Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/d6dd00045b

Identification of multi-transcriptomic prognostic biomarkers to explore natural therapeutics for lung cancer integrating machine learning

Md Ahad Ali, †*[ab] Hriddhi Sarker, †[ac] Marguba Kamrun,[ad] Humaira Sheikh, [ae] Bilkis Akter Shifa, [af] Siam Ahmed,[a] Tarikul Islam, [g] Sujoy Banik[h] and Neeraj Kumar[i]

Lung cancer remains the leading cause of cancer-related mortality worldwide, underscoring the urgent need for novel therapeutic strategies. Cyclin-dependent kinase 1 (CDK1), a central cell-cycle regulator, has emerged as an oncogenic driver and potential target in lung adenocarcinoma. This study aimed to integrate transcriptomics, machine learning (ML), and advanced *in silico* approaches to identify natural product-derived potential inhibitors targeting CDK1. To identify robust differentially expressed genes, first we analyzed four different datasets (GSE19804, GSE10072, GSE18842, and GSE10799). Protein–protein interaction network and topological analysis highlighted CDK1 as a primary key hub gene (pKHG) enriched in cell-cycle and p53 pathways. Target validation confirmed CDK1 overexpression, prognostic significance, immune infiltration links, and mutation associations. In addition, a collected library of 9667 natural phytocompounds was reduced through ML-based bioactivity (pIC50) prediction targeting pKHG to discover potential lead molecules. Then, the selected top lead molecules were considered for further evaluation *via* molecular docking, molecular dynamics simulations, ADMET analysis, and binding free-energy calculations (MM-GBSA). Among the selected phytochemicals, CID_14218027 (−7.69 kcal mol^{-1}), CID_487089 (−6.80 kcal mol^{-1}), and CID_174880 (−6.70 kcal mol^{-1}) showed the highest binding affinity score (GLIDE_XP score) and stable molecular interactions. Furthermore, MD simulations confirmed the conformational stability of ligand–protein complexes, supporting their potential as CDK1 inhibitors. This integrated omics-to-*in silico* pipeline identifies CDK1 as a robust therapeutic target and highlights natural product-derived inhibitors with favorable pharmacological and physicochemical properties. Therefore, these findings present a viable framework for accelerating precision drug discovery, with experimental validation underway. However, these findings are based solely on computational analyses and require further experimental validation to confirm CDK1 inhibitory activity, anticancer efficacy, and safety.

Received 27th January 2026
Accepted 24th April 2026

DOI: 10.1039/d6dd00045b

rsc.li/digitaldiscovery

1. Introduction

Lung cancer (LC) is the deadliest cancer in the world because it causes more deaths annually than all breast, colon and prostate cancers put together.[1,2] In 2022, LC was responsible for about 1.82 million deaths, representing 18.7% of all cancer fatalities, far exceeding the mortality from colorectal cancer (9.3%) and breast cancer (6.9%). Approximately 2.48 million new LC diagnoses were made that same year, along with an age-standardized

Md. Ahad Ali is a computational chemist and bioinformatics researcher affiliated with Panacea Research Center and the University of Rajshahi, where he works on transcriptomics, machine learning, and structure-based drug discovery.

[a]*Computational Chemistry and Drug Design Division, Panacea Research Center, Rajshahi 6206, Bangladesh*
[b]*Department of Chemistry, University of Rajshahi, Rajshahi 6205, Bangladesh. E-mail: ahad.chembd@gmail.com*
[c]*Department of Biochemistry and Molecular Biology, University of Rajshahi, Rajshahi 6205, Bangladesh*
[d]*Department of Chemistry and Biochemistry, University of Oklahoma Norman, OK 73019, USA*
[e]*Department of Chemistry, Gopalganj Science and Technology University, Gopalganj 8100, Bangladesh*
[f]*Department of Biochemistry and Molecular Biology, University of Dhaka, Dhaka 1000, Bangladesh*
[g]*Department of Chemistry, University of Barishal, Barishal 8254, Bangladesh*
[h]*Department of Computer Science & Engineering, Rajshahi University of Engineering & Technology, Rajshahi, Bangladesh*
[i]*Department of Pharmaceutical Chemistry, Bhupal Nobles' College of Pharmacy, Udaipur 313001, Rajasthan, India*
† These authors contributed equally as first author to this work.

mortality rate of around 16.8 per 100 000, a pattern that tends to rise in countries with higher Human Development Index scores. If today's rates stay unchanged, population aging and growth could push the numbers to about 4.62 million new cases and roughly 3.55 million deaths by 2050, marking increases in total cases and deaths rather than in standardized mortality rates.[2,3] LC is highly lethal due to late-stage diagnosis, inherent molecular heterogeneity, and low sensitivity to the current treatment.[4,5] Though there has been a great deal of development of synthetic FDA-approved drugs, including cytotoxic chemotherapy, molecularly targeted agents, and immune checkpoint inhibitors,[6–8] the overall prognosis of the patients of LC remains challenging. However, drug–drug interactions, off-target toxicity, inadequate potency, and low pharmacokinetics have limited the translation of these insights into effective therapies.[9–11] Though standard synthetic drugs have played a significant role in the treatment of cancer, these drugs are usually restricted by their adverse side effects, toxicities, resistance and high cost of production.[12–15] In particular, conventional chemotherapy is effective but frequently causes clinically meaningful adverse effects (notably myelosuppression and gastrointestinal toxicity), which can reduce quality of life and limit dosing intensity.[16] The number of patients who initially respond eventually develop acquired resistance (*e.g.*, resistance after EGFR-TKI/osimertinib treatment), which drives disease progression and the need for next-line options.[17,18] Overall, these limitations justify exploration of alternative vulnerabilities such as cell-cycle control: cyclin-dependent kinase is a central mitotic kinase/enzyme and is reported to be overexpressed and prognostically relevant in LC. Abnormal patterns in oncogenic signaling pathways, DNA repair mechanisms, cell death, and uncontrolled cell-cycle progression collectively contribute to tumor initiation and progression.[19–22] These molecular abnormalities suggest that cell-cycle regulators, mitotic proteins, and other important modulators represent promising points of therapeutic intervention. Transcriptomics and multiomics studies help to identify significant regulatory genes and signaling networks with clinical relevance. Network pharmacology-based methodologies, including protein–protein interaction (PPI) networking, hub gene identification, transcriptomics factors, and pathway enrichment analysis, are some effective ways to discover the key molecular regulators.[23–26]

On the other hand, natural compounds—to be more specific, medicinal plant derived phytochemicals—have emerged as an attractive alternative to synthetic drug molecules, because their chemical diversity and evolved bioactivity can provide novel scaffolds and mechanisms that differ from current synthetic libraries.[27] Previous studies show that more than 60 percent of the present anticancer agents are natural, including famous drugs such as paclitaxel, camptothecin, or vincristine.[28–30] The IMPATT database is one of the large databases that contain information on 4010 medicinal plants and 17 967 phytochemicals, along with their properties, which are often not well represented in other databases.[31] Thus, this database is a useful source of lung cancer drugs.

Nowadays, to reduce the time and cost, scientists are considering computational screening using the *in silico* methodology prior to evaluating the study through experimental (*in vivo* and *in vitro*) validation. This research developed a holistic computational workflow integrating transcriptomics, machine learning (ML)-based lead screening, molecular docking and a dynamic simulation study to screen a vast set of compounds in a library and develop therapeutic drug molecules with reduced toxicity and enhanced efficacy. Previous studies have used transcriptomic and machine learning (ML) approaches in LUAD primarily to identify potential biomarkers and therapeutic targets or to predict responses to approved and investigational drugs.[32–36] Similarly, several studies have developed ML models using LUAD mRNA and mutation profiles to predict sensitivity to existing targeted and chemotherapeutic agents, achieving good predictive performance across dozens of drugs.[37–40] In contrast to these studies, which mainly focus on biomarker/drug target identification, validation, and response prediction for existing drugs, this study integrates LUAD transcriptomics-based drug target identification with classical ML-based QSAR modelling to predict pIC50 values for plant-derived phytocompounds, followed by docking and molecular dynamics (MD) simulations to prioritize potent natural candidates.

Therefore, this study builds upon transcriptomic information from various publicly available GEO datasets to construct PPI networks, a classical ML-based regression model to predict the bioactivity (pIC50), docking validation integrating different algorithms, evaluation of the binding strength by calculating post-docking MM-GBSA, large-scale MD simulation, and pharmacokinetics analysis to create a comprehensive approach to natural product-based drug discovery for lung cancer treatment. Compared with conventional docking-centered studies, this biologically informed workflow allows both disease-relevant target prioritization and more systematic lead selection from a large natural compound library. Thus, the present work provides a comprehensive and translationally oriented computational strategy for prioritizing CDK1-targeting phytochemicals in lung adenocarcinoma and supports future experimental validation of the identified candidate compounds. A complete guideline for this work is given below in Fig. 1.

2. Materials and methods

2.1. Target identification

In this section, we present a systematic analysis of lung cancer gene expression profiles to uncover differentially expressed genes (DEGs) and highlight the primary key hub gene (pKHG), which plays a central role in understanding the underlying disease mechanisms.

2.1.1. Microarray expression dataset acquisition. We used four publicly available microarray datasets from the Gene Expression Omnibus (GEO) database of National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov/geo/). The selected datasets were GSE19804, GSE10072, GSE18842, and GSE10799, which contain gene expression profiles of lung cancer and corresponding normal tissues. Details of these datasets, including platform and sample distribution, are summarized in Table 1.

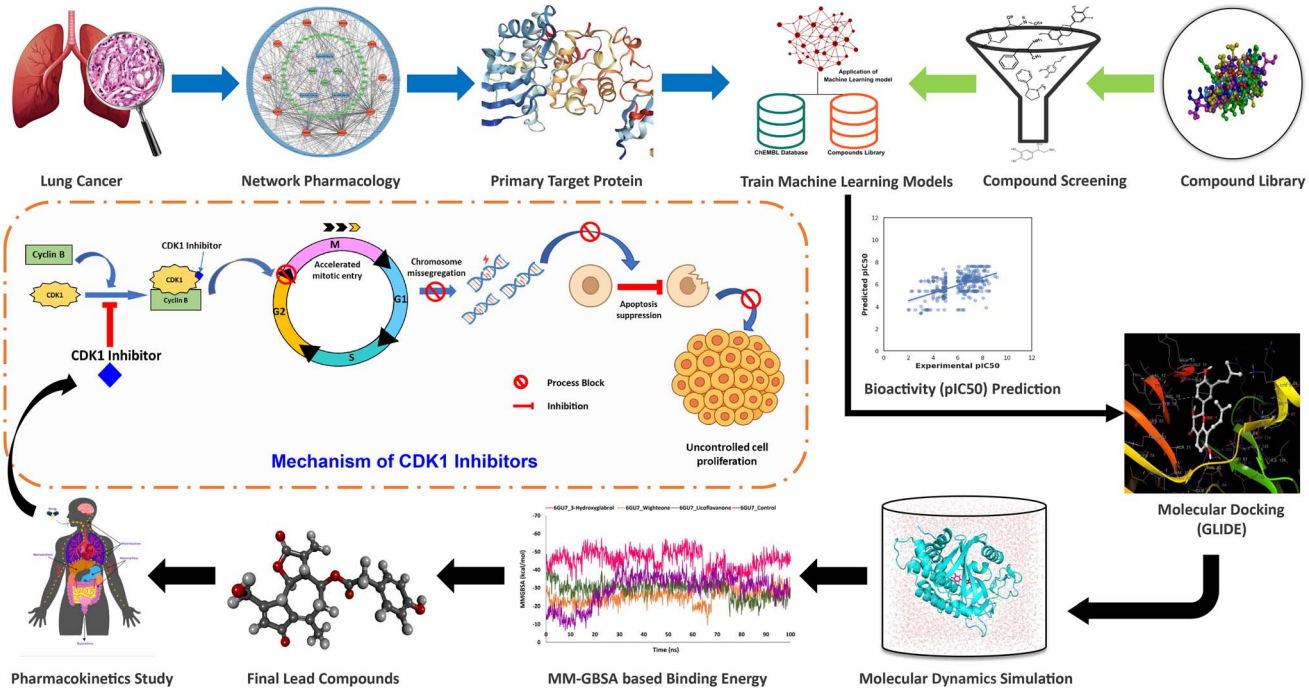

Fig. 1 A complete graphical representation of this study.

Table 1 Summary of lung cancer GEO datasets used in this study

GEO dataset	Number of samples	Cancer	Control	Platform	Ref.
GSE19804	120	60	60	GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array	41,42
GSE10072	107	58	49	GPL96[HG-U133A] Affymetrix Human Genome U133A Array	43
GSE18842	91	46	45	GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array	44
GSE10799	19	16	3	GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array	45

2.1.2. DEG analysis based on microarray gene expression datasets. To identify differentially expressed genes (DEGs) in lung cancer, we analyzed the four selected datasets (GSE19804, GSE10072, GSE18842, and GSE10799) using GEO2R (https://www.ncbi.nlm.nih.gov/geo/geo2r/), an online tool provided by NCBI. In each dataset, samples were categorized into "control" and "cancer" groups, and the data were normalized using log 2 transformation. The limma package[46] was applied to detect DEGs, while the Benjamini–Hochberg false discovery rate (FDR) method was used to adjust for multiple testing. Genes with an adjusted p-value < 0.05 and log FC > +1.0 were considered significantly upregulated, whereas those with log FC < −1.0 were considered significantly downregulated.

2.1.3. Identification of common DEGs (cDEGs). Finally, we identified the common DEGs (cDEGs) by intersecting the DEG lists from all four datasets, representing potential candidate genes associated with lung cancer. The cDEGs were extracted using the "dplyr" package[47] in R, and their overlap across datasets was visualized through Venn diagrams constructed with the R packages "ggplot2"[48] and "ggvenn".[49]

2.1.4. Identification of key hub genes *via* PPI network analysis. Having determined the common DEGs (cDEGs), we investigated the interactivity of the relevant proteins with each other to reveal major molecular targets in lung cancer. The screening of a protein–protein interaction (PPI) network was performed through the STRING database,[50] and only interactions that were experimentally validated were considered. To concentrate the analysis on direct interactions among the cDEGs, we defined the minimum interaction score as low confidence (0.150) and did not add any additional interactors to the initial shell in order to reduce the results of the analysis to direct interactions. It was then visualized in Cytoscape (v3.10.4), with each node being a protein and each edge being the interaction between the two proteins.[51] In order to uncover the most significant proteins, we applied the CytoHubba[52] application and analyzed hub genes in eight diverse topological approaches: Degree, Maximum Neighborhood Component (MNC), Maximal Clique Centrality (MCC),

Density of Maximum Neighborhood Component (DMNC), Edge Percolated Component (EPC), Bottleneck, EcCentricity, and Closeness. Using these several topological analyses, the proteins with the largest percentage of interactions were chosen as the most important key hub genes (KHGs), which guaranteed a high value in identifying the key central target proteins in lung cancer-related networks.

2.1.5. Analysis of transcriptional and post-transcriptional regulation of KHGs. In order to understand the regulatory processes of the identified KHGs at an upstream level, we conducted a combined analysis of regulatory networks to identify the transcriptional and post-transcriptional regulators. The prediction of transcription factors was performed using the TF–target interaction within the JASPAR database[53] and miRNA–target interaction within miRTarBase.[54] The regulatory networks were built with the help of the web platform NetworkAnalyst,[55] (https://www.networkanalyst.ca/) which combines these interactions and enables topological analysis to be performed to discover essential regulators. The networks that resulted were then visualized and analyzed further using Cytoscape (v3.10.3) where nodes are regulators or target genes and edges are interactions that regulate a gene. Significant regulators were also prioritized according to their topological features of networks, thus bringing light to transcription factors and miRNAs that could be at the center of regulating lung cancer-related KHGs.

2.1.6. GO and pathway enrichment analysis of KHGs. We first used the Gene Ontology (GO) and KEGG pathway enrichment analysis through the DAVID database to understand the biological functions of the identified DEGs.[56] GO grouped genes based on molecular functions, biological processes and cellular components, whereas KEGG identified the relevant signaling pathways. For GO enrichment analysis, terms with a minimum gene count of 2 and a p-value of less than 0.05 were considered significant and KEGG pathways were selected with the default EASE score in DAVID. In order to guarantee the reliability of such results, DAVID-enriched GO terms and KEGG pathways were further verified with the help of Enrichr[57] and GeneCloudOmics.[58] The terms found in DAVID and those terms that we regularly found in both DAVID and the other validation platforms were retained, allowing us to obtain a comprehensive and reliable set of functional categories based on which lung cancer KHGs were associated.

2.2. Target validation

To validate the identified prime key hub gene (pKHG) in lung cancer, we focused on the lung adenocarcinoma (LUAD) dataset from the available databases.

2.2.1. Transcriptional and proteomic expression analysis of the pKHG. The pKHG was selected from the candidate hub genes based on a combination of topological scoring and enrichment analysis, ensuring that the most biologically relevant and network-central target was prioritized for downstream validation. To strengthen the reliability of our findings, we next validated the expression pattern of the pKHG in LUAD using multiple publicly available resources. The Tumor Immune Estimation Resource (TIMER 2.0) (https://timer.comp-genomics.org/)[59] was first used to examine the expression of the pKHG between lung adenocarcinoma tissues and normal controls based on TCGA data. For further validation, the GEPIA2 platform (https://gepia2.cancer-pku.cn)[60] was employed, integrating TCGA and GTEx datasets. Boxplot parameters were set as follows: log 2 FC cutoff = 1, p-value cutoff = 0.01, jitter size = 0.4, and values were log 2 (TPM + 1)-transformed for visualization. The "Stage Plot" module of GEPIA2 was also utilized to determine the correlation between the expression of genes and the pathological stages (I–IV) of LUAD. In addition, two levels were applied to the UALCAN database (https://ualcan.path.uab.edu):[61] (i) the study utilized TCGA RNA-seq data to validate the differences in expression of mRNA between the normal and LUAD tissues, and (ii) the study used Clinical Proteomic Tumor Analysis Consortium (CPTAC) data to test the levels of protein expression of the pKHG. This validation was consistent both at the transcriptomic and proteomic levels with this combined approach.

2.2.2. Survival analysis of the pKHG. In order to examine the prognostic role of the pKHG in LUAD, the Kaplan–Meier survival analysis was performed on the GEPIA2 stage. The median cutoff value of 50 percent was used to classify patients in high- and low-expression groups. The overall survival (OS) and disease-free survival (DFS) were measured, and the difference between the two was tested by the use of the log-rank test. The survival plots display time in months on the X-axis and survival probability (%) on the Y-axis, and the dotted line indicates the 95 percent confidence intervals. This discussion presented valuable information regarding the prognostic value of the pKHG in LUAD.

2.2.3. Immune infiltration associated with the pKHG. The correlation between the pKHG and immune cell infiltration on LUAD alone was explored using the module named "Immune-Gene" in TIMER2.0. We concentrated on CD8+ T cells, macrophages and CD4+ T cells to determine the strongest positive and negative correlations. The deconvolution algorithms were applied to give an estimate of the immune cell infiltration, and Spearman correlation with adjustment of tumor purity was used to provide both the correlation coefficient and the significance value. The resulting associations were plotted with heatmaps, which show in which type(s) of immune cells the hub gene in lung cancer is most tightly associated.

2.2.4. Investigation of mutations and alterations in the pKHG. To continue investigating the problem of cancer-related genomic changes, the cBioPortal platform (https://www.cbioportal.org/) was used.[62] The patterns of genetic alteration of the identified prime key hub gene (pKHG) were systematically analyzed using this resource. The analyses of the data in the TCGA PanCancer Atlas studies (including 32 tumor types and 10 957 samples of patients) were performed with the help of the "Query by Gene" module. The resulting "Cancer Type Summary" was a summary of mutation and copy number change types of the pKHG in various cancers. Moreover, the application of the "Mutations" module was used to produce a schematic diagram of the exact mutation sites in the gene.

2.3. pKHG guided *in silico* drug discovery

To explore therapeutic opportunities, we performed structure-based drug discovery on the validated pKHG. This included retrieval and preparation of target protein's 3D crystal structure, ligand collection, virtual screening of ligands, molecular docking, ADMET, molecular dynamics (MD) simulation, MMGBSA, PCA (Principal Component Analysis) and 3D-FEL (Free Energy Landscape).

2.3.1. Retrieval and preparation of target protein. The pKHG crystallographic structure was derived in the RCSB protein database.[63] To enable the subsequent analysis of the protein, all the existing heteroatoms, ligands and water molecules were eliminated with the help of BIOVIA Discovery Studio 2021.[64] The protein structure was then energy-minimized using Swiss-PDB Viewer (spdbv) 4.1.0.[65]

2.3.2. Compound library construction. In order to detect possible inhibitors, we built a universal phytochemical library with the aid of the IMPPAT 3.0 database.[31] We took a total of 33 traditionally used medicinal plants depending on their ethnopharmacological significance, and extracted phytocompounds in various parts of the plants, such as roots, leaves, and seeds. Out of these chosen plants, 9667 phytochemicals were obtained, which were collected together to form a comprehensive compound library to be used in future virtual screening and molecular docking studies.

2.3.3. Physicochemical property-based ligand screening. Virtual screening (VS) is an essential part of contemporary drug discovery, which also provides a computational method of finding the possible bioactive compounds in large chemical libraries.[66] Pharmacokinetic and toxicity-related parameters can help in this screening process with the help of ADMETlab 2.0,[67] a web-based platform. In order to narrow the selection down, the Rule of Five (RO5) by Lipinski[68] is often used as a criterion in determining drug-likeness. Under this rule, the compound has a higher chance of being orally bioavailable; therefore, it has no more than five hydrogen bond donors, no more than ten hydrogen bond acceptors, a molecular weight less than 500 Da, and the logP less than five. Phytochemicals that fulfil this requirement are said to be good leads for further computational and experimental studies. Duplicate compounds were eliminated, and those that survived were those which had 3D structure available and would be used in further studies.

2.3.4. ML-based bioactivity prediction (pIC50) of the selected compounds. The potent inhibitory concentration (pIC50) value serves as an important indicator of a drug's potency, showing the concentration needed to reduce the activity of a biological target by half.[69] Over time, many studies have worked to improve these prediction strategies, making it easier to focus on the most promising drug candidates.[70–72] In this context, the present study applied machine learning (ML) techniques to forecast the bioactivity of the selected compounds against the pKHG.

2.3.4.1. Dataset curation and preparation. For dataset construction, bioactivity records and chemical structures were obtained from the ChEMBL database, a well-established platform for QSAR investigations. Compounds were then curated by retaining only those entries that reported IC50 values against pKHG. For selecting the final ChEMBL database, we mainly focused on the types of protein, organism, availability of the compound activities, including standard type (IC50) and standard unit (nm). This careful refinement produced a dataset of reliable and biologically relevant molecules, which served as the basis for regression modeling and subsequent analyses.

2.3.4.2. Molecular descriptor calculations. Molecular descriptors play a central role in QSAR modeling because they translate the structural and physicochemical features of compounds into measurable values, allowing meaningful patterns to be recognized.[73] In this study, RDKit was used to calculate Morgan fingerprints and ECFP4 descriptors for each compound. These descriptors offered a detailed numerical profile of the compounds' structural characteristics, which served as essential inputs for building predictive models to estimate bioactivity in QSAR modeling.

2.3.4.3. Data splitting into train and test. The final dataset of 987 compounds was split into training and test subsets using the train_test_split function from scikit-learn with stratification on the class labels to preserve the original class distribution in both subsets. It was split into 789 molecules in the training set and 198 molecules in the test set, typical among cheminformatics and other machine learning tasks to use 20–30% of the data to test the algorithm. Random or stratified random splits are widely used as a baseline in cheminformatics modeling, although recent work has highlighted that more stringent splitting strategies (*e.g.*, scaffold or clustering-based splits) may provide more realistic assessments of external generalization for chemical datasets.[74,75]

2.3.4.4. ML Model development and validation. For predictive modeling, we selected the best regression model, with a high-performance machine learning algorithm valued for its speed, accuracy, interpretability, and the coefficient of determination (R^2) values of test and train datasets. To enhance the model's effectiveness, Recursive Feature Elimination (RFE) was applied to remove less informative features, reducing noise and making the model more interpretable. Here, we used optuna to tune the hyperparameter. The predictive performance of the model was assessed using widely adopted metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and coefficient of determination (R^2). These metrics were calculated according to previously published formulae.

Mean Absolute Error (MAE):

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

Mean Squared Error (MSE):

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

Root Mean Squared Error (RMSE):

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

Coefficient of determination (R^2):

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

Here, y_i refers to the experimentally observed IC50 values, $\hat{y}_i$ represents the predicted values generated by the model, and $\bar{y}$ indicates the mean of the observed IC50 values.

2.3.4.5. Model implementation for bioactivity (pIC50) predictions. After model development, we applied our trained and fine-tuned model to predict the pIC50 values of the collected compounds against the pKHG. Compounds showing the highest predicted pIC50 values (>6.5), reflecting greater potential potency, were considered for further investigation. These selected compounds were then analyzed through molecular docking studies to evaluate their binding affinities and interactions with the pKHG.

2.3.5. Molecular docking *via* AutoDock Vina. Molecular docking is important in the discovery and optimization of possible drug candidates. It can be used to estimate the possibility of two molecules, such as a protein and a ligand, interacting with each other and can provide information on their binding behavior and possible efficacy.[76] Therefore, in this study PyRx was used[77] for molecular docking of the chosen compounds with the target protein. PyRx offers an easy-to-use graphical interface with AutoDock Vina to make docking a lot easier. It further combines AutoDock Vina and Open Babel, which provides a complete package of molecular docking and analysis.[78] Lastly, we analyzed the interaction between the ligands and the target receptor through PyMOL and BIOVIA Discovery Studio 2021, which enabled us to see the interaction and to have a better perspective on their binding patterns.

2.3.6. Molecular docking validation *via* Schrödinger software. Following the first docking in AutoDock Vina, the shortlisted compounds were further filtered based on their highest binding affinity and were selected for further docking re-scoring using the GLIDE module. First, these potential hit compounds were generated and optimized with the LigPrep module of Schrodinger suites.[79] The protein structure was preprocessed, and the grid was generated through the protein preparation wizard and receptor grid generation module, respectively. This involved the correction of missing hydrogen atoms, the closure of side-chain gaps and loops using Prime, the elimination of water molecules that are distant to the active site and the creation of suitable protonation states using Epik at a physiological pH of 7.0 $\pm$ 2.0. Lastly, the validated ligands were re-docked into the active sites of the target proteins using the GLIDE module of the Maestro (Version 11.8.012) software in Extra Precision (XP) mode, ensuring that only reliable targets were studied in more detail in terms of their binding interactions.

2.3.7. Post docking MMGBSA. To gain a better idea of ligand–protein interactions, post-docking refinement methods, especially Molecular Mechanics Generalized Born Surface Area (MM-GBSA), were used. Although docking scores provide an initial value of binding affinity, they do not include all of the solvation effects or entropic contributions, which may affect the binding stability and strength of ligand binding.[80] MM-GBSA overcomes these drawbacks by offering a more accurate estimate of the binding free energy (ΔG), which provides a more accurate image of the affinity between the ligand and receptor.[81] In this research, the MM-GBSA analysis was only applied to the ligands that yielded good results in extra precision (XP) mode, using the OPLS4 force field in the Schrodinger suite (PRIME module). This selective approach was necessary because not all ligands produce reliable or meaningful scores during XP docking due to differences in binding modes and structural flexibility.[82] The binding free energy for each ligand was calculated using the formula:

$$\Delta G_{\text{bind}} = \Delta G_{\text{complex}} - (\Delta G_{\text{receptor}} + \Delta G_{\text{ligand}})$$

Here, $\Delta G_{\text{complex}}$ represents the free energy of the ligand–protein complex, $\Delta G_{\text{receptor}}$ is the receptor's free energy, and ΔG_{ligand} is the ligand's free energy. More negative ΔG values correspond to stronger binding interactions.[83]

2.3.8. Molecular dynamics simulation and post simulation MM-GBSA calculation. The molecular dynamics (MD) simulations were performed using the Desmond module of Schrödinger to examine the stability and conformational dynamics of the protein–ligand complexes.[84] System preparation was performed using the System Builder wizard. Each complex was solvated in a simple point charge (SPC) water model inside an orthorhombic simulation box, ensuring a minimum distance of 10 Å between the protein surface and the box edges. To emulate physiological ionic conditions, the systems were neutralized and supplemented with 0.15 M Na^+ and Cl^- ions. Energy minimization was conducted for 100 ps to remove unfavorable contacts, followed by equilibration under NVT and NPT ensembles at 300 K and 1 atm. The production MD simulations were conducted for 100 ns with a 2 fs time step. Trajectories were saved at 20 ps intervals, yielding 5000 frames in total. We used the OPLS3e force field,[85] which delivers improved parameterization for biomolecules as well as drug-like compounds. The simulation trajectories were analyzed for RMSD, RMSF, SASA, radius of gyration, hydrogen bonding, and principal component dynamics to assess conformational stability.

Then the MM-GBSA-based binding free energy (ΔG_bind) of the selected complex was calculated with the gmx_MMPBSA package.[86,87] This approach integrates van der Waals, electrostatic, and solvation energy components, along with solvent-accessible surface area (SASA) contributions, providing a thermodynamic estimate of ligand affinity. The Desmond trajectory files were converted to GROMACS-compatible formats using Schrödinger utilities, while the corresponding topology files were generated through InterMol conversion of *.cms files to *.gro and *.top formats.[88] The free energy associated with

binding (ΔG_bind) was determined based on the following relationship:

$$\Delta G_{\text{bind}} = \langle G_{\text{PL}} \rangle - \langle G_{\text{P}} \rangle - \langle G_{\text{L}} \rangle$$

In this equation, $\langle G_{\text{PL}} \rangle$, $\langle G_{\text{P}} \rangle$, and $\langle G_{\text{L}} \rangle$ correspond to the average free energies of the complex, the unbound protein, and the free ligand, respectively.

Accordingly, the overall binding energy is given by:

$$\Delta G_{\text{bind}} = \Delta E_{\text{MM}} + \Delta G_{\text{SOLV}} - T\Delta S$$

where ΔE_{MM} represents the gas-phase molecular mechanics energy (including van der Waals and electrostatic components), ΔG_{SOLV} is the change in solvation free energy, and $T\Delta S$ reflects the contribution from entropy.

2.3.9. Pharmacokinetics (ADME & toxicity) evaluation. After completing molecular docking and simulation studies, it became essential to examine whether the top-performing compounds possessed properties suitable for real-world drug development. The pharmacokinetic analysis has been performed based on the SwissADME web tool (https://www.swissadme.ch/), which predicts the important ADME parameters (absorption, distribution, metabolism, and excretion) and the physicochemical properties including solubility, lipophilicity, and molecular flexibility.[89] In this analysis, short listing was done on compounds that had an optimal balance of potency and pharmacokinetic feasibility.

In order to supplement these results, the ProTox-III server (https://tox-new.charite.de/protox_III) was employed to make predictions regarding different types of toxicity, such as hepatotoxicity, nephrotoxicity, cardiotoxicity, and neurotoxicity.[90] It uses deep machine learning algorithms to train on experimental data to give confidence in estimating toxicity.

3. Results

3.1. Target identification

3.1.1. Differential gene expression analysis and DEG selection. In order to examine the transcriptional changes in lung cancer, four independent GEO microarray datasets were analyzed (GSE19804, GSE10072, GSE18842 and GSE10799). We found extensive transcriptional changes in lung cancer *versus* normal tissues, as many of the genes were significantly upregulated or downregulated in all the four datasets. Table S1 contains the lists of these genes for each dataset. DEG visualization was performed as a volcano plot (Fig. 2A), revealing obvious differences in significantly upregulated and downregulated genes in the state of lung cancer and normal tissues. As a step to find strong molecular signatures, we then identified common DEGs among the four datasets. The intersection analysis showed that there are common DEGs (cDEGs), which were integrated in a Venn diagram (Fig. 2B). Table S2 gives the detailed list of these cDEGs. These shared genes are the possible candidates that can be important to the pathogenesis of lung cancer and they were prioritized to be further analyzed with respect to their functions.

3.1.2. Identification of key hub proteins *via* PPI network analysis. Having built the protein–protein interaction (PPI) network using the common DEGs (cDEGs), we obtained the network having 375 nodes and 349 edges with an average node degree of 1.86 and an average local clustering coefficient of 0.299. The expected number of edges was 222, and the PPI enrichment p-value was 2.55×10^{-15}, which meant that the interactions observed were considerably more than should have occurred by chance. All of these PPI networks are depicted in Fig. 3A, in which nodes correspond to individual proteins and edges represent experimentally confirmed interactions.

To identify the most influential proteins within the network, we applied eight topological methods using the CytoHubba plugin. For each method, the top 10 hub genes were determined. The results of these analyses are summarized in Fig. 3B, highlighting the proteins that consistently appear as central hubs across multiple metrics. These key hub genes (kHGs) are likely to play critical roles in lung cancer biology and represent promising targets for further investigation. Among all the identified KHGs, CDK1 exhibited the highest number of protein–protein connections, and therefore its interaction network is highlighted in Fig. 3A.

3.1.3. Analysis of transcriptional and post-transcriptional regulation of kHGs. After identifying hub genes using multiple topological algorithms in the PPI network (eight different methods), we observed that although several hub genes were common across most methods, some unique genes were identified by individual algorithms. Each centrality method emphasizes different network properties and therefore captures complementary aspects of network biology.[91,92] To minimize the bias associated with any single algorithm and to ensure that potentially relevant candidates were not excluded, we considered the union set of hub genes ($n = 26$) obtained from all methods. To better understand how the identified kHGs are regulated, we carried out an integrated analysis focusing on their interactions with transcription factors (TFs) and microRNAs (miRNAs). As illustrated in Fig. 4A and Fig. 4B, the networks highlight kHG–TF and kHG–miRNA interactions, respectively. From the network's topological evaluation, we pinpointed five key TFs (FOXC1, GATA2, YY1, E2F1, and HINFP) along with five major miRNAs (hsa-miR-192-5p, hsa-miR-92a-3p, hsa-miR-193b-3p, hsa-miR-215-5p, and hsa-miR-155-5p) as central regulators. These molecules appear to play crucial roles in controlling gene expression at both the transcriptional and post-transcriptional levels.

3.1.4. GO and pathway enrichment analysis of KHGs. In order to shed light on the biological significance of the identified key hub genes (KHGs) in lung cancer, the enrichment analysis of Gene Ontology (GO) and KEGG pathways was conducted in DAVID, EnrichR, and GeneCloudOmics. All of the terms detected in DAVID and those that were present throughout the validation platforms were kept, which led to a full range of functional categories of the KHGs of lung cancer. The enriched categories were classified into biological processes (BP), cellular components (CC), molecular functions

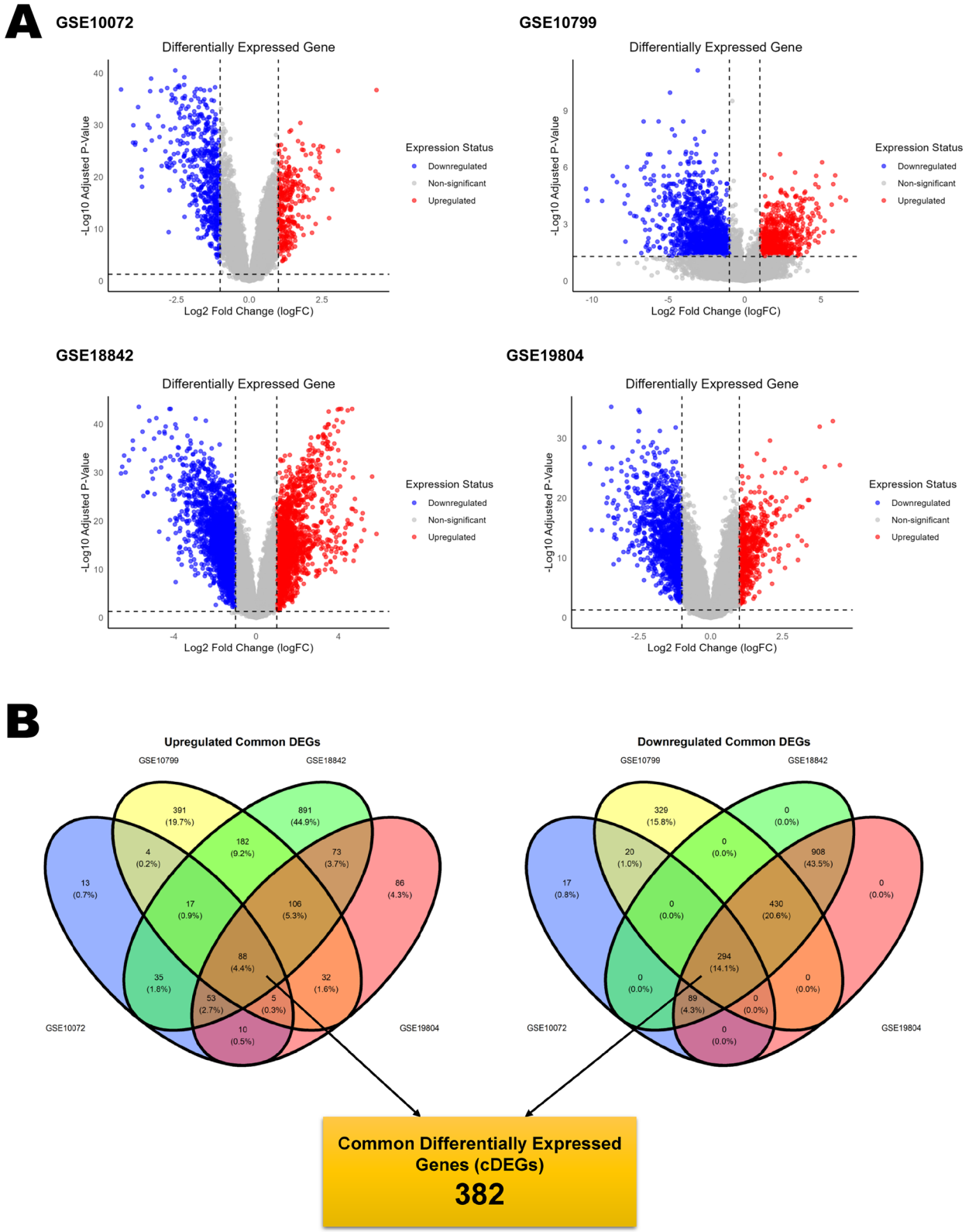

Fig. 2 (A) Volcano plots illustrating the distribution of significantly upregulated and downregulated genes across each dataset. (B) Venn diagram showing the overlap of DEGs among the four datasets and the common DEGs (cDEGs).

(MF), and signaling pathways, and the corresponding detailed lists are presented in the SI (Table S3). In GO categories, enrichment of functions related to the cell cycle was observed to be strong. Mitotic cell cycle progression, spindle organization, chromosome segregation, and checkpoint signalling terms were highly enriched in biological processes (BP), and AURKA,

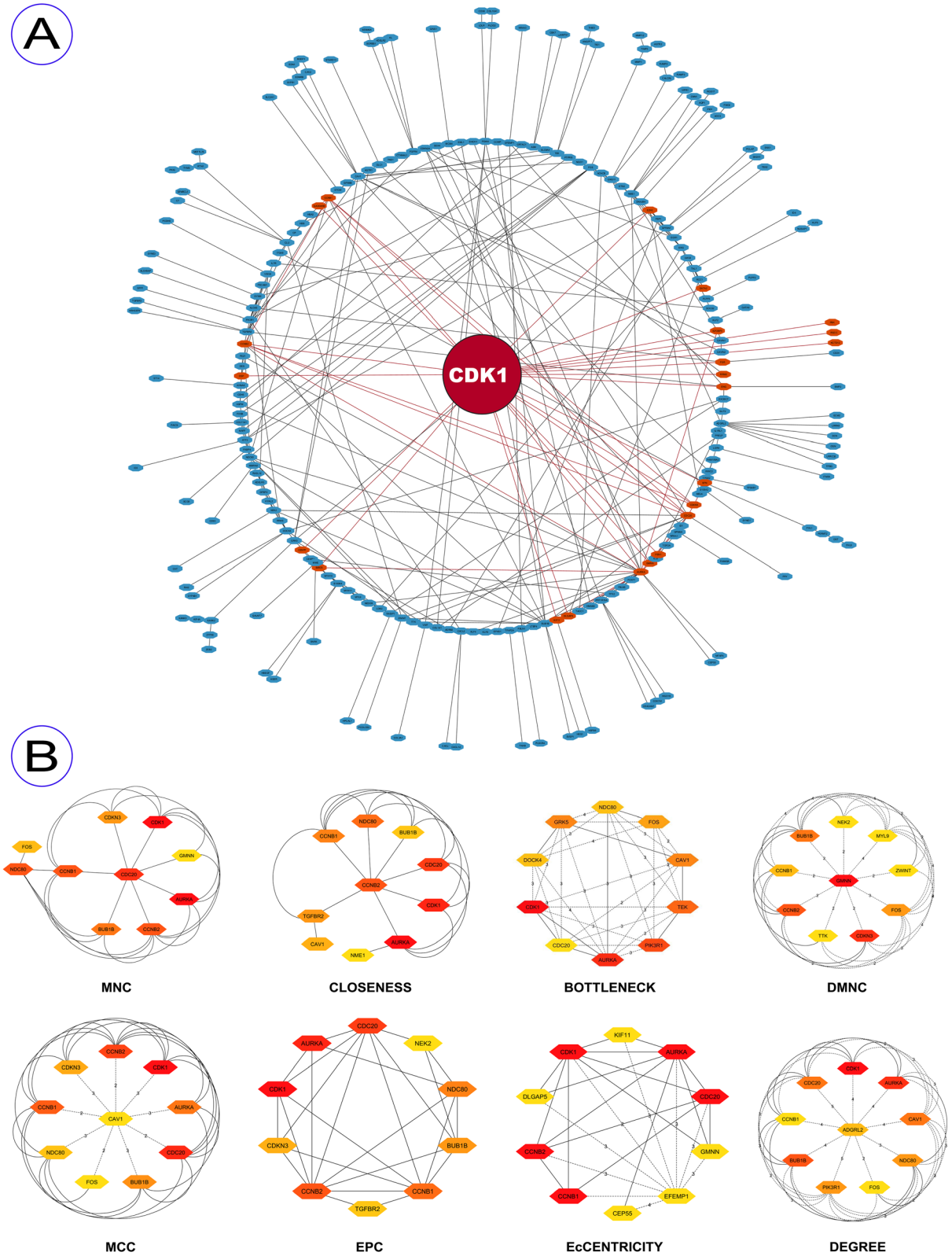

Fig. 3 Protein–protein interaction (PPI) network of common DEGs (cDEGs) in lung cancer. (A) Overall PPI network showing nodes as proteins and edges as experimentally validated interactions. (B) Top 10 hub genes identified using eight different topological methods (Degree, MNC, MCC, DMNC, EPC, Bottleneck, EcCentricity, and Closeness).

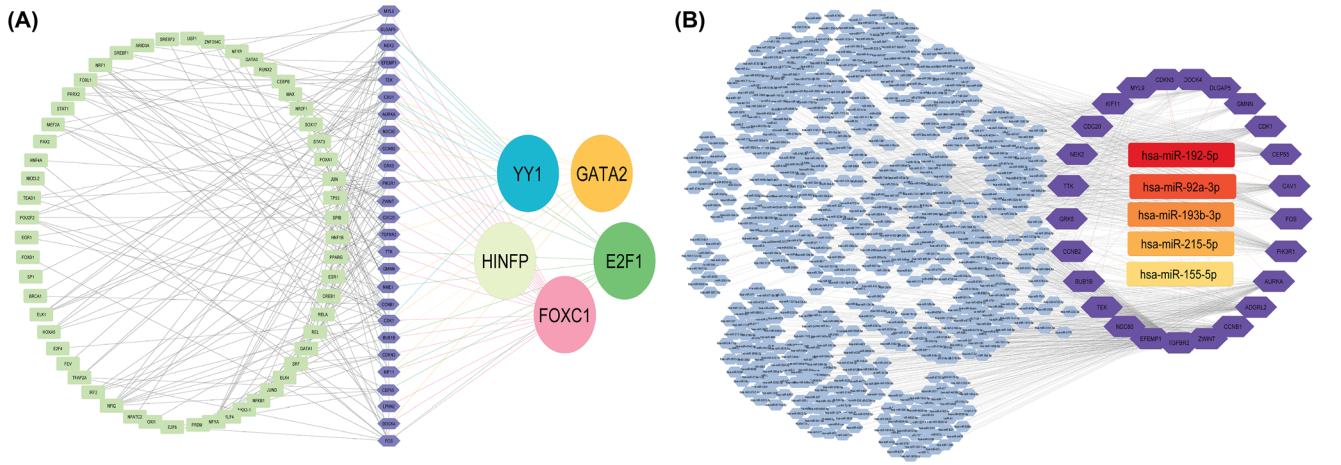

Fig. 4 Integrated regulatory network of DEGs with transcription factors (A) and microRNAs (B).

CDK1, CCNB1, NDC80 and CDC20 among others have been repeatedly involved. These results indicate that the destabilization of mitotic fidelity is one of the key markers in lung cancer progression. KHGs in the cellular components (CC) group were enriched to known mitotic structures, *e.g.* kinetochore, centrosome, spindle pole, and mitotic spindle, and KHGs such as NDC80, BUB1B and AURKA were identified as major drivers of these enrichments. This indicates that there is structural dysregulation of chromosome segregation machinery in the lung cancer cells. In molecular functions (MF), the representation of kinase activities (protein kinase and serine/threonine kinase) and binding ATP functions was most enriched, with AURKA, CDK1, NEK2, and TTK occupying central positions. Multiple histone kinase activities also arose, and these hub genes might play a role in mitosis-related epigenetic regulation. In accordance with the GO results, KEGG pathway analysis showed 14 highly enriched pathways, with the strongest association made by the cell cycle pathway ($p = 3.83 \times 10^{-7}$) with multiple KHGs (CDK1, CCNB1, CCNB2, BUB1B, CDC20, NDC80, and TTK). Pathways connected to tumor suppressive functions and cellular stress reactions, including p53 signaling and FoxO signaling, were additionally greatly enriched and include these genes as part of pathways that regulate apoptosis, senescence, and genomic stability. Interestingly, a number of infection-related pathways (*e.g.*, HTLV-1, HIV-1 and Hepatitis B) were also enriched, implying that there are common molecular pathways between virus infection and oncogenesis that could play a role in the pathology of lung cancer.

3.2. Target validation

Based on the integrated evidence from hub gene ranking within the PPI network and the enrichment analysis highlighting its strong involvement in significant pathways, CDK1 was prioritized as the prime key hub gene (pKHG) for downstream validation.

3.2.1. Evaluation of transcriptional and proteomic expression levels of CDK1 in lung cancer. To validate CDK1 as the prime key hub gene, we first assessed its expression profile across pan-cancer datasets using TIMER2.0. As shown in Fig. 5A, CDK1 expression was significantly elevated in multiple tumor types, including lung adenocarcinoma (LUAD), with $p < 0.001$ indicating strong statistical significance. Further validation in LUAD using GEPIA2 confirmed the upregulation of CDK1. The boxplot analysis (Fig. 5B-i) demonstrated a significant increase in CDK1 transcript levels in LUAD tissues compared to normal controls, while the stage plot (Fig. 5B-ii) revealed that CDK1 expression remained significantly high across pathological stages (I–IV) ($p = 0.000579$). These findings suggest that CDK1 overexpression is maintained throughout disease progression. In addition, UALCAN-based analyses provided consistent evidence at both transcriptomic and proteomic levels. TCGA RNA-seq data showed markedly higher CDK1 expression in LUAD tumors compared with normal samples (Fig. 5C-i), and CPTAC proteomic data similarly confirmed elevated protein expression of CDK1 in tumor tissues relative to controls (Fig. 5C-ii). Collectively, these results demonstrate that CDK1 is robustly upregulated in LUAD at both mRNA and protein levels, reinforcing its candidacy as a biologically relevant prime key hub gene for downstream validation.

3.2.2. Survival analysis of CDK1 in lung cancer. To explore the clinical relevance of CDK1 expression in LUAD, we performed Kaplan–Meier survival analysis using the GEPIA2 platform. Patients were stratified into high- and low-expression groups based on the median cutoff value. In the overall survival (OS) analysis, patients with high CDK1 expression showed a markedly poorer outcome compared to those with low expression (Fig. 6A). The difference was statistically significant (log-rank $p = 2.6 \times 10^{-5}$), with a hazard ratio (HR) of 1.9, suggesting that elevated CDK1 expression nearly doubled the risk of death in LUAD. Consistent with this, the disease-free survival (DFS) analysis also indicated that patients with higher CDK1 levels experienced earlier recurrence and shorter DFS than those in the low-expression group (Fig. 6B). The association was significant (log-rank $p = 0.027$; HR = 1.4), reinforcing the unfavorable role of CDK1 overexpression in disease progression. Together, these findings highlight CDK1 as a potential

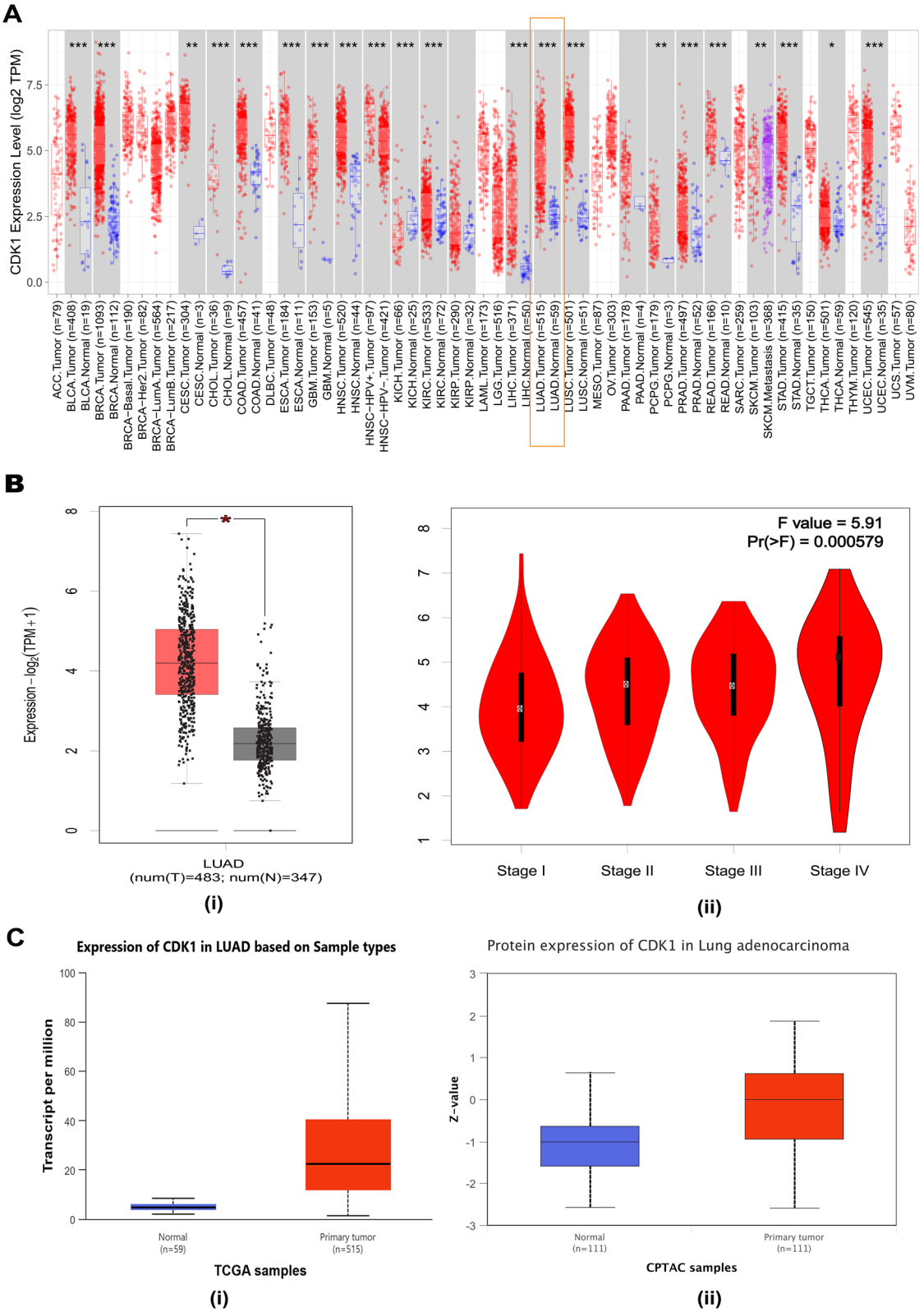

Fig. 5 Transcriptional and proteomic expression profiles of CDK1 in lung adenocarcinoma (LUAD). (A) Pan-cancer analysis of CDK1 expression levels across TCGA datasets using TIMER2.0, showing significantly higher expression in LUAD (***$p < 0.001$). (B) Validation of CDK1 expression in LUAD using GEPIA2: (i) boxplot analysis of tumor *versus* normal tissues and (ii) stage plot showing expression across pathological stages ($p = 0.000579$). (C) UALCAN-based validation of CDK1 in LUAD: (i) TCGA RNA-seq data confirming higher transcript levels in tumors and (ii) CPTAC proteomic data showing elevated protein levels in primary tumors compared to normal samples.

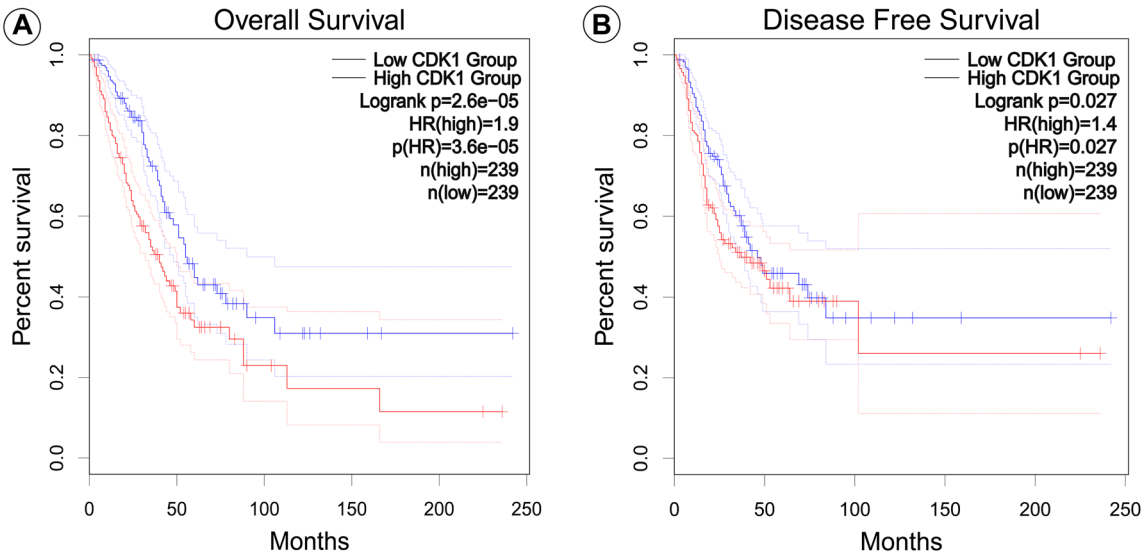

Fig. 6 Kaplan–Meier survival analysis of LUAD patients stratified by CDK1 expression. (A) Overall survival (OS). (B) Disease-free survival (DFS).

prognostic marker in LUAD, where its overexpression is linked to both reduced survival and increased risk of relapse.

3.2.3. Evaluation of immune infiltration associated with CDK1. Using TCGA information, we examined the connection between the expression of CDK1 and the immune cell infiltration in LUAD. Data of CD8+ T cell infiltration analysis in LUAD indicated that the levels of CDK1 expression and T cell CD8+ infiltration had a significant positive correlation as determined by various deconvolution algorithms. A clear band in the heatmap (Fig. S1A) showed the LUAD, and this band showed a significant association. The best positive relationship was found with T cell CD8+ (naïve_XCELL) infiltration with a Rho of 0.235 and a p-value of 1.35×10^{-7}. On the other hand, the most significant negative relationship was observed with T cell CD8+ (EPIC), with a Rho of -0.207 and a p-value of 3.46×10^{-6}. In CD4+ T cell infiltration in LUAD, the heatmap (Fig. S1B) once again was showing an orange band that was representative of LUAD-specific associations. The most significant correlation was found with T cell CD4+ (Th2XCELL) with a Rho value of 0.813 and a p-value of 2.96×10^{-117}. The most negative correlation was found with T cell CD4+ (central memory_XCELL) with a Rho value of -0.322 and a p-value of 2.27×10^{-13}. When it comes to the macrophage infiltration in LUAD, the heatmap (Fig. S1C) was once again represented by an orange band, pointing to the relevant associations. The best positive correlation was with macrophage (M1_CIBERSORT) infiltration, with a Rho of 0.367 and a p-value of 3.39×10^{-17}. The strongest negative correlation was with macrophages (M2_TIDE), which has a Rho value of -0.346 and a p-value of 2.45×10^{-15}. Combined with the analysis, we have found that there is a strong association between the expression of CDK1 and immune cell infiltration in LUAD. In particular, there were positive and negative correlations between CD8 + and CD4 + T cell subsets, and a significant correlation between CD4 + Th2 cells (a subtype that is a humoral immunity). On the other hand, the negative correlation with the CD4+ central memory cells (long-term immune memory) showed a condition-dependent effect. In addition, the expression of CDK1 was positively associated with M1 macrophage infiltration (pro-inflammatory/anti-tumor phenotype) and negatively correlated with M2 macrophages (immunosuppressive/pro-tumor phenotype). This movement between M1 and M2, also known as macrophage polarization, is the functional re-programming of tumor microenvironment macrophages. Taken together, these results indicate that CDK1 can have an effect on the immune microenvironment of LUAD, and it could have an impact on tumor progression and sensitivity to therapy.

3.2.4. Investigation of mutations and alterations in CDK1. To investigate the genomic landscape of CDK1 in human cancers, we queried the cBioPortal database using data from the TCGA Pan Cancer Atlas. The highest alteration frequency was observed in uterine corpus endometrial carcinoma (UCEC), approaching 8%, predominantly driven by amplification. This was followed by skin cutaneous melanoma at approximately 3%, cholangiocarcinoma and uterine corpus endometrial carcinoma at 2.5–2.8%, and lower rates in breast invasive carcinoma (Fig. S2A). Overall, mutations were the most prevalent alteration type across cohorts, with structural variants (purple), amplifications (red), deep deletions (blue), and multiple alterations (black) occurring less frequently but notably in cancers such as LUAD and BRCA. The presence of these alterations, particularly amplification, positions CDK1 as a significant oncogenic driver in a subset of LUAD patients. This suggests that targeted therapies against CDK1 (such as selective CDK inhibitors) could be a viable therapeutic strategy for those patients whose tumors harbor these specific genetic alterations. Further analysis *via* the "Mutations" module generated a schematic lollipop diagram of CDK1 mutations (Fig. S2B), mapping alterations onto the 297-amino acid protein sequence. Missense mutations (green) dominate, while truncating (black), splice

(brown), and fusion (purple) variants were less common. Detailed mutation distributions are presented in Table S4. These data show that CDK1 can be mutated in LUAD, specifically through missense mutations that are predicted to be functionally damaging. This analysis identified and characterizes three non-synonymous somatic mutations in the CDK1 gene (S178L, I136N, and E57V) within a cohort of lung adenocarcinoma (LUAD) patients from The Cancer Genome Atlas (TCGA). While none were currently annotated as known drivers in major clinical databases (OncoKB and CIViC), *in silico* functional prediction tools suggest that two of these mutations (I136N and E57V) are likely pathogenic. These mutations may represent a candidate biomarker for sensitivity to CDK inhibitors in a subset of LUAD patients.

3.3. Structure-based drug discovery targeting CDK1

3.3.1. Structural preparation and optimization of the target protein. The three-dimensional crystal structure of the CDK1 protein (PDB ID: 6GU7) was obtained from the RCSB Protein Data Bank.[93] Prior to its use in computational experiments, the structure underwent a refinement process as outlined in the methodology section. This preparation involved the removal of heteroatoms and water molecules, followed by energy minimization to ensure a stable and optimized conformation suitable for subsequent molecular modeling analyses.

3.3.2. Compound filtering and selection results. In this study, we selected a total of 33 natural medicinal plants as a source of compounds to identify the potential inhibitors of CDK1 from the IMPPAT database (Table S5). Initially we retrieved almost 9 667 compounds from the database with their IMPPAT IDs and SMILES. The Lipinski's Rule of Five (RO5) results reveal that 4236 compounds had zero (0) violations out of 5786 phytochemicals, indicating favorable drug-likeness profiles. Following the removal of duplicates, 2113 unique compounds were retained from all of that selected medicinal plants (Table S6). These compounds were then selected for further computational analyses.

3.3.3. Bioactivity prediction (pIC50) using the ML approach

3.3.3.1. Model development, optimization and validation. To calculate the bioactivity of our selected compounds targeting CDK1, we choose the ChEMBL dataset (CHEMBL308) containing 3348 experimentally validated inhibitors and 4536 activities against CDK1 with their IC50 information (accessed on November 20, 2025). The aim was to develop a classical regression model for CDK1 inhibitors, so that we can calculate the IC50 and pIC50 values of our compound library based on the trained model. According to our study, we get random forest (RF), LighGBM, XGBoost, and Stacking as the best-performing models. The evaluation metrics indicated that the difference in R^2 value for RF, LightGBM, XGBoost, and Stacking were 0.177, 0.117, 0.241, and 0.214, respectively (Fig. S3A). Thus, based on these findings, we observed that LightGBM showed the lowest difference in R^2 values between the test and train datasets, indicating the best fit for prediction outcomes.

The LightGBM regression model among other machine-learning methods has been receiving considerable attention due to its excellent predictive accuracy, computational efficiency, and the capacity to handle large data volumes.[94,95] As an ensemble of decision trees, LightGBM is well suited to capture nonlinear structure–activity relationships while maintaining fast training on large molecular descriptor and fingerprint sets. The LightGBM model was trained on the training set using optimized hyperparameters, including n_estimators, num_-leaves, max_depth, learning_rate, subsample, colsample_by-tree, min_child_samples, reg_alpha, and reg_lambda. Additionally, the LightGBM model tuned to improve generalization and reduce overfitting. Model performance was then evaluated on the test set using standard QSAR regression metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and coefficient of determination (R^2). These parameters offered a comprehensive assessment of how well the model's predictions aligned with the actual experimental data. The test model demonstrated impressive performance, with MAE, MSE, RMSE, and R^2 values of 0.520, 0.544, 0.737, and 0.758, respectively, shown in Fig. S3A. The overall results suggest that the accuracy is comparatively higher than the others. Therefore, finally we selected the LightGBM model for predicting the bioactivity (pIC50) value of our selected natural compounds.

3.3.3.2. Application to bioactivity calculation and compound screening. Once the model's predictive performance was validated, it was used to virtually screen a curated phytocompound library of 2113 phytocompounds by predicting their pIC50 values using the optimized ML model. A total of 380 compounds showed predicted pIC50 > 6.5 (~IC50 < 1 μM), suggesting potential inhibitory activity against the target and prioritizing them for follow-up validation (see Table S7).[96] As the experimental and predicted values lie close to each other in the scatter plot (Fig. S3B), it indicates that the QSAR model is strong and has excellent predictive accuracy, as most data points closely align along the regression line. These results indicate the extent to which the model is useful in identifying potentially promising molecules that can be subjected to further investigation either with the use of molecular docking or through experimental studies to confirm their utility as a possible drug target.

3.3.4. Molecular docking analysis *via* AutoDock Vina. Out of 380 previously selected phytochemicals, a total of 358 with available 3D SDF structures were extracted from the PubChem database and subjected to molecular docking against the CDK1 protein (6GU7) using AutoDock Vina implemented in PyRx. The docking was performed using a grid box centered at the protein's active site with coordinates $X = 22.0351$, $Y = 12.6323$, and $Z = 4.7655$. Additionally, we also compare the docking binding affinity score with the known inhibitor of CDK1 (FB8/AZD5438) and we set this binding affinity of AZD5438 as our cut-off value (-8.8 kcal mol^{-1}). Finally, we selected 29 compounds for further analysis and evaluation of their binding strength (Table S8).

3.3.5. Molecular docking validation *via* Schrödinger software. For further accuracy assessment, extra precision (XP) docking was carried out using the GLIDE module within the

Table 2 Docking scores (binding affinity) in kcal mol^{-1} of the selected ligands and positive control compound using AutoDock Vina and GLIDE_XP mode

Compound CID	Compounds name	Source	Binding affinity by AutoDock	Binding affinity by GLIDE_XP
115196	Brassinolide	*Camellia sinensis*	−9.2	−7.85
5317764	Glycyrrhisoflavon	*Glycyrrhiza glabra*	−9.5	−7.70
14218027	Licoflavanone	*Glycyrrhiza glabra*	−9.5	−7.69
487089	3-(3,4-Dihydroxyphenyl)-5,7-dihydroxy-6, 8-bis(3-methylbut-2-enyl)chroman-4-one	*Glycyrrhiza glabra*	−9.2	−6.80
174880	Lactupicrin	*Cichorium intybus*	−9.4	−6.70
16747683	AZD5438	—	−8.8	−5.15

Schrödinger's Maestro environment for the selected 29 compounds. From the initially docked compounds, top 5 ligands (CID_115196, CID_5317764, CID_14218027, CID_487089, and CID_174880) exhibiting binding affinities below −9.2 kcal mol^{-1} in AutoDock Vina were shortlisted for this validation step, along with the control compound. Their XP docking results revealed that their binding affinity ranged from −6.80 to −7.85 kcal mol^{-1} (Table 2). In particular, CID_115196, CID_5317764, and CID_14218027 showed the highest Maestro XP docking scores of −7.85, −7.70, and −7.69 kcal mol^{-1}, respectively, indicating the strongest predicted interactions with the target protein among these ligands. Also, CID_487089 (−6.80 kcal mol^{-1}) and CID_174880 (−6.70 kcal mol^{-1}) indicated better binding than the reference compound AZD5438 (−5.15 kcal mol^{-1}). Thus, these docking results suggested that these top 5 compounds could be selected as candidate drug molecules for further computational studies.

3.3.6. Post docking MMGBSA analysis. To evaluate the docking-score-based screening results, we further estimated the binding free energies (ΔG_bind) of the selected protein–ligand complexes using the MM-GBSA approach. The analysis revealed that CID_174880 exhibited the most favorable binding energy (−50.44 kcal mol^{-1}), indicating the highest predicted binding stability among all evaluated compounds. Similarly, CID_487089 also demonstrated strong binding affinity with a ΔG_bind value of −46.78 kcal mol^{-1}. CID_14218027 (−36.28 kcal mol^{-1}) and CID_5317764 (−36.15 kcal mol^{-1}) showed moderate but stable binding energies, comparable to each other. In contrast, CID_115196 exhibited the weakest binding among the tested phytochemicals (−27.81 kcal mol^{-1}). Notably, the reference inhibitor AZD5438 displayed a strong binding free energy of −44.58 kcal mol^{-1}, consistent with its reported inhibitory activity against the target protein. Importantly, several candidate compounds, particularly CID_174880 and CID_487089, showed more favorable MM-GBSA energies than the control, while others such as CID_14218027 and CID_5317764 demonstrated comparable binding profiles. Overall, these findings suggest that the top-ranked compounds possess binding affinities equal to or stronger than the reference inhibitor, supporting their potential as promising candidates for further stability and dynamic analyses (Table 3). Lastly, based on the above findings, we selected top 3 compounds for further analysis to validate their binding strength and stability.

3.3.7. Assessment of molecular interactions between protein and ligands. All docked ligands exhibited stable binding within the CDK1 active site through a combination of hydrogen bonding and hydrophobic interactions, closely resembling the interaction behavior of the control compound AZD5438. The control formed four well-oriented hydrogen bonds (THR14, LYS130, ASN133, and LEU83) along with diverse hydrophobic contacts, establishing a strong benchmark profile. Among the phytochemicals, 6GU7_487089 showed the most competitive interaction pattern, forming multiple hydrogen bonds with key residues such as LYS130, ASP86, LEU83, and ASN133, along with dense hydrophobic interactions involving ILE10, ALA31, VAL64, and LEU135, indicating a highly stable and well-packed binding mode. Similarly, 6GU7_174880 demonstrated strong anchoring through four hydrogen bonds (LEU83, LYS89, ASP86, and GLU81) supported by π–alkyl and alkyl interactions with ALA31, ILE10, and LEU135, although its hydrophobic network was comparatively less extensive than that of 6GU7_487089. In contrast, 6GU7_14218027 exhibited a balanced but slightly weaker profile, with hydrogen bonds

Table 3 Post docking MM-GBSA binding free energies of top docked compounds compared to control (AZD5438)

Compound CID	Compounds name	Post-docking MMGBSA (ΔG_{Bind})
174880	Lactupicrin	−50.44
487089	3-(3,4-Dihydroxyphenyl)-5,7-dihydroxy-6,8-bis(3-methylbut-2-enyl)chroman-4-one	−46.78
14218027	Licoflavanone	−36.28
5317764	Glycyrrhisoflavon	−36.15
115196	Brassinolide	−27.81
16747683	AZD5438	−44.58

primarily centered on LYS33, LYS89, and GLN132 and hydrophobic interactions involving ALA31, VAL18, and PHE80. Overall, while AZD5438 maintains the most diverse interaction pattern, 6GU7_487089 emerges as the closest competitor in terms of binding stability and interaction density, followed by 6GU7_174880, whereas 6GU7_14218027 shows comparatively moderate interaction strength. These findings suggest that certain phytochemicals, particularly 6GU7_487089, have the potential to mimic the binding efficiency of the control ligand within the CDK1 active site. Full interaction details are provided in Table S9, and the corresponding 3D and 2D interaction maps are illustrated in Fig. 7.

3.3.8. Docking protocol validation by re-docking. Docking protocol validation is typically carried out by re-docking the co-crystallized ligand back into the active site of the target protein to check how reliable the docking method is. The accuracy of this process is measured by calculating the root mean square deviation (RMSD) between the experimentally determined ligand position and the re-docked pose. In this study, the co-crystal ligand AZD5438 was re-docked into the binding pocket of CDK1 (PDB ID: 6GU7) using the Maestro platform. The best docking pose showed an RMSD value of 1.360 Å when superimposed with the original ligand conformation (Fig. 8). Since this RMSD value is well below the commonly accepted threshold of ≤2.0 Å, it indicates that the docking protocol can accurately reproduce the experimentally observed binding mode. Overall, this validation confirms that the docking setup is reliable and suitable for further analysis of binding affinities and molecular interactions.

3.3.9. Molecular dynamics simulation. To understand the behavior of all four CDK1 complexes—6GU7_487089, 6GU7_174880, 6GU7_14218027, and 6GU7_ AZD5438 (control)—a 100 ns molecular dynamics (MD) simulation to each of the four complexes was conducted to understand their behavior under conditions similar to the human body. With this simulation, we could track the evolution of the structures with time, and how the ligands remained bound to the protein. In order to have a clear understanding of their stability, we analysed some of the important indicators such as RMSD, RMSF, radius of gyration (R_g), and solvent-accessible surface area (SASA). Besides that, principal component analysis (PCA) and free energy landscape (FEL) mapping were conducted to describe the large-scale motions and found the most stable conformational states that each complex visited throughout the trajectory of 100 ns.

3.3.9.1. Root mean square deviation (RMSD). The root mean square deviation (RMSD) evaluates the structural integrity and conformational shifts in the docked complex,[70] with elevated values signaling reduced stability and lower values reflecting stable system performance.[97] The RMSD analysis showed that the apo protein (6GU7) presented the highest average structural deviation (3.245 Å), which is the highest conformational flexibility that is shown in the absence of a ligand. In contrast, all ligand-bound complexes exhibited markedly lower average RMSD values, indicating enhanced structural stability upon ligand binding. Among the screened compounds, CID_487089 was the compound with the lowest average RMSD (2.446 Å), which was then followed by CID_14218027 (2.470 Å), indicating that the compounds have a strong stabilizing interaction in the binding pocket. CID_174880 showed comparable stability to CID_14218027, with an average RMSD of 2.476 Å. The control ligand showed an average RMSD of 3.002 Å, comparable to the apo form, implying minimal stabilizing influence. Overall, the average RMSD results confirm that CID_487089 forms the most stable complex with 6GU7 during the MD simulation (Fig. 9A).

3.3.9.2. Root mean square fluctuation (RMSF). The residue-level dynamics of the system were examined through RMSF analysis over the equilibrated phase of the simulation. As expected, the apo protein showed the highest average fluctuation (1.571 Å), reflecting its greater structural freedom in the absence of a bound ligand. In contrast, all ligand-bound complexes exhibited noticeably lower average RMSF values, suggesting that ligand interaction contributes to a more stable and ordered residue environment. Among the screened compounds, CID_174880 exhibited the strongest stabilizing effect, as reflected by its lowest average RMSF value (1.181 Å), suggesting tight binding and effective restriction of residue mobility. The control compound showed a moderate decrease in flexibility (1.390 Å), indicating a reasonable but less pronounced stabilizing interaction. Meanwhile, CID_487089 and CID_14218027 displayed intermediate stabilization, with average RMSF values of 1.253 Å and 1.375 Å, respectively, implying balanced binding interactions that reduce flexibility without overly restricting protein dynamics (Fig. 9B).

3.3.9.3. Solvent-accessible surface area (SASA). SASA reflects the solvent-exposed surface of a protein, with lower values generally indicating greater ligand burial and potentially stronger binding.[98] Analysis of the 6GU7 protein complexes revealed notable differences in solvent-accessible surface area (SASA) among the ligands. The 6GU7_487089 complex exhibited the highest average SASA value (148.17 $Å^2$), indicating a relatively larger solvent-exposed surface and suggesting a less compact binding conformation. In contrast, the 6GU7_174880 complex showed the lowest average SASA value (126.15 $Å^2$), implying that the ligand is more deeply buried within the binding pocket and forms a more compact and stable protein–ligand interface. The 6GU7_14218027 (130.14 $Å^2$) and control complex (131.81 $Å^2$) displayed intermediate SASA values, reflecting moderately compact structural arrangements with balanced solvent exposure. Overall, 6GU7_174880 demonstrated the most favorable SASA profile in terms of structural compactness, highlighting its potential as the most effective stabilizing ligand among the tested compounds (Fig. 9C).

3.3.9.4. The radius of gyration (R_g). The radius of gyration (R_g) is a key indicator of protein structural compactness and stability in molecular dynamics simulations. In this study, the control complex (6GU7_Control) exhibited a relatively high average R_g value (4.60 Å), indicating a less compact conformation. Similarly, the 6GU7_487089 complex showed the highest R_g value among the ligand-bound systems (4.69 Å), suggesting comparatively lower compactness and weaker structural stabilization. In contrast, 6GU7_14218027 (4.40 Å) and 6GU7_174880 (4.44 Å) displayed lower average R_g values, reflecting more compact and stable conformations. Among these, 6GU7_14218027 exhibited the lowest R_g value, indicating the

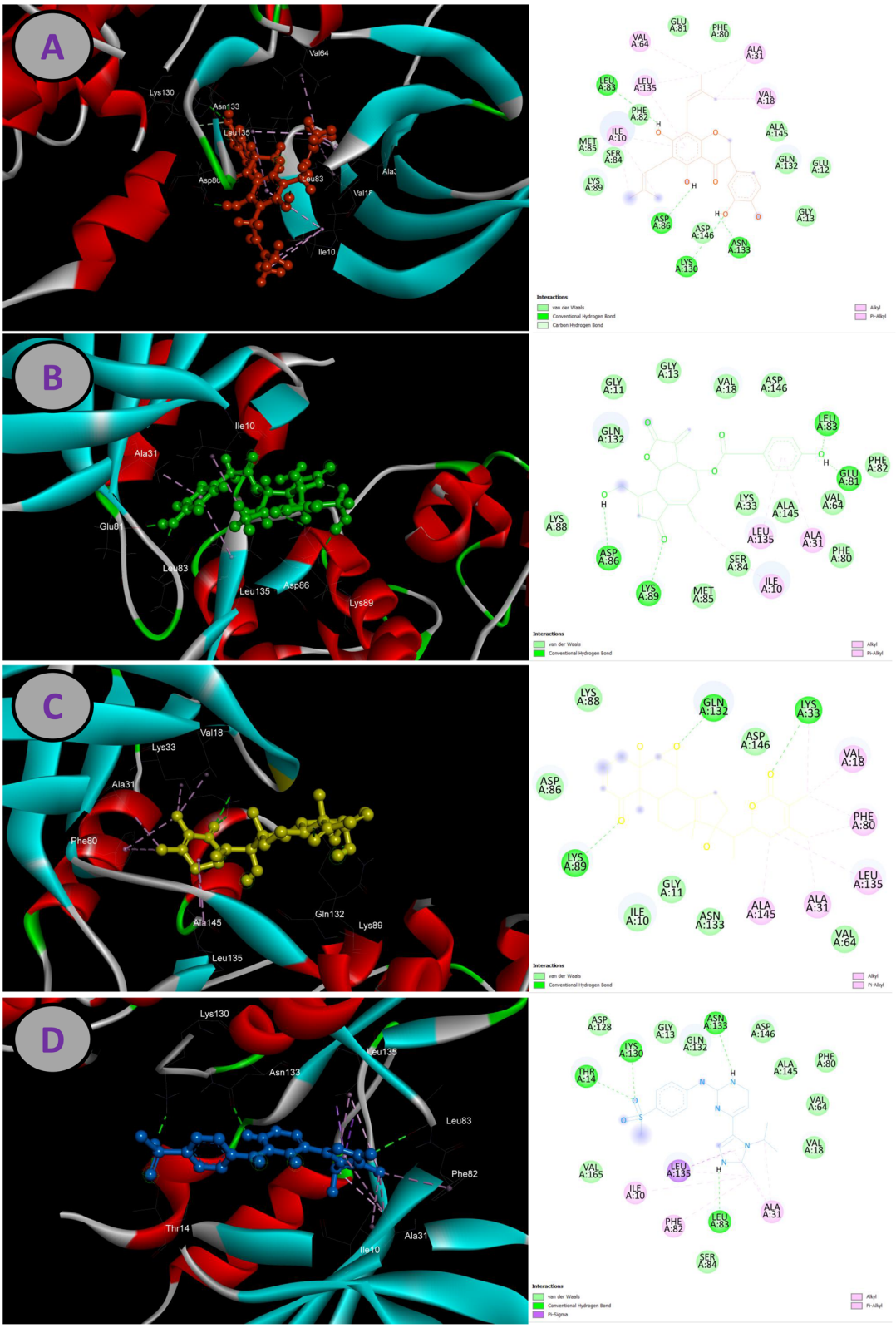

Fig. 7 3D (left) and 2D (right) interaction representations of CDK1 complexes: (A) 6GU7_487089, (B) 6GU7_174880, (C) 6GU7_14218027, and (D) control (AZD5438), highlighting key hydrogen-bonding and hydrophobic interactions within the binding pocket.

highest degree of structural compactness and suggesting a stronger stabilizing effect on the protein. Overall, ligand binding generally enhanced protein compactness compared to the control, with 6GU7_14218027 emerging as the most effective stabilizer based on R_g analysis, followed closely by 6GU7_174880 (Fig. 9D).

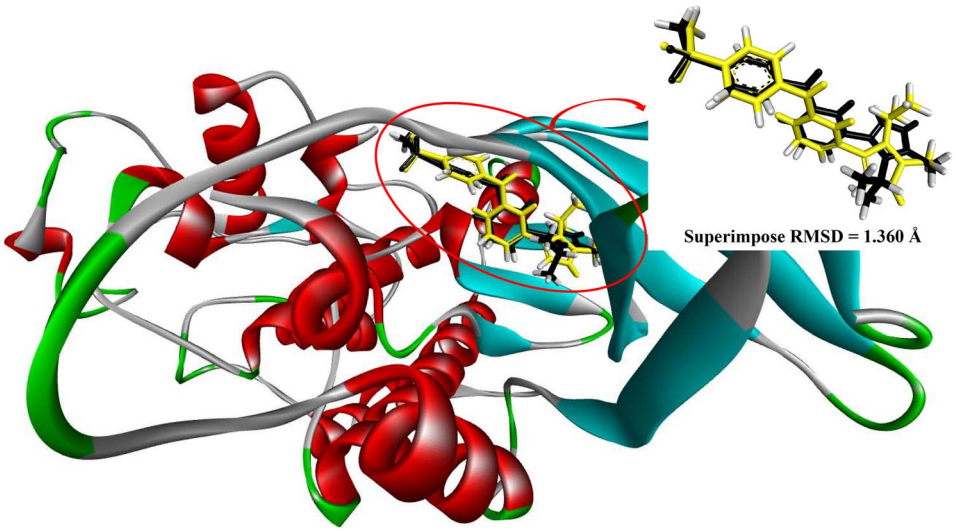

Fig. 8 Graphical illustration of the re-docking procedure and superimposition between the co-crystalize ligand (green) and the re-docked co-crystal ligand (blue) using GLIDE docking.

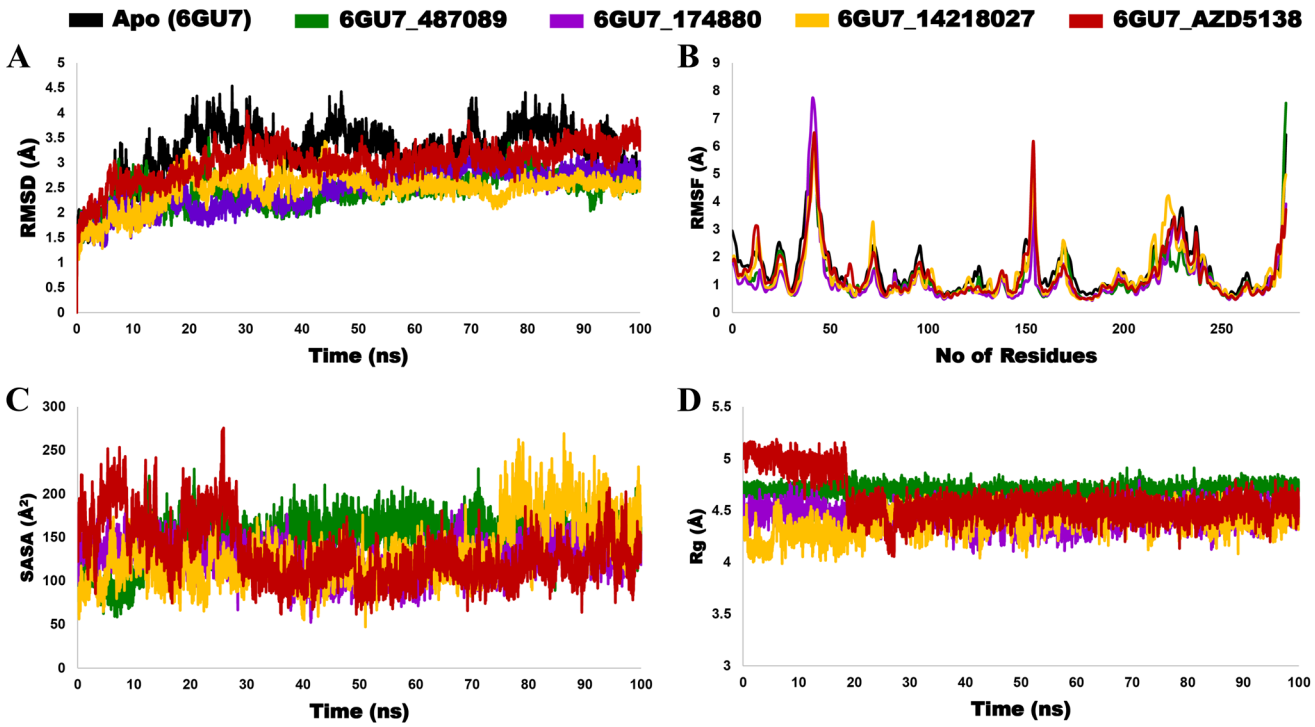

Fig. 9 Molecular dynamics analysis of CDK1 (PDB ID: 6GU7) in complex with three test ligands (6GU7_487089, 6GU7_174880, and 6GU7_14218027) and the reference ligand (6GU7_Control) over 100 ns. (A) RMSD, (B) RMSF, (C) SASA, and (D) R_g profiles.

3.3.10. Post simulation MMGBSA analysis. The post simulation MMGBSA analysis of 6GU7_487089, 6GU7_174880, 6GU7_14218027, and the reference ligand complex (6GU7_AZD5438) against CDK1 (PDB ID: 6GU7) was performed over the 100 ns MD simulation (Fig. 10). The 6GU7_487089 complex exhibited the most favorable binding free energy (-40.29 kcal mol^{-1}), indicating the strongest and most stable interaction with the target protein. This was followed by 6GU7_174880 (-36.06 kcal mol^{-1}), which also demonstrated a relatively strong binding affinity. In contrast, 6GU7_14218027 (-29.13 kcal mol^{-1}) and the control complex (-29.53 kcal mol^{-1}) showed comparatively higher ΔG values, suggesting weaker binding interactions. Notably, the binding affinity of 6GU7_14218027 was slightly lower than that of the control, indicating limited improvement over the reference ligand. Overall, 6GU7_487089 emerged as the most promising

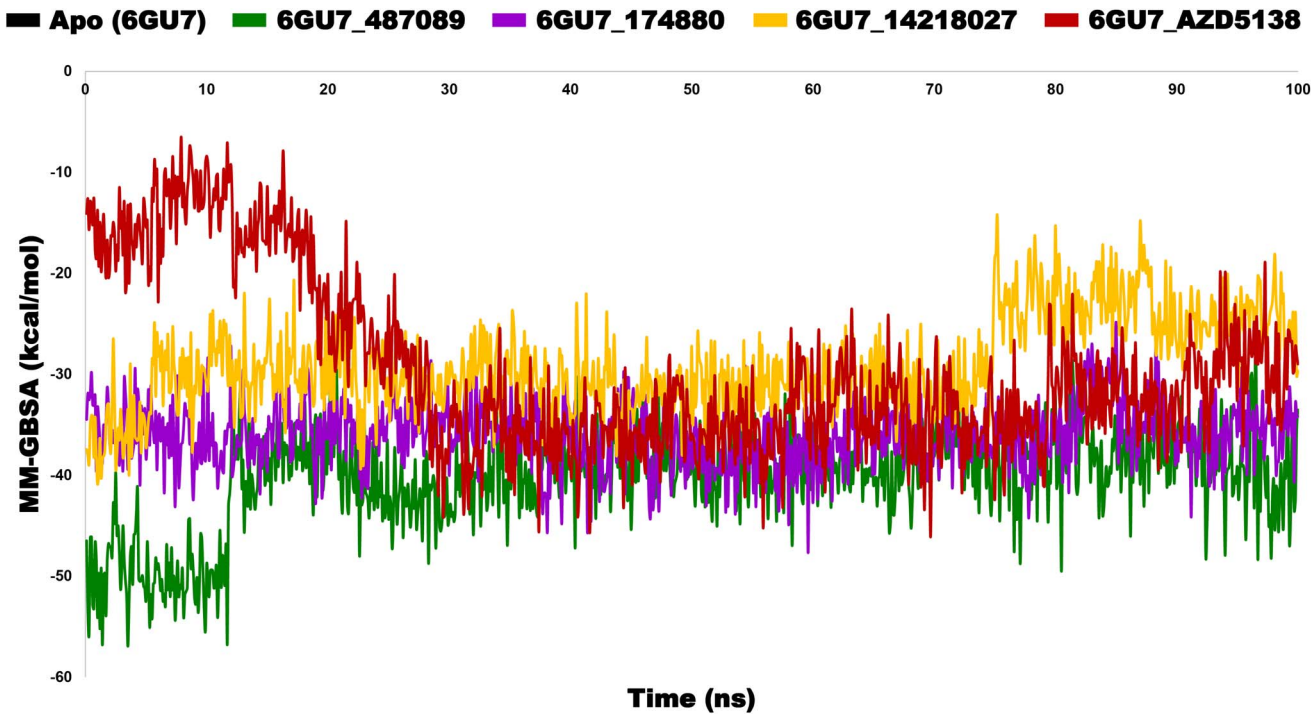

Fig. 10 Comparative MM/GBSA free energy analysis of ligand-bound and control CDK1 systems.

candidate with the highest binding affinity, followed by 6GU7_174880. These findings are consistent with their favorable stability profiles observed in molecular dynamics simulations, further supporting their potential as effective inhibitors.

3.3.11. Principal component analysis. To understand how each ligand influenced the overall motion of the protein, we performed a Principal Component Analysis (PCA). Among the studied complexes, 6GU7_174880 showed the highest contribution in PC1 (44.97%), which was very close to the control complex, 6GU7_AZD5438 (44.18%). This suggests that most of the protein's motion in the presence of ligand CID_174880 is mainly concentrated in one dominant direction. Such a pattern usually indicates a more stable and well-defined conformational movement, similar to that observed for the reference inhibitor. On the other hand, 6GU7_487089 and 6GU7_14218027 showed lower PC1 values, with 31.37% and 30.43%, respectively. At the same time, these two complexes had relatively higher contributions in PC2 and PC3, especially PC2 values of 16.13% and 17.88%. This may indicate that the protein motion is spread across more than one direction, suggesting comparatively higher flexibility and the possibility of multiple conformational changes during the simulation period. For the control complex, the contributions of the first three principal components were 44.18% (PC1), 10.85% (PC2), and 7.21% (PC3), which reflects a stable dynamic behavior of the protein–ligand complex. Interestingly, the PCA profile of 6GU7_174880 was found to be quite similar to that of the control, which may support its potential as a promising ligand with favorable binding stability. In contrast, the higher PC2 and PC3 values observed for 6GU7_487089 and 6GU7_14218027 suggest that these complexes may have relatively more flexible interaction patterns with the target protein (Fig. S4).

3.3.12. Gibbs free energy landscape (FEL) analysis. The Gibbs free energy landscape (FEL) analysis was carried out to explore the conformational stability and energy minima of the 6GU7 protein in complex with the selected ligands and the reference control. In the FEL plots, the dark blue regions represent the lowest energy states, indicating the most stable conformations sampled during the simulation. Among the studied complexes, 6GU7_174880 (Fig. S5B) and the control complex, 6GU7_AZD5438 (Fig. S5D), showed a well-defined and deep energy basin with a comparatively compact distribution. This suggests that both complexes remained in a more stable conformational state throughout the simulation and experienced less conformational fluctuations. The similarity between the ligand CID_174880 and the control indicates that this compound may induce a stable binding conformation comparable to the reference inhibitor. In contrast, 6GU7_487089 (Fig. S5A) displayed a broader low-energy basin with a relatively wider spread across the conformational space. This may indicate that the complex explored multiple conformational states during the simulation, suggesting a comparatively more flexible dynamic behavior. Similarly, 6GU7_14218027 (Fig. S5C) also showed an extended energy basin with noticeable spreading of the low-energy region. Such a pattern may reflect the presence of multiple metastable conformations and greater structural flexibility compared with the control complex. Overall, the FEL results suggest that 6GU7_174880 exhibited a more stable conformational landscape that closely resembles the control, whereas 6GU7_487089 and 6GU7_14218027 showed relatively

broader energy minima, indicating more flexible binding-associated motions (Fig. S5).

3.4. Pharmacokinetics study

3.4.1. Physicochemical and ADME properties prediction. The physicochemical and ADME properties of the selected phytochemicals were evaluated and compared with the reference control compound, AZD5438, to assess their drug-likeness and pharmacokinetic suitability. The molecular weight of all selected compounds was found to be within an acceptable range for oral drug candidates, ranging from 340.37 to 424.49 g mol^{-1}, which is comparable to the control (371.46 g mol^{-1}). Similarly, the topological polar surface area (TPSA) values of the phytochemicals (86.99–110.13 $Å^2$) were also within the favorable range, suggesting good membrane permeability and absorption potential. The consensus log *P* values varied among the compounds, where CID_487089 (4.35) showed relatively higher lipophilicity compared to the control (2.51), while CID_174880 (1.87) and CID_14218027 (3.33) remained within an acceptable range. All selected phytochemicals showed high gastrointestinal (GI) absorption, similar to the control compound, indicating their potential suitability for oral administration. In addition, CID_487089 and CID_14218027 were predicted as non-substrates of P-glycoprotein (P-gp), similar to the control, whereas CID_174880 was identified as a P-gp substrate, which may slightly influence its cellular efflux behavior. None of the compounds violated Lipinski's rule of five, Veber's rule, Egan rule, or Ghose filter (0 violations of all compounds), and this once again indicates that they had good drug-likeness properties.[99–102] None of the compounds, including the control, were predicted to be BBB permeant, which may reduce the possibility of central nervous system-related side effects.[103] The bioavailability score was 0.55 for all selected compounds, which is comparable to the control and suggests favorable oral bioavailability. Regarding structural alerts, CID_14218027 demonstrated a more favorable profile with no PAINS alert and only one Brenk alert, whereas CID_487089 and CID_174880 showed relatively higher alert counts. Among the evaluated compounds, CID_14218027 exhibited broad-spectrum inhibition across all major CYP isoforms (CYP1A2, CYP2C19, CYP2C9, CYP2D6, and CYP3A4), indicating a high likelihood of metabolic interference and an increased risk of drug–drug interactions. In contrast, CID_487089 showed selective inhibition of CYP2C9 and CYP3A4, suggesting a comparatively moderate interaction risk. Notably, CID_174880 demonstrated no inhibitory activity against any of the assessed CYP enzymes, reflecting a more favorable metabolic profile with minimal risk of CYP-mediated interactions. The control compound also inhibited CYP2C9 and CYP3A4, further emphasizing that compounds with minimal or no CYP inhibition, such as CID_174880, are more desirable from a pharmacokinetic and safety perspective. The detailed physicochemical and ADME parameters of all compounds are summarized in Table S10, and their overall ADME profiles are visualized in Fig. 11.

3.4.2. Toxicity analysis. The toxicity assessment revealed that all three selected phytochemicals exhibited a favorable safety profile with no predicted hepatotoxicity, neurotoxicity, cardiotoxicity, mutagenicity, or cytotoxicity. However, all compounds showed potential respiratory toxicity, similar to the reference drug AZD5438, suggesting a possible target-related adverse effect. Notably, the phytochemicals demonstrated immunotoxicity, whereas the control compound was inactive in this regard, indicating a limitation of the selected ligands. Importantly, none of the compounds were predicted to be carcinogenic, in contrast to the control drug, highlighting

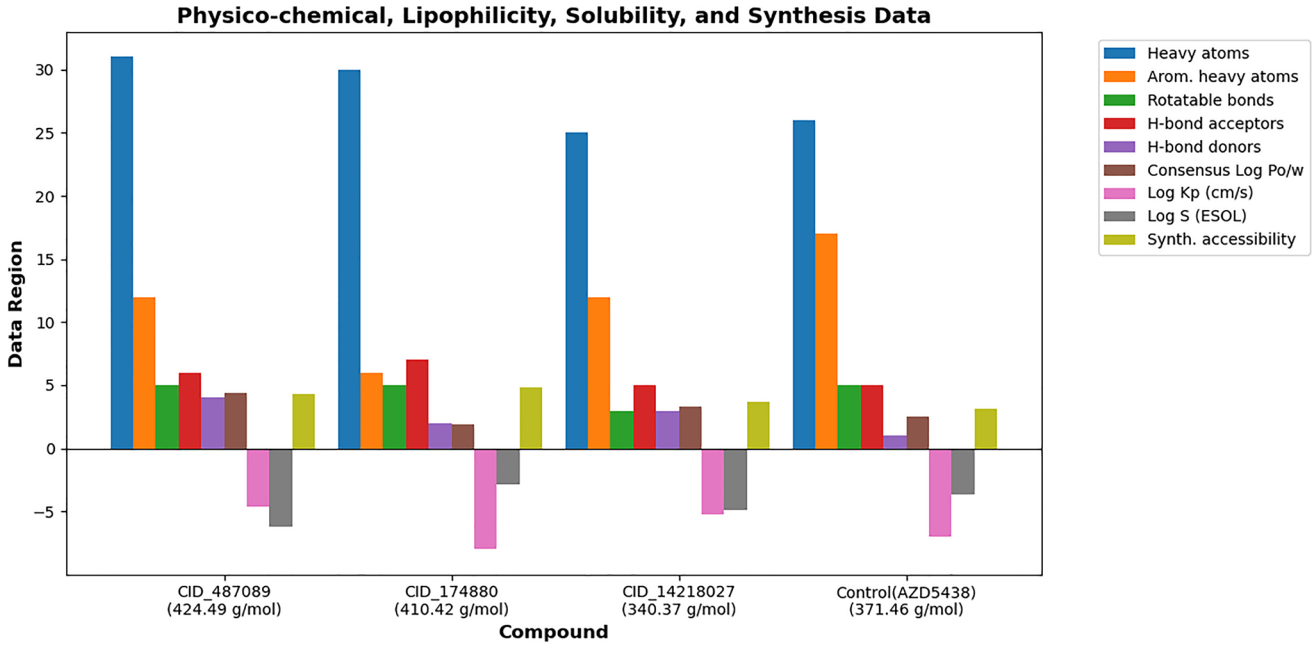

Fig. 11 Physicochemical and ADME-related property comparison of CID_487089, CID_174880, CID_14218027 and control AZD5438.

Table 4 Predicted toxicity profiles of selected lung cancer drug candidates (CID_487089, CID_174880, and CID_14218027) and control (AZD5438)

Target	CID_487089	CID_174880	CID_14218027	Control (AZD5438)
Hepatotoxicity	Inactive	Inactive	Inactive	Inactive
Neurotoxicity	Inactive	Inactive	Inactive	Inactive
Cardiotoxicity	Inactive	Inactive	Inactive	Inactive
Respiratory toxicity	Active	Active	Active	Active
Immunotoxicity	Active	Active	Active	Inactive
Carcinogenicity	Inactive	Inactive	Inactive	Active
Mutagenicity	Inactive	Inactive	Inactive	Inactive
Cytotoxicity	Inactive	Inactive	Inactive	Inactive
LD_{50} mg kg^{-1}	10	300	2000	500
Toxicity class	2	3	4	4

a significant safety advantage. Based on LD_{50} values and toxicity classification, CID_14218027 emerged as the safest candidate with low acute toxicity (Class 4), whereas CID_487089 showed high toxicity (Class 2), making it less suitable for further development (Table 4). Therefore, comprehensive experimental validation is essential to confirm the safety and therapeutic applicability of these compounds.

4. Discussion

Lung cancer is one of the most detrimental malignancies in the entire world. It grows fast, is often detected late, and does not respond well to available treatments.[2] These challenges highlight the need for new treatments guided by a clear strategy for finding targets and developing drugs. Using transcriptomic profiling, network analysis, pathway studies, and computer-based drug discovery, this research systematically identified a key molecular target in lung cancer and proposed a drug to target it. Transcriptomics is the study of all RNA molecules produced in a cell, helping us understand which genes are active and how their activity changes under different conditions.[104] In cancer research, it's often used to find important targets because cancer cells tend to show unusual patterns of gene activity compared to normal cells.[105] However, the results from a single dataset can be affected by technical differences, experimental platforms, or variations among populations. To address this, four independent GEO microarray datasets (GSE19804, GSE10072, GSE18842, and GSE10799) were analyzed across different patient groups (case and control). From this combined analysis, 88 genes were consistently upregulated and 294 genes were downregulated. Repeated changes in these certain genes suggest a real connection to this cancer. After identifying these common differentially expressed genes, we studied how their proteins interact. The strong enrichment of the protein–protein interaction (PPI) network indicates that these connections are not random, showing a real biological relationship between these genes in lung cancer. Then, based on this PPI network we identified the major key hub genes (kHGs) using several analytical methods, which ensured that our results were robust and not dependent on a single approach. Among these interactions, cyclin dependent kinase 1 (CDK1) showed the highest number of interactions, suggesting that it may play a central role in regulating important cancer-related processes. To clarify the regulation of kHGs, we analyzed their interactions with transcription factors (TFs) and microRNAs (miRNAs). Network analysis (Fig. 4) identified FOXC1, GATA2, YY1, E2F1, and HINFP as major TFs, while hsa-miR-192-5p, hsa-miR-92a-3p, hsa-miR-193b-3p, hsa-miR-215-5p, and hsa-miR-155-5p were identified as central miRNA regulators. The study of TFs and miRNAs revealed important hints on the regulation of kHGs at the different levels. Transcription factors are known to activate the expression of genes,[106] whereas the miRNAs are known to regulate the activity of genes once they are transcribed.[107] This combined method enabled us to determine vital regulatory factors which could contribute to dysregulated gene expression in lung cancer and gives a clear understanding for further functional and therapeutic studies. Moreover, we performed functional enrichment analyses (GO and KEGG) using multiple pathway databases to understand the role of these hub genes in lung cancer cells. These enrichment results indicate that lung cancer progression is driven by widespread disruption of cell cycle regulation, mitotic structure, and signaling control rather than by isolated gene alterations. The overall involvement of CDK1 in key mitotic events, strong kinase activity, and extensive interaction with other cell cycle regulators suggest that it functions as a major coordinator of uncontrolled cell division in lung cancer. Thus, this central and recurring role makes CDK1 the most promising candidate for further functional validation and therapeutic exploration.

The reason why LUAD was chosen as a validation is because it is the most prevalent type of lung cancer and it accounts for a substantial percentage of the cases of lung cancer, compared to other types.[108] Multiple analyses of expressions in different independent platforms indicated that CDK1 is not only transcriptionally upregulated but also overexpressed at the protein level in LUAD. In addition, the patients whose CDK1 level was higher had poorer survival and earlier recurrence, which is an obvious clinical effect. In addition to expression, the correlation of CDK1 and immune cell infiltration indicates that its activity moves beyond the expression to influence the tumor microenvironment, especially at the T-cell subsets and macrophage polarization. This implies that CDK1 can regulate tumor growth and immune response. Lastly, the somatic mutations and copy

number changes observed in CDK1 contribute to the oncogenic significance of this protein in a group of LUAD patients. From a biological perspective, CDK1 is more than a statistically prioritized hub gene. As a key regulator of the G2/M transition and mitotic entry, CDK1 promotes sustained tumor cell proliferation when aberrantly activated. Prior studies in lung cancer have also linked elevated CDK1 expression with poor survival, enhanced cell-cycle and DNA-repair signaling, and altered immune-associated pathways in LUAD.[109,110] These observations strengthen the view that CDK1 may act as a functionally important driver of lung cancer progression and a rational candidate for therapeutic targeting.

In this study, we used a structure-based drug discovery approach to efficiently screen a large library of phytochemicals.[111] In this field, the integration of machine learning (ML), molecular docking, and molecular dynamics (MD) simulation studies has transformed the identification and optimization of novel drug candidates.[112,113] We performed cheminformatics-based screening procedure to screen physicochemical and drug-likeness properties of 9577 phytocompounds. After removing duplicate entries and unwanted sub-structures, we finally obtained 2113 phytocompounds, among which 1802 showed no violations of RO5 and Veber's rule. ML is revolutionizing the computational drug discovery approach by reducing traditional experimental time and cost. ML is the best option to calculate the bioactivity of a large dataset with high accuracy within a very short time.[114] To identify the most promising candidates, we applied a ML-based pIC50 prediction model, which helped us identify compounds with higher chances of biological activity. This step reduced the dataset to identify most potential 380 phytocompounds with pIC50 > 6.5 for further computational analysis, including molecular docking, molecular dynamics simulation, and pharmacokinetics analysis. The docking results and post docking MM-GBSA binding energy (ΔG) highlighted the top-ranked three phytochemicals (CID_487089, CID_174880, and CID_14218027) as candidate drug molecules.

Furthermore, molecular dynamics (MD) simulations were used to understand how the selected phytochemicals interact with CDK1 over time. While the apo protein and AZD5438-bound complex showed noticeable structural fluctuations, all phytochemical-bound systems exhibited improved stability throughout the 100 ns simulation. Notably, CID_174880 consistently outperformed the control compound by maintaining lower RMSD and RMSF values (Fig. 9), indicating a stronger and more stable binding mode. Structural compactness analyses further highlighted this difference. The CID_174880 complex showed the lowest SASA and a tightly maintained R_g, reflecting deeper ligand burial and a more compact protein structure than AZD5438.[115] In contrast, the control complex remained relatively solvent-exposed and flexible, suggesting weaker stabilization of CDK1. Post-simulation MM-GBSA results further supported these findings, with CID_487089 showing the most favorable binding free energy (-40.29 kcal mol^{-1}), followed by CID_174880 (-36.06 kcal mol^{-1}), both outperforming the control AZD5438 (-29.53 kcal mol^{-1}). Dynamic motion analysis using PCA showed clear differences between the control compound (AZD5438) and the phytochemicals,[116] which suggests a better fit between the ligand and the binding site. Also, the FEL analysis further supported these findings by demonstrating the stability and conformational states of the ligand–protein interactions. Among the studied compounds, CID_174880 showed a dynamic and energy profile closely comparable to the control, suggesting favorable binding stability. The control compound was predicted to be carcinogenic, while none of the phytochemicals showed this risk, indicating better overall safety. However, all compounds exhibited some potential respiratory toxicity, and the phytochemicals also showed immunotoxic effects, which may limit their development. Based on LD_{50} values and toxicity classes, CID_14218027 was the safest (Class 4), whereas CID_487089 showed higher toxicity (Class 2). CID_174880 displayed a toxicity profile close to the control but with slightly improved safety. Overall, CID_174880 demonstrated promising inhibitory potential with comparatively better safety profiles than the control compound. Therefore, this study could be a useful resource to identify natural CDK1 inhibitors after being validated through the experimental (*in vivo* & *in vitro*) study.

5. Conclusion

This paper displays CDK1 as an important lung cancer driver, disrupting the cell cycle, p53 pathway, and immune microenvironment, such as augmented M1 macrophages, diminished M2 polarization, and unfavorable prognosis owing to overexpression and mutations, including I136N and E57V. These discoveries indicate a shared vulnerability in lung cancer, which puts mitotic dysregulation as one of the prospective treatment targets. Among the phytochemicals evaluated, lactupicrin (CID_174880), derived from *Cichorium intybus*, emerged as the most promising *in silico* prioritized candidate compared with the reference compound AZD5438, based on its favorable docking performance, structural stability, dynamic behavior, and predicted pharmacokinetic and toxicity profiles. Although minor fluctuations were observed in some metrics, all CDK1 complexes maintained overall stability and compatibility. Docking results, MD simulations, MM-GBSA calculations and ADMET analyses consistently support lactupicrin as a computationally prioritized candidate for further development. Further experimental validation, including *in vitro* biochemical assays, cell-based functional studies, and *in vivo* investigations, is required to confirm CDK1 inhibitory activity, anticancer efficacy, and safety. Overall, this research provides a clear workflow for identifying effective CDK1-targeting compounds and offers a roadmap for developing broad-spectrum, mechanism-based therapies for lung cancer, bridging computational predictions with potential real-world applications.

6. Limitations of this study

Although this study applied a comprehensive multiomics and *in silico* approach, several limitations should be considered. First, the transcriptomic analysis was based solely on publicly

available microarray datasets, which may be affected by batch effects, platform-related biases, and limited clinical information. While using four independent cohorts improves reliability, the absence of RNA-seq data may limit the ability to capture finer gene expression variability. Second, the drug discovery pipeline, including machine learning predictions, molecular docking, MM-GBSA analysis, MD simulations, and ADEMT, relies entirely on computational models. Although these methods are useful for prioritizing candidates, they cannot fully represent the complexity of biological systems, tumor heterogeneity, or real pharmacodynamic behavior. Finally, the identified drug candidates, especially lactupicrin, have not yet been validated *in vitro* or *in vivo* models. Therefore, the therapeutic potential, optimal dosage, and safety profiles of the candidate drugs still need to be confirmed through experimental validation.

Author contributions

conceptualization – Md. Ahad Ali and Hriddhi Sarker; methodology –Md. Ahad Ali, Hriddhi Sarker, and Humaira Sheikh; data curation – Hriddhi Sarker, Marguba Kamrun, Bilkis Akter Shifa, and Sujoy Banik; formal analysis – Md. Ahad Ali, Hriddhi Sarker, Humaira Sheikh, and Siam Ahmed; visualization – Hriddhi Sarker, Md. Ahad Ali, Marguba Kamrun, and Tarikul Islam; validation – Md. Ahad Ali, Hriddhi Sarker, and Tarikul Islam; project administration – Md. Ahad Ali; software & resources – Neeraj Kumar, Sujoy Banik, and Bilkis Akter Shifa; supervision – Md. Ahad Ali; writing – original draft, Hriddhi Sarker, Md. Ahad Ali, and Humaira Sheikh; writing – review & editing, Md. Ahad Ali, Marguba Kamrun, and Neeraj Kumar.

Conflicts of interest

All authors declare no conflict of interests.

Data availability

The original data and contributions presented in this study are included in the article and its supplementary information (SI) (Tables S1–S10 and Fig. S1–S5). The training set of our selected compounds, the test set, the chEMBL datasets (containing the consensus predictions) and the related python code for running the QSAR models to predict the bioactivity of the selected compounds can be found on our GitHub repository (https://github.com/ahad004/LUAD_ML_QSAR_Modeling) or can be accessed using the following DOI: https://doi.org/10.5281/zenodo.19841200.

Supplementary information is available. See DOI: https://doi.org/10.1039/d6dd00045b.

References

1 Y. Ji, Y. Zhang, S. Liu, J. Li, Q. Jin, J. Wu, H. Duan, X. Liu, L. Yang and Y. Huang, The Epidemiological Landscape of Lung Cancer: Current Status, Temporal Trend and Future Projections Based on the Latest Estimates from GLOBOCAN 2022, *J. Natl. Cancer Cent.*, 2025, **5**, 278–286, DOI: 10.1016/j.jncc.2025.01.003.

2 F. Bray, M. Laversanne, H. Sung, J. Ferlay, R. L. Siegel, I. Soerjomataram and A. Jemal, Global Cancer Statistics 2022: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries, *CA Cancer J. Clin.*, 2024, **74**, 229–263, DOI: 10.3322/caac.21834.

3 J. Zhou, Y. Xu, J. Liu, L. Feng, J. Yu and D. Chen, Global Burden of Lung Cancer in 2022 and Projections to 2050: Incidence and Mortality Estimates from GLOBOCAN, *J. Cancer Epidemiol.*, 2024, **93**, 102693, DOI: 10.1016/j.canep.2024.102693.

4 S. Bharadwaj, J. M. Mierzwicka, L. Vaňková and P. Malý, Unraveling the Molecular-Pathological Characteristics and Cellular Complexity of the Tumor Immune Microenvironment in Metastatic Non-Small Cell Lung Cancer, *Cell Commun. Signal.*, 2025, **23**, 400, DOI: 10.1186/s12964-025-02410-w.

5 Z. Gu, Y. Heng, R. Fan, J. Luo and L. Ju, Single-Cell RNA Sequencing Reveals Cellular and Molecular Heterogeneity in Extensive-Stage Small Cell Lung Cancer with Different Chemotherapy Responses, *Cancer Cell Int.*, 2025, **25**, 157, DOI: 10.1186/s12935-025-03785-z.

6 A. Jachowski, M. Marcinkowski, J. Szydłowski, O. Grabarczyk, Z. Nogaj, Ł. Marcin, A. Pławski, P. P. Jagodziński and B. K. Słowikowski, Modern Therapies of Nonsmall Cell Lung Cancer, *J. Appl. Genet.*, 2023, **64**, 695–711, DOI: 10.1007/s13353-023-00786-4.

7 M. Araghi, R. Mannani, A. Heidarnejad maleki, A. Hamidi, S. Rostami, S. H. Safa, F. Faramarzi, S. Khorasani, M. Alimohammadi, S. Tahmasebi, *et al.*, Recent Advances in Non-Small Cell Lung Cancer Targeted Therapy; an Update Review, *Cancer Cell Int.*, 2023, **23**, 162, DOI: 10.1186/s12935-023-02990-y.

8 P.-L. Su, N. Furuya, A. Asrar, C. Rolfo, Z. Li, D. P. Carbone and K. He, Recent Advances in Therapeutic Strategies for Non-Small Cell Lung Cancer, *J. Hematol. Oncol.*, 2025, **18**, 35, DOI: 10.1186/s13045-025-01679-1.

9 L. Valdez Capuccino, T. Kleitke, B. Szokol, L. Svajda, F. Martin, F. Bonechi, M. Krekó, S. Azami, A. Montinaro, Y. Wang, *et al.*, CDK9 Inhibition as an Effective Therapy for Small Cell Lung Cancer, *Cell Death Dis.*, 2024, **15**, 345, DOI: 10.1038/s41419-024-06724-4.

10 A. Osoegawa, Y. Takumi, T. Hashimoto, S. Nakatsuji, M. Hori, M. Sakai, T. Karashima, M. Abe, M. Miyawaki and K. Sugio, Cyclin-Dependent Kinase (CDK) 4/6 Inhibition in Non-Small Cell Lung Cancer with Epidermal Growth Factor Receptor (EGFR) Mutations, *Invest. New Drugs*, 2023, **41**, 183–192, DOI: 10.1007/s10637-023-01337-8.

11 E. Panagiotou, G. Gomatou, I. P. Trontzas, N. Syrigos and E. Kotteas, Cyclin-Dependent Kinase (CDK) Inhibitors in Solid Tumors: A Review of Clinical Trials, *Clin. Transl. Oncol.*, 2022, **24**, 161–192, DOI: 10.1007/s12094-021-02688-5.

12 X. Huang, Y. Yin, G. Saha, I. Francis and S. C. Saha, A Comprehensive Numerical Study on the Transport and Deposition of Nasal Sprayed Pharmaceutical Aerosols in

a Nasal-To-Lung Respiratory Tract Model, *Part. Part. Syst. Charact.*, 2025, **42**, 2400004, DOI: 10.1002/ppsc.202400004.
13 X.-Q. Li, X.-J. Cheng, J. Wu, K.-F. Wu and T. Liu, Targeted Inhibition of the PI3K/AKT/MTOR Pathway by (+)-Anthrabenzoxocinone Induces Cell Cycle Arrest, Apoptosis, and Autophagy in Non-Small Cell Lung Cancer, *Cell. Mol. Biol. Lett.*, 2024, **29**, 58, DOI: 10.1186/s11658-024-00578-6.
14 H. Kitai, P. H. Choi, Y. C. Yang, J. A. Boyer, A. Whaley, P. Pancholi, C. Thant, J. Reiter, K. Chen, V. Markov, *et al.*, Combined Inhibition of KRASG12C and MTORC1 Kinase Is Synergistic in Non-Small Cell Lung Cancer, *Nat. Commun.*, 2024, **15**, 6076, DOI: 10.1038/s41467-024-50063-z.
15 J.-L. Huang, L.-M. Wu, S.-Q. Wu, F.-Y. Yuan, H.-Z. Weng, D. Huang, L. Gan, S.-B. Chen, G.-H. Tang and S. Yin, A Small Molecule Targets LIC1 to Suppress Lung Tumor Growth by Inducing Autophagy, *Nat. Chem. Biol.*, 2026, **22**, 459–470, DOI: 10.1038/s41589-025-02040-w.
16 J. Crawford, D. Herndon, K. Gmitter and J. Weiss, The Impact of Myelosuppression on Quality of Life of Patients Treated with Chemotherapy, *Future Oncol.*, 2024, **20**, 1515–1530, DOI: 10.2217/fon-2023-0513.
17 C. Lazzari, V. Gregorc, N. Karachaliou, R. Rosell and M. Santarpia, Mechanisms of Resistance to Osimertinib, *J. Thorac. Dis.*, 2020, **12**, 2851–2858, DOI: 10.21037/jtd.2019.08.30.
18 L. Astolfi, S. Ghiselli, V. Guaran, M. Chicca, E. Simoni, E. Olivetto, G. Lelli and A. Martini, Correlation of Adverse Effects of Cisplatin Administration in Patients Affected by Solid Tumours: A Retrospective Evaluation, *Oncol. Rep.*, 2013, **29**, 1285–1292, DOI: 10.3892/or.2013.2279.
19 E. S. Alsatari, K. R. Smith, S. P. L. Galappaththi, E. A. Turbat-Herrera and S. Dasgupta, The Current Roadmap of Lung Cancer Biology, Genomics and Racial Disparity, *Int. J. Mol. Sci.*, 2025, **26**, 3818, DOI: 10.3390/ijms26083818.
20 G. Gupta, V. P. Samuel, M. M. Rekha, B. Rani, Y. Sasikumar, P. P. Nayak, P. Sudan, K. Goyal, B. G. Oliver, A. Chakraborty, *et al.*, Caspase-Independent Cell Death in Lung Cancer: From Mechanisms to Clinical Applications, *Naunyn-Schmiedebergs Arch Pharmacol*, 2025, **398**, 13031–13048, DOI: 10.1007/s00210-025-04149-0.
21 J. Zhang, X. Zeng, Q. Guo, Z. Sheng, Y. Chen, S. Wan, L. Zhang and P. Zhang, Small Cell Lung Cancer: Emerging Subtypes, Signaling Pathways, and Therapeutic Vulnerabilities, *Exp. Hematol. Oncol.*, 2024, **13**, 78, DOI: 10.1186/s40164-024-00548-w.
22 N. MISHRA, A. SONI, M. KUMARI, G. SINGH, S. K. SHARMA and S. K. SINGH, Targeting Cell Cycle Regulators: A New Paradigm in Cancer Therapeutics, *Biocell*, 2024, **48**, 1639–1666, DOI: 10.32604/biocell.2024.056503.
23 A. Alibakhshi, A. Alagheband Bahrami, E. Mohammadi, S. Ahangarzadeh and M. Mobasheri, In-Silico Design of a New Multi-Epitope Vaccine Candidate against SARS-CoV-2, *Acta Virol.*, 2024, **67**, 12481, DOI: 10.3389/av.2023.12481.
24 F. S. Abdel Razek, S. D. Ibrahim, A. A. Megahed, A. S. Sadik and S. S. A. El-Masry, In Silico Molecular Docking Analysis of Green Tea Bioactive Compounds Targeting Banana Bunchy Top Virus Proteins, *Discov. Appl. Sci.*, 2025, **7**, 1232, DOI: 10.1007/s42452-025-07466-4.
25 Q. M. S. Jamal, S. Khan, M. Khan, A. A. Ansai, J. M. Ashraf, M. Habibullah, A. Farasani, A. M. Madkhali and M. Lohani, Smoking May Increase the Risk of COVID-19 Infection: Evidence from In Silico Analysis, *J. Pharm. Res. Int.*, 2021, 12–21, DOI: 10.9734/jpri/2021/v33i22B31394.
26 Z. Sajid, T. Akhtar, K. Ahmad and M. Haroon, Molecular Docking Simulation and ADMET/Pharmacokinetic Screening of Newly Designed 2-(2-(Aryl)-4-oxo-4,5-dihydrothiazol-5-yl)Acetohydrazides as Potential Antitubercular Agents, *ChemistrySelect*, 2024, **9**(44), e202403715, DOI: 10.1002/slct.202403715.
27 P. Chunarkar-Patil, M. Kaleem, R. Mishra, S. Ray, A. Ahmad, D. Verma, S. Bhayye, R. Dubey, H. N. Singh and S. Kumar, Anticancer Drug Discovery Based on Natural Products: From Computational Approaches to Clinical Studies, *Biomedicines*, 2024, **12**(1), 201, DOI: 10.3390/biomedicines12010201.
28 M. Huang, J.-J. Lu and J. Ding, Natural Products in Cancer Therapy: Past, Present and Future, *Nat. Prod. Bioprospect.*, 2021, **11**, 5–13, DOI: 10.1007/s13659-020-00293-7.
29 S. T. Asma, U. Acaroz, K. Imre, A. Morar, S. R. A. Shah, S. Z. Hussain, D. Arslan-Acaroz, H. Demirbas, Z. Hajrulai-Musliu, F. R. Istanbullugil, *et al.*, Natural Products/Bioactive Compounds as a Source of Anticancer Drugs, *Cancers*, 2022, **14**, 6203, DOI: 10.3390/cancers14246203.
30 G. M. Cragg and D. J. Newman, Natural Products: A Continuing Source of Novel Drug Leads, *Biochim. Biophys. Acta Gen. Subj.*, 2013, **1830**, 3670–3695, DOI: 10.1016/j.bbagen.2013.02.008.
31 K. Mohanraj, B. S. Karthikeyan, R. P. Vivek-Ananth, R. P. B. Chand, S. R. Aparna, P. Mangalapandi and A. Samal, IMPPAT: A Curated Database of Indian Medicinal Plants, Phytochemistry and Therapeutics, *Sci. Rep.*, 2018, **8**, 4329, DOI: 10.1038/s41598-018-22631-z.
32 S. M. Soyer, P. Ozbek and C. Kasavi, Lung Adenocarcinoma Systems Biomarker and Drug Candidates Identified by Machine Learning, Gene Expression Data, and Integrative Bioinformatics Pipeline, *OMICS J. Integr. Biol.*, 2024, **28**, 408–420, DOI: 10.1089/omi.2024.0121.
33 Y. Li, Y. Cai, L. Ji, B. Wang, D. Shi and X. Li, Machine Learning and Bioinformatics Analysis of Diagnostic Biomarkers Associated with the Occurrence and Development of Lung Adenocarcinoma, *PeerJ*, 2024, **12**, e17746, DOI: 10.7717/peerj.17746.
34 W. Roh, Y. Geffen, H. Cha, M. Miller, S. Anand, J. Kim, D. I. Heiman, J. F. Gainor, P. W. Laird, A. D. Cherniack, *et al.*, High-Resolution Profiling of Lung Adenocarcinoma Identifies Expression Subtypes with Specific Biomarkers and Clinically Relevant Vulnerabilities, *Cancer Res.*, 2022, **82**, 3917–3931, DOI: 10.1158/0008-5472.CAN-22-0432.
35 R. X. Huang, D. Siriwanna, W. C. Cho, T. K. Wan, Y. R. Du, A. N. Bennett, Q. E. He, J. D. Liu, X. T. Huang and K. H. K. Chan, Lung Adenocarcinoma-Related Target Gene Prediction and Drug Repositioning, *Front.*

Pharmacol, 2022, **13**, 936758, DOI: 10.3389/fphar.2022.936758.

36 Y. Li, K. Ma, H. Wang, Z. Liu and Z. Li, Identification of Therapeutic Targets in Lung Adenocarcinoma Using Mendelian Randomization and Multi-Omics, *Discov. Oncol.*, 2025, **16**(1), 10258, DOI: 10.1007/s12672-025-02835-2.

37 K. Jia, Y. Wang, Q. Cao and Y. Wang, Extensive Prediction of Drug Response in Mutation-Subtype-Specific LUAD with Machine Learning Approach, *Oncol. Res.*, 2024, **32**, 409–419, DOI: 10.32604/or.2023.042863.

38 R. Qureshi, S. A. Basit, J. A. Shamsi, X. Fan, M. Nawaz, H. Yan and T. Alam, Machine Learning Based Personalized Drug Response Prediction for Lung Cancer Patients, *Sci. Rep.*, 2022, **12**, 18935, DOI: 10.1038/s41598-022-23649-0.

39 M. Sobhan and A. M. Mondal, Explainable Machine Learning to Identify Patient-Specific Biomarkers for Lung Cancer, in. *Proc. – 2022 IEEE Int. Conf. Bioinforma. Biomed. BIBM*, 2022, pp. 3152–3159, DOI: 10.1109/BIBM55620.2022.9995516.

40 T. M. Okyay, I. Yilmaz and M. Koldas, Machine Learning-Based Bioactivity Prediction of Porphyrin Derivatives: Molecular Descriptors, Clustering, and Model Evaluation, *Photochem. Photobiol. Sci.*, 2025, **24**, 923–937, DOI: 10.1007/s43630-025-00733-8.

41 T.-P. Lu, M.-H. Tsai, J.-M. Lee, C.-P. Hsu, P.-C. Chen, C.-W. Lin, J.-Y. Shih, P.-C. Yang, C. K. Hsiao, L.-C. Lai, *et al.*, Identification of a Novel Biomarker, SEMA5A, for Non-Small Cell Lung Carcinoma in Nonsmoking Women, *Cancer Epidemiol. Biomarkers Prev.*, 2010, **19**, 2590–2597, DOI: 10.1158/1055-9965.EPI-10-0332.

42 T.-P. Lu, C. K. Hsiao, L.-C. Lai, M.-H. Tsai, C.-P. Hsu, J.-M. Lee and E. Y. Chuang, Identification of Regulatory SNPs Associated with Genetic Modifications in Lung Adenocarcinoma, *BMC Res. Notes*, 2015, **8**, 92, DOI: 10.1186/s13104-015-1053-8.

43 M. T. Landi, T. Dracheva, M. Rotunno, J. D. Figueroa, H. Liu, A. Dasgupta, F. E. Mann, J. Fukuoka, M. Hames, A. W. Bergen, *et al.*, Gene Expression Signature of Cigarette Smoking and Its Role in Lung Adenocarcinoma Development and Survival, *PLoS One*, 2008, **3**, e1651, DOI: 10.1371/journal.pone.0001651.

44 A. Sanchez-Palencia, M. Gomez-Morales, J. A. Gomez-Capilla, V. Pedraza, L. Boyero, R. Rosell and M. E. Fárez-Vidal, Gene Expression Profiling Reveals Novel Biomarkers in Nonsmall Cell Lung Cancer, *Int. J. Cancer*, 2011, **129**, 355–364, DOI: 10.1002/ijc.25704.

45 M. Wrage, S. Ruosaari, P. P. Eijk, J. T. Kaifi, J. Hollmén, E. F. Yekebas, J. R. Izbicki, R. H. Brakenhoff, T. Streichert, S. Riethdorf, *et al.*, Genomic Profiles Associated with Early Micrometastasis in Lung Cancer: Relevance of 4q Deletion, *Clin. Cancer Res.*, 2009, **15**, 1566–1574, DOI: 10.1158/1078-0432.CCR-08-2188.

46 G. K. Smyth, Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments, *Stat. Appl. Genet. Mol. Biol.*, 2004, **3**, 3, DOI: 10.2202/1544-6115.1027.

47 G. Singh and B. Soman, *Data Transformation Using Dplyr Package in R*, 2019.

48 P. M. Valero-Mora, Ggplot2: Elegant Graphics for Data Analysis, *J. Stat. Softw.*, 2010, **35**(1), 1–3, DOI: 10.18637/jss.v035.b01.

49 C. H. Gao, G. Yu and P. Cai, GgVennDiagram: An Intuitive, Easy-to-Use, and Highly Customizable R Package to Generate Venn Diagram, *Front. Genet.*, 2021, **12**, 706907, DOI: 10.3389/fgene.2021.706907.

50 D. Szklarczyk, J. H. Morris, H. Cook, M. Kuhn, S. Wyder, M. Simonovic, A. Santos, N. T. Doncheva, A. Roth, P. Bork, *et al.*, The STRING Database in 2017: Quality-Controlled Protein–Protein Association Networks, Made Broadly Accessible, *Nucleic Acids Res.*, 2017, **45**, D362–D368, DOI: 10.1093/nar/gkw937.

51 P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski and T. C. Ideker, A Software Environment for Integrated Models of Biomolecular Interaction Networks, *Genome Res.*, 2003, **13**, 2498–2504, DOI: 10.1101/gr.1239303.

52 C.-H. Chin, S.-H. Chen, H.-H. Wu, C.-W. Ho, M.-T. Ko and C.-Y. Lin, CytoHubba: Identifying Hub Objects and Sub-Networks from Complex Interactome, *BMC Syst. Biol.*, 2014, **8**, S11, DOI: 10.1186/1752-0509-8-S4-S11.

53 A. Khan, O. Fornes, A. Stigliani, M. Gheorghe, J. A. Castro-Mondragon, R. Van Der Lee, A. Bessy, J. Chèneby, S. R. Kulkarni, G. Tan, *et al.*, JASPAR 2018: Update of the Open-Access Database of Transcription Factor Binding Profiles and Its Web Framework, *Nucleic Acids Res.*, 2018, **46**, D260–D266, DOI: 10.1093/nar/gkx1126.

54 S.-D. Hsu, F.-M. Lin, W.-Y. Wu, C. Liang, W.-C. Huang, W.-L. Chan, W.-T. Tsai, G.-Z. Chen, C.-J. Lee, C.-M. Chiu, *et al.*, MiRTarBase: A Database Curates Experimentally Validated MicroRNA–Target Interactions, *Nucleic Acids Res.*, 2011, **39**, D163–D169, DOI: 10.1093/nar/gkq1107.

55 J. Xia, E. E. Gill and R. E. W. Hancock, NetworkAnalyst for Statistical, Visual and Network-Based Meta-Analysis of Gene Expression Data, *Nat. Protoc.*, 2015, **10**, 823–844, DOI: 10.1038/nprot.2015.052.

56 B. T. Sherman, M. Hao, J. Qiu, X. Jiao, M. W. Baseler, H. C. Lane, T. Imamichi and W. Chang, DAVID: A Web Server for Functional Enrichment Analysis and Functional Annotation of Gene Lists (2021 Update), *Nucleic Acids Res.*, 2022, **50**, W216–W221, DOI: 10.1093/nar/gkac194.

57 M. V. Kuleshov, M. R. Jones, A. D. Rouillard, N. F. Fernandez, Q. Duan, Z. Wang, S. Koplev, S. L. Jenkins, K. M. Jagodnik, A. Lachmann, *et al.*, Enrichr: A Comprehensive Gene Set Enrichment Analysis Web Server 2016 Update, *Nucleic Acids Res.*, 2016, **44**, W90–W97, DOI: 10.1093/nar/gkw377.

58 M. Helmy, R. Agrawal, J. Ali, M. Soudy, T. T. Bui and K. Selvarajoo, GeneCloudOmics: A Data Analytic Cloud Platform for High-Throughput Gene Expression Analysis, *Front. Bioinform.*, 2021, **1**, 693836, DOI: 10.3389/fbinf.2021.693836.

59 T. Li, J. Fu, Z. Zeng, D. Cohen, J. Li, Q. Chen, B. Li and X. S. Liu, TIMER2.0 for Analysis of Tumor-Infiltrating Immune Cells, *Nucleic Acids Res.*, 2020, **48**(W1), W509–W514, DOI: 10.1093/NAR/GKAA407.

60 Z. Tang, B. Kang, C. Li, T. Chen and Z. Zhang, GEPIA2: An Enhanced Web Server for Large-Scale Expression Profiling and Interactive Analysis, *Nucleic Acids Res.*, 2019, **47**, W556–W560, DOI: 10.1093/nar/gkz430.

61 D. S. Chandrashekar, S. K. Karthikeyan, P. K. Korla, H. Patel, A. R. Shovon, M. Athar, G. J. Netto, Z. S. Qin, S. Kumar, U. Manne, *et al.*, UALCAN: An Update to the Integrated Cancer Data Analysis Platform, *Neoplasia*, 2022, **25**, 18–27, DOI: 10.1016/j.neo.2022.01.001.

62 J. Gao, B. A. Aksoy, U. Dogrusoz, G. Dresdner, B. Gross, S. O. Sumer, Y. Sun, A. Jacobsen, R. Sinha, E. Larsson, *et al.*, Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the CBioPortal, *Sci. Signal.*, 2013, **6**(269), pl1, DOI: 10.1126/scisignal.2004088.

63 S. K. Burley, C. Bhikadiya, C. Bi, S. Bittrich, L. Chen, G. V. Crichlow, C. H. Christie, K. Dalenberg, L. Di Costanzo, J. M. Duarte, *et al.*, RCSB Protein Data Bank: Powerful New Tools for Exploring 3D Structures of Biological Macromolecules for Basic and Applied Research and Education in Fundamental Biology, Biomedicine, Biotechnology, Bioengineering and Energy Sciences, *Nucleic Acids Res.*, 2021, **49**, D437–D451, DOI: 10.1093/nar/gkaa1038.

64 BIOVIA, D. S., *Discovery Studio Visualizer V21.1.0.20298. BIOVIA*, Dassault Systèmes, 2005.

65 N. Guex and M. C. Peitsch, SWISS-MODEL and the Swiss-Pdb Viewer: An Environment for Comparative Protein Modeling, *Electrophoresis*, 1997, **18**, 2714–2723, DOI: 10.1002/elps.1150181505.

66 A. A. Puelles, L. L. Bastos, V. M. Paixão, S. C. Araujo and R. C. de Melo Minardi In. *Virtual Screening*. 2024, pp. 209–236.

67 G. Xiong, Z. Wu, J. Yi, L. Fu, Z. Yang, C. Hsieh, M. Yin, X. Zeng, C. Wu, A. Lu, *et al.*, ADMETlab 2.0: An Integrated Online Platform for Accurate and Comprehensive Predictions of ADMET Properties, *Nucleic Acids Res.*, 2021, **49**, W5–W14, DOI: 10.1093/nar/gkab255.

68 C. A. Lipinski, Lead- and Drug-like Compounds: The Rule-of-Five Revolution, *Drug Discov. Today Technol.*, 2004, **1**, 337–341, DOI: 10.1016/j.ddtec.2004.11.007.

69 W. Caldwell, Z. Yan, W. Lang and A. Masucci, J. The IC50 Concept Revisited, *Curr. Top. Med. Chem.*, 2012, **12**(11), 1282–1290, DOI: 10.2174/156802612800672844.

70 M. A. Ali, H. Sarker, T. Khan, H. Sheikh, A. Saif, F. B. Farid, S. Afrin, M. A. Khatun and N. Kumar, Multi-Omics Pan-Cancer Profiling of CDK2 and in Silico Identification of Plant-Derived Inhibitors Using Machine Learning Approaches, *RSC Adv.*, 2025, **15**, 36938–36968, DOI: 10.1039/D5RA05535K.

71 A. Saif, M. T. Islam, M. O. Raihan, N. Yousefi, M. A. Rahman, H. Faridi, A. R. Hasan, M. M. Hossain, R. M. Saleem, G. M. Albadrani, *et al.*, Pan-Cancer Analysis of CDC7 in Human Tumors: Integrative Multi-Omics Insights and Discovery of Novel Marine-Based Inhibitors through Machine Learning and Computational Approaches, *Comput. Biol. Med.*, 2025, **190**, 110044, DOI: 10.1016/j.compbiomed.2025.110044.

72 H. Gubler, U. Schopfer and E. Jacoby, Theoretical and Experimental Relationships between Percent Inhibition and IC50 Data Observed in High-Throughput Screening, *J. Biomol. Screen*, 2013, **18**, 1–13, DOI: 10.1177/1087057112455219.

73 K. Danishuddin and A. U. Descriptors, Their Selection Methods in QSAR Analysis: Paradigm for Drug Design, *Drug Discov. Today*, 2016, **21**, 1291–1302, DOI: 10.1016/j.drudis.2016.06.013.

74 Q. Guo, S. Hernandez-Hernandez and P. J. Ballester, UMAP-Based Clustering Split for Rigorous Evaluation of AI Models for Virtual Screening on Cancer Cell Lines, *J. Cheminform.*, 2025, **17**, 94, DOI: 10.1186/s13321-025-01039-8.

75 S. Majumdar and S. C. Basak, Beware of External Validation! - A Comparative Study of Several Validation Techniques Used in QSAR Modelling, *Curr. Comput. Aided Drug Des.*, 2018, **14**, 284–291, DOI: 10.2174/1573409914666180426144304.

76 X.-Y. Meng, H.-X. Zhang, M. Mezei and M. Cui, Molecular Docking: A Powerful Approach for Structure-Based Drug Discovery, *Curr. Comput. Aided Drug Des.*, 2012, **7**, 146–157, DOI: 10.2174/157340911795677602.

77 S. K. Kondapuram, S. Sarvagalla and M. S. Coumar, Docking-Based Virtual Screening Using PyRx Tool: Autophagy Target Vps34 as a Case Study, In *Molecular Docking for Computer-Aided Drug Design*, Elsevier, 2021, pp. 463–477.

78 A. V. Danish Ahmad, S. W. Khan, Q. Yasar, M. S. Shaikh and M. M. Khan, Computational Biology Approach to Predict Molecular Mechanism in Cancer, *Oral Oncol Rep.*, 2024, **12**, 100651, DOI: 10.1016/j.oor.2024.100651.

79 A. J. Owoloye, F. C. Ligali, O. A. Enejoh, A. Z. Musa, O. Aina, E. T. Idowu and K. M. Oyebola, Molecular Docking, Simulation and Binding Free Energy Analysis of Small Molecules as PfHT1 Inhibitors, *PLoS One*, 2022, **17**, e0268269, DOI: 10.1371/journal.pone.0268269.

80 U. C. Ogbodo, O. A. Enejoh, C. H. Okonkwo, P. Gnanasekar, P. W. Gachanja, S. Osata, H. C. Atanda, E. A. Iwuchukwu, I. Achilonu and O. I. Awe, Computational Identification of Potential Inhibitors Targeting Cdk1 in Colorectal Cancer, *Front. Chem.*, 2023, **11**, 1264808, DOI: 10.3389/fchem.2023.1264808.

81 P. A. Kollman, I. Massova, C. Reyes, B. Kuhn, S. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, *et al.*, Calculating Structures and Free Energies of Complex Molecules: Combining Molecular Mechanics and Continuum Models, *Acc. Chem. Res.*, 2000, **33**, 889–897, DOI: 10.1021/ar000033j.

82 R. A. Friesner, R. B. Murphy, M. P. Repasky, L. L. Frye, J. R. Greenwood, T. A. Halgren, P. C. Sanschagrin and D. T. Mainz, Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for

Protein–Ligand Complexes, *J. Med. Chem.*, 2006, **49**, 6177–6196, DOI: 10.1021/jm051256o.

83 M. K. Gilson, J. A. Given, B. L. Bush and J. A. McCammon, The Statistical-Thermodynamic Basis for Computation of Binding Affinities: A Critical Review, *Biophys. J.*, 1997, **72**, 1047–1069, DOI: 10.1016/S0006-3495(97)78756-3.

84 H. Arya and T. K. Bhatt, Molecular Dynamics Simulations. In *The Design & Development of Novel Drugs and Vaccines*, Elsevier, 2021, pp. 65–81.

85 K. Roos, C. Wu, W. Damm, M. Reboul, J. M. Stevenson, C. Lu, M. K. Dahlgren, S. Mondal, W. Chen, L. Wang, *et al.*, OPLS3e: Extending Force Field Coverage for Drug-Like Small Molecules, *J. Chem. Theory Comput.*, 2019, **15**, 1863–1874, DOI: 10.1021/acs.jctc.8b01026.

86 M. S. Valdés-Tresanco, M. E. Valdés-Tresanco, P. A. Valiente and E. Moreno, Gmx_MMPBSA: A New Tool to Perform End-State Free Energy Calculations with GROMACS, *J. Chem. Theory Comput.*, 2021, **17**, 6281–6291, DOI: 10.1021/acs.jctc.1c00645.

87 B. R. Miller, T. D. McGee, J. M. Swails, N. Homeyer, H. Gohlke and A. E. Roitberg, MMPBSA.Py: An Efficient Program for End-State Free Energy Calculations, *J. Chem. Theory Comput.*, 2012, **8**, 3314–3321, DOI: 10.1021/ct300418h.

88 M. R. Shirts, C. Klein, J. M. Swails, J. Yin, M. K. Gilson, D. L. Mobley, D. A. Case and E. D. Zhong, Lessons Learned from Comparing Molecular Dynamics Engines on the SAMPL5 Dataset, *J. Comput. Aided Mol. Des.*, 2017, **31**, 147–161, DOI: 10.1007/s10822-016-9977-1.

89 A. Daina, O. Michielin and V. Zoete, SwissADME: A Free Web Tool to Evaluate Pharmacokinetics, Drug-Likeness and Medicinal Chemistry Friendliness of Small Molecules, *Sci. Rep.*, 2017, **7**, 42717, DOI: 10.1038/srep42717.

90 P. Banerjee, E. Kemmler, M. Dunkel and R. Preissner, ProTox 3.0: A Webserver for the Prediction of Toxicity of Chemicals, *Nucleic Acids Res.*, 2024, gkae303, DOI: 10.1093/nar/gkae303.

91 M. Ashtiani, A. Salehzadeh-Yazdi, Z. Razaghi-Moghadam, H. Hennig, O. Wolkenhauer, M. Mirzaie and M. Jafari, A Systematic Survey of Centrality Measures for Protein-Protein Interaction Networks, *BMC Syst. Biol.*, 2018, **12**, 80, DOI: 10.1186/s12918-018-0598-2.

92 M. Wang, H. Wang and H. Zheng, A Mini Review of Node Centrality Metrics in Biological Networks, *Int. J. Intell. Netw.*, 2022, 99–110, DOI: 10.53941/ijndi0101009.

93 D. J. Wood, S. Korolchuk, N. J. Tatum, L.-Z. Wang, J. A. Endicott, M. E. M. Noble and M. P. Martin, Differences in the Conformational Energy Landscape of CDK1 and CDK2 Suggest a Mechanism for Achieving Selective CDK Inhibition, *Cell Chem. Biol.*, 2019, **26**, 121–130.e5, DOI: 10.1016/j.chembiol.2018.10.015.

94 G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye and T.-Y. Liu, LightGBM: A Highly Efficient Gradient Boosting Decision Tree, In *Proceedings of the Advances in Neural Information Processing Systems*, ed. Guyon, I., Luxburg, U. Von, Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S. and Garnett, R., Curran Associates, Inc., 2017, vol. 30.

95 X. Zhao, Y. Liu and Q. Zhao, Improved LightGBM for Extremely Imbalanced Data and Application to Credit Card Fraud Detection, *IEEE Access*, 2024, **12**, 159316–159335, DOI: 10.1109/ACCESS.2024.3487212.

96 A. P. Bento, A. Gaulton, A. Hersey, L. J. Bellis, J. Chambers, M. Davies, F. A. Krüger, Y. Light, L. Mak, S. McGlinchey, *et al.*, The ChEMBL Bioactivity Database: An Update, *Nucleic Acids Res.*, 2014, **42**, D1083–D1090, DOI: 10.1093/nar/gkt1031.

97 L. Zhao, W. Jiang, Z. Zhu, F. Pan, X. Xing, F. Zhou and L. Zhao, Rosemarinic Acid-Induced Destabilization of Aβ Peptides: Insights from Molecular Dynamics Simulations, *Foods*, 2024, **13**(24), 4170, DOI: 10.3390/foods13244170.

98 A. S. Abouzied, S. Alqarni, K. M. Younes, S. M. Alanazi, D. M. Alrsheed, R. K. Alhathal, B. Huwaimel and A. M. Elkashlan, Structural and Free Energy Landscape Analysis for the Discovery of Antiviral Compounds Targeting the Cap-Binding Domain of Influenza Polymerase PB2, *Sci. Rep.*, 2024, **14**, 25441, DOI: 10.1038/s41598-024-69816-3.

99 D. F. Veber, S. R. Johnson, H. Y. Cheng, B. R. Smith, K. W. Ward and K. D. Kopple, Molecular Properties That Influence the Oral Bioavailability of Drug Candidates, *J. Med. Chem.*, 2002, **45**, 2615–2623, DOI: 10.1021/jm020017n.

100 F. T. Ndombera, Revisiting Cheminformatics and Mechanisms of Action of Chloroquine and Hydroxychloroquine in Targeting Covid-19, *J. Biol. Chem.*, 2020, **3**, 1–11, DOI: 10.17303/jbcg.2020.3.101.

101 M. Rashid, O. Afzal and A. S. A. Altamimi, Benzimidazole Molecule Hybrid With Oxadiazole Ring As Antiproliferative Agents: In-Silico Analysis, Synthesis And Biological Evaluation, *J. Chil. Chem. Soc.*, 2021, **66**, 5164–5182, DOI: 10.4067/S0717-97072021000205164.

102 L. Z. Benet, C. M. Hosey, O. Ursu and T. I. Oprea, BDDCS, the Rule of 5 and Drugability, *Adv. Drug Deliv. Rev.*, 2016, **101**, 89–98, DOI: 10.1016/j.addr.2016.05.007.

103 Y. Sun, M. Zabihi, Q. Li, X. Li, B. J. Kim, E. E. Ubogu, S. N. Raja, U. Wesselmann and C. Zhao, Drug Permeability: From the Blood–Brain Barrier to the Peripheral Nerve Barriers, *Adv. Ther.*, 2023, **6**, 2200150, DOI: 10.1002/adtp.202200150.

104 M. K. Heavey, D. Durmusoglu, N. Crook and A. C. Anselmo, Discovery and Delivery Strategies for Engineered Live Biotherapeutic Products, *Trends Biotechnol.*, 2022, **40**, 354–369, DOI: 10.1016/j.tibtech.2021.08.002.

105 M. Sibai, S. Cervilla, D. Grases, E. Musulen, R. Lazcano, C.-K. Mo, V. Davalos, A. Fortian, A. Bernat, M. Romeo, *et al.*, The Spatial Landscape of Cancer Hallmarks Reveals Patterns of Tumor Ecological Dynamics and Drug Sensitivity, *Cell Rep.*, 2025, **44**, 115229, DOI: 10.1016/j.celrep.2024.115229.

106 A. Boija, I. A. Klein, B. R. Sabari, A. Dall'Agnese, E. L. Coffey, A. V. Zamudio, C. H. Li, K. Shrinivas, J. C. Manteiga, N. M. Hannett, *et al.*, Transcription Factors Activate Genes through the Phase-Separation Capacity of Their

Activation Domains, *Cell*, 2018, **175**, 1842–1855.e16, DOI: 10.1016/j.cell.2018.10.042.

107 C. Catalanotto, C. Cogoni and G. Zardo, MicroRNA in Control of Gene Expression: An Overview of Nuclear Functions, *Int. J. Mol. Sci.*, 2016, **17**(10), 1712, DOI: 10.3390/ijms17101712.

108 Y. Chen, L. Jin, Z. Jiang, S. Liu and W. Feng, Identifying and Validating Potential Biomarkers of Early Stage Lung Adenocarcinoma Diagnosis and Prognosis, *Front. Oncol.*, 2021, **11**, 644426, DOI: 10.3389/fonc.2021.644426.

109 Q. Du, W. Liu, T. Mei, J. Wang, T. Qin and D. Huang, Prognostic and Immunological Characteristics of CDK1 in Lung Adenocarcinoma: A Systematic Analysis, *Front. Oncol.*, 2023, **13**, 1128443, DOI: 10.3389/fonc.2023.1128443.

110 S. Li, H. Li, Y. Cao, H. Geng, F. Ren, K. Li, C. Dai and N. Li, Integrated Bioinformatics Analysis Reveals CDK1 and PLK1 as Potential Therapeutic Targets of Lung Adenocarcinoma, *Medicine*, 2021, **100**, e26474, DOI: 10.1097/MD.0000000000026474.

111 M. Batool, B. Ahmad and S. Choi, A Structure-Based Drug Discovery Paradigm, *Int. J. Mol. Sci.*, 2019, **20**, 2783, DOI: 10.3390/ijms20112783.

112 O. A. Enejoh, C. H. Okonkwo, H. Nortey, O. A. Kemiki, A. Moses, F. N. Mbaoji, A. S. Yusuf and O. I. Awe, Machine Learning and Molecular Dynamics Simulations Predict Potential TGR5 Agonists for Type 2 Diabetes Treatment, *Front. Chem.*, 2024, **12**, 1503593, DOI: 10.3389/fchem.2024.1503593.

113 M. Di Stefano, S. Galati, G. Ortore, I. Caligiuri, F. Rizzolio, C. Ceni, S. Bertini, G. Bononi, C. Granchi, M. Macchia, *et al.*, Machine Learning-Based Virtual Screening for the Identification of Cdk5 Inhibitors, *Int. J. Mol. Sci.*, 2022, **23**(18), 10653, DOI: 10.3390/ijms231810653.

114 R. S. Bhimanwar, K. B. Lokhande, A. Shrivastava, A. Singh, S. S. Chitlange and A. Mittal, Identification of Potential Drug Candidates as TGR5 Agonist to Combat Type II Diabetes Using in Silico Docking and Molecular Dynamics Simulation Studies, *J. Biomol. Struct. Dyn.*, 2023, **41**, 13314–13331, DOI: 10.1080/07391102.2023.2173654.

115 V. Hurmach, V. Karaushu, S. Prylutska, Z. Klestova, S. Vyzhva, Y. Prylutskyy, U. Ritter and V. Garamus, In Silico Analysis of C60 Fullerene Interaction with TMPRSS2: Toward Novel COVID-19 Prevention Approaches, *Molecules*, 2025, **30**(23), 4586, DOI: 10.3390/molecules30234586.

116 C. C. David and D. J. Jacobs, In, *Principal Component Analysis: A Method for Determining the Essential Dynamics of Proteins*. 2014. pp. 193–226.