactome revealed that the identified hub genes were involved in G1/S checkpoint control, mitotic spindle formation, and chromosomal segregation, to name a few. These findings suggest that dysregulation of mitotic fidelity may be a significant, though not exclusive, contributor to NSCLC metastatic potential. A possible explanation could be that compromised checkpoint control permits chromosomal instability, facilitating malignant transformation. This interpretation finds some support in a study that demonstrated that alterations in oxidative stress response and DNA damage repair pathways stratify LUAD subtypes with poor prognosis.[20] Additionally, enrichment of RNA splicing, mitochondrial translation, and proteasome pathways aligns with the findings of a study that identified similar biological programs in LUSC cohorts associated with immune infiltration and poor survival.[21]

With this mechanistic backdrop established, virtual screening revealed several candidate compounds with strong predicted binding affinities toward BUB1, including the known kinase inhibitors sanguinarine and AT-7519. These findings are promising given the existing evidence of *BUB1* overexpression in NSCLC and its role in tumor proliferation. Although direct inhibitors of BUB1 remain limited in clinical use, several BUB1-targeted compounds are currently under preclinical evaluation for their potential to disrupt mitotic fidelity and tumor growth. Related kinase inhibitors have shown efficacy in preclinical studies. This supports the potential of BUB1-targeted compounds as viable anticancer agents. It remains to be tested, however, whether such ligands retain efficacy in complex in vivo tumor microenvironments. The identification of phytochemical ligands like sanguinarine also aligns with prior efforts exploring phytochemical inhibition of mitotic regulators, although direct targeting of BUB1 remains unverified.

To distill these docking outputs into a prioritized list, an XGBoost classifier was trained on molecular descriptors and refined using SHAP value interpretation. The model highlighted XLogP, TPSA, and rotatable bonds as key predictors of binding affinity, confirming established structure–activity relationships in medicinal chemistry. Pashaei *et al*. similarly demonstrated the efficacy of machine learning coupled with network-based analysis for high-throughput drug screening and compound interpretation. While our model exhibited robust feature interpretability, its predictions are necessarily constrained by the resolution of available docking scores. Of the 1,869 compounds screened, 367 were prioritized as high-confidence candidates, providing a tractable shortlist for future validation. Future applications should include experimental validation of predicted targets as well as ligands, for instance, with CRISPR-mediated gene knockout for functional characterization or SPR assay for confirmation of compound binding affinity.

This study acknowledges some limitations that it is primarily computational and relies on *in silico* predictions, so results should be interpreted as hypothesis-generating rather than definitive evidence of biological causality. No wet lab validation was performed, meaning the proposed role of BUB1 and the prioritized natural compounds require experimental confirmation (for example, cell-based assays of migration or invasion and target engagement). In addition, docking scores and the chosen activity threshold are approximations that depend on the docking setup and scoring function, so binding affinity rankings may shift under alternative parameterizations or more rigorous free energy methods.

Conclusion

Overall, this work demonstrates the power of merging transcriptomic integration, network biology, dynamical systems models, and AI-improved virtual screening to identify regulatory hubs alongside therapeutic leads for NSCLC metastasis. Concurrence of biologically reasonable conclusions for an extended set of analytical spaces demonstrates internal validity of the analytical pipeline. Such multidimensional concurrence reinforces confidence in prioritized targets alongside their translational validity. At a high level, our systems-level approach potentially offers a starting point for stabilization of biomarkers alongside target fidelity improvement for molecularly heterogeneous cancers like NSCLC, wherein traditional single-layer analyses are suboptimal.

Acknowledgments

Prachet Patakula would like to acknowledge Aashna Saraf, Founder of CreatED, for providing valuable feedback and guidance throughout the project. I would also like to acknowledge my parents for their unwavering support and motivation throughout.

DOI: 10.36838/IJHSR88.90

References

1. Mehta, A. A.; Pavithran, K.; Nair, P. K.; Vazhoor, V.; Gutjahr, G.; Lakshmi Priya, V. P., Epidemiological and histopathological profile of lung Cancer: Insights from a 15-year cross-sectional study at a tertiary care centre in South India. *Global Epidemiology* **2025**, *9*, 100208.
2. Liao, J.; Li, X.; Gan, Y.; Han, S.; Rong, P.; Wang, W.; Li, W.; Zhou, L., Artificial intelligence assists precision medicine in cancer treatment. *Frontiers in Oncology* **2023**, *Volume 12* - 2022.
3. Weaver, I. N.; Weaver, D. F., Drug design and discovery: translational biomedical science varies among countries. *Clin Transl Sci* **2013**, *6* (5), 409-13.
4. Siegel, R. L.; Miller, K. D.; Fuchs, H. E.; Jemal, A., Cancer Statistics, 2021. *CA: A Cancer Journal for Clinicians* **2021**, *71* (1), 7-33.
5. Collaborators, I. S.-L. D. B. I. C., The burden of cancers and their variations across the states of India: the Global Burden of Disease Study 1990-2016. *Lancet Oncol* **2018**, *19* (10), 1289-1306.
6. Jamil, A.; Kasi, A., Lung Metastasis. In StatPearls, StatPearls Publishing Copyright © 2025, StatPearls Publishing LLC.: Treasure Island (FL) ineligible companies. Disclosure: Anup Kasi declares no relevant financial relationships with ineligible companies. 2025.
7. Chunarkar-Patil, P.; Kaleem, M.; Mishra, R.; Ray, S.; Ahmad, A.; Verma, D.; Bhayye, S.; Dubey, R.; Singh, H. N.; Kumar, S., Anticancer Drug Discovery Based on Natural Products: From Computational Approaches to Clinical Studies. *Biomedicines* **2024**, *12* (1).
8. Bi, W. L.; Hosny, A.; Schabath, M. B.; Giger, M. L.; Birkbak, N. J.; Mehrtash, A.; Allison, T.; Arnaout, O.; Abbosh, C.; Dunn, I. F.; Mak, R. H.; Tamimi, R. M.; Tempany, C. M.; Swanton, C.; Hoffmann, U.; Schwartz, L. H.; Gillies, R. J.; Huang, R. Y.; Aerts, H., Artificial intelligence in cancer imaging: Clinical challenges and applications. *CA Cancer J Clin* **2019**, *69* (2), 127-157.
9. Tomczak, K.; Czerwińska, P.; Wiznerowicz, M., The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol (Pozn)* **2015**, *19* (1a), A68-77.
10. Agarwal, A.; Agrawal, P.; Sharma, A.; Kumar, V.; Mugdal, C.; Dhall, A.; Raghava, G. P. S., A repository of web-based bioinformatics resources developed in India. *bioRxiv* **2020**, 2020.01.21.855627.
11. Siddiqui, A. J.; Jahan, S.; Singh, R.; Saxena, J.; Ashraf, S. A.; Khan, A.; Choudhary, R. K.; Balakrishnan, S.; Badraoui, R.; Bardakci, F.; Adnan, M., Plants in Anticancer Drug Discovery: From Molecular Mechanism to Chemoprevention. *Biomed Res Int* **2022**, *2022* (1), 5425485.
12. Paul, D.; Sanap, G.; Shenoy, S.; Kalyane, D.; Kalia, K.; Tekade, R. K., Artificial intelligence in drug discovery and development. *Drug Discov Today* **2021**, *26* (1), 80-93.
13. Bhinder, B.; Gilvary, C.; Madhukar, N. S.; Elemento, O., Artificial Intelligence in Cancer Research and Precision Medicine. *Cancer Discov* **2021**, *11* (4), 900-915.
14. Serrano, D. R.; Luciano, F. C.; Anaya, B. J.; Ongoren, B.; Kara, A.; Molina, G.; Ramirez, B. I.; Sánchez-Guirales, S. A.; Simon, J. A.; Tomietto, G.; Rapti, C.; Ruiz, H. K.; Rawat, S.; Kumar, D.; Lalatsa, A., Artificial Intelligence (AI) Applications in Drug Discovery and Drug Delivery: Revolutionizing Personalized Medicine. *Pharmaceutics* **2024**, *16* (10).
15. Medina-Franco, J. L., Computational Approaches for the Discovery and Development of Pharmacologically Active Natural Products. *Biomolecules* **2021**, *11* (5).
16. Chen, C.; Wang, J.; Pan, D.; Wang, X.; Xu, Y.; Yan, J.; Wang, L.; Yang, X.; Yang, M.; Liu, G. P., Applications of multi-omics analysis in human diseases. *MedComm (2020)* **2023**, *4* (4), e315.
17. Zheng, W.; Zhou, C.; Xue, Z.; Qiao, L.; Wang, J.; Lu, F., Integrative analysis of a novel signature incorporating metabolism and stemness-related genes for risk stratification and assessing clinical outcomes and therapeutic responses in lung adenocarcinoma. *BMC Cancer* **2025**, *25* (1), 591.
18. Ma, Y.; Huang, J.; He, L.; Du, J.; Liu, L.; Li, X.; Jiao, P.; Wu, X.; Zhou, W.; Xu, X.; Yang, L.; Di, J.; Zhu, C.; Li, L.; Liu, D.; Wang, Z., Evaluation of two algorithms measuring homologous recombination deficiency status in prognostic assessment for treatment-naïve non-small cell lung cancer. *Chin J Cancer Res* **2025**, *37* (3), 352-364.
19. Pashaei, E.; Liu, S.; Li, K.; Zang, Y.; Yang, L.; Lautenschlaeger, T.; Huang, J.; Lu, X.; Wan, J., DiCE: differential centrality-ensemble analysis based on gene expression profiles and protein-protein interaction network. *bioRxiv* **2025**.
20. Rao, W.; Zhang, Q.; Dai, X.; Yang, Y.; Lei, Z.; Kuang, X.; Xiao, H.; Zhu, J.; Xiong, Y.; Wang, D.; Yang, L., A three-subtype prognostic classification based on base excision repair and oxidative stress genes in lung adenocarcinoma and its relationship with tumor microenvironment. *Sci Rep* **2025**, *15* (1), 16647.
21. Chen, Y.; Wang, M., Revealing roles of PANoptosis-related genes in prognosis and molecular subtypes in lung squamous cell carcinoma by integrated bioinformatic analyses and experiments. *Clin Exp Med* **2025**, *25* (1), 154.

Authors

Patakula Patakula is presently a Grade 12 student at Manthan School, passionate about computational biology and mathematical modeling. His research on vaccine decision modelling using game theory has been recognized with the CREST Gold Award and IEOM publication honors. He plans to pursue bioengineering and medicine to advance translational research.

Nirupma Singh is a Bioinformatics Scientist with a doctorate in Biotechnology and Bioinformatics from the University of Delhi, with six years of hands-on research and development experience. Her journey is marked by a robust foundation in Machine Learning and Python, with five years of expertise. Proficient in Linux and cloud servers like AWS. She excels in structural biology, systems biology, protein/gene network analysis, data mining, and computational genomics.