Multi-Scale Computational Investigation of Heliotropium indicum Phytochemicals as Multi-Target Therapeutics Against Breast Cancer: Integration of Network Pharmacology, Molecular Dynamics with Free- Energy Calculation, and Machine Learning

Shanmugavani S
Vels Institute of Science Technology and Advanced Studies
Nishandhini M
Vels Institute of Science Technology and Advanced Studies
Abhirami L
BioCogniz
Anoosha Paruchuri
BioCogniz
Sivakumar Dakshinamurthy
sivabioinfo@gmail.com

BioCogniz
Radha Mahendran
Vels Institute of Science Technology and Advanced Studies

Research Article

Keywords: Heliotropium indicum, breast cancer, network pharmacology, consensus docking, molecular dynamics, pharmacophore modelling, ADMET, machine learning, polypharmacology, natural products

Posted Date: May 29th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9732760/v1

Additional Declarations: No competing interests reported.

Multi-Scale Computational Investigation of *Heliotropium indicum* Phytochemicals as Multi-Target Therapeutics Against Breast Cancer: Integration of Network Pharmacology, Molecular Dynamics with Free-Energy Calculation, and Machine Learning

Shanmugavani S[1], M.Nishandhini[1], Abhirami L[2], Anoosha Paruchuri[2,3], Dakshinamurthy Sivakumar[2,*], Radha Mahendran[1,*]

[1]Department of Bioinformatics , School of Life Sciences, Vels Institute of Science Technology and Advanced Studies, Pallavaram, Chennai-600117, Tamilnadu, India.

[2]Research and Development, BioCogniz, Chennai 600101, Tamilnadu, India.

[3]Department of Biotechnology, Prathyusha Engineering College, Tiruvallur 602025, Tamil Nadu, India.

Address for correspondence: siva@biocogniz.com, mahenradha@gmail.com

Abstract

Background. Breast cancer (BC) remains a highly heterogeneous malignancy characterised by subtype-specific therapeutic vulnerabilities, pathway redundancy and adaptive resistance mechanisms that necessitate multi-target therapeutic strategies. *Heliotropium indicum* L. (Boraginaceae), a medicinal plant traditionally associated with diverse pharmacological activities, has not previously been investigated using an integrated systems-level computational framework against clinically relevant BC-associated targets.

Methods. A total of 1,365 phytochemical records were compiled from PubMed, IMPPAT, NPACT, KNApSAcK, COCONUT and Dr. Duke's databases. Structural curation and multi-parameter drug-likeness filtering yielded 95 prioritised phytochemicals. A subtype-relevant 13-target BC panel (ERα, PR, HER2, EGFR, PARP1, PD-L1, TOP2A, PI3Kα, AKT1, CDK4, BCL-2, mTOR and Aromatase) was evaluated using TCGA-BRCA transcriptomic profiling, Kaplan–Meier survival analysis and network-pharmacology prioritisation. Multi-engine consensus docking integrating AutoDock Vina, AutoDock4, GNINA-CNN and PLANTS was performed alongside pharmacophore modelling and multi-target enrichment analysis. Translational filtering was conducted using integrated SwissADME, pkCSM, ADMETlab 3.0 and ProTox prediction platforms prior to molecular-dynamics evaluation. Three prioritised complexes underwent 200 ns molecular-dynamics simulation with PCA, free-energy landscape and SASA analyses. Orthogonal machine-learning validation using Random Forest, XGBoost and SVM classifiers trained on ChEMBL-derived datasets was subsequently performed across the selected therapeutic targets.

Results. Transcriptomic and network analyses identified prominent involvement of PI3K–AKT, ErbB, endocrine and apoptosis-associated signalling pathways within the selected target panel. Consensus docking demonstrated recurrent multi-target enrichment of Pestalamide B, Kaempferol, Apigenin, Dihydroxyisoflavone and Lasiocarpine across multiple oncogenic proteins. ADMET-guided translational filtering subsequently prioritised Pestalamide B, Kaempferol and Apigenin for detailed downstream analysis. Molecular-dynamics simulations demonstrated stable ligand occupancy and conformational behaviour across all three prioritised complexes, with AKT1–Pestalamide B exhibiting the strongest overall dynamic stability. Pharmacophore analysis identified conserved donor–acceptor–hydrophobic feature architectures shared across structurally diverse phytochemicals. Orthogonal machine-learning analysis demonstrated selective convergence with docking-derived prioritisation, particularly for PI3K, PD-L1 and PR-associated targets, supporting the robustness of the multi-layer prioritisation framework.

Conclusions. The present study establishes the first integrated systems-level computational evaluation of *H. indicum* phytochemicals against subtype-relevant breast-cancer-associated targets. Pestalamide B, Kaempferol, Apigenin and Dihydroxyisoflavone emerged as the most promising multi-target phytochemical leads through convergent evidence from transcriptomics, network pharmacology, consensus docking, translational ADMET filtering, molecular dynamics and orthogonal machine-learning validation. These findings provide a biologically prioritised foundation for future biochemical, cellular and preclinical investigation of *H. indicum*-derived anti-breast-cancer compounds.

Keywords: *Heliotropium indicum; breast cancer; network pharmacology; consensus docking; molecular dynamics; pharmacophore modelling; ADMET; machine learning; polypharmacology; natural products.*

1. Introduction

Breast cancer (BC) is the most commonly diagnosed cancer in women worldwide and remains the leading cause of cancer-related mortality, with an estimated 2.3 million new cases and 685,000 deaths annually (Sung et al., 2021). Although precision oncology has transformed clinical management of hormone-receptor-positive and HER2-amplified disease, molecular heterogeneity, dynamic signalling plasticity, and subtype-specific therapeutic vulnerabilities continue to limit cure rates. The four major molecular subtypes - luminal A, luminal B, HER2-enriched and basal-like / triple-negative (TNBC) - possess distinct genomic, transcriptomic and biochemical signatures that determine drug response (Perou et al., 2000; Sørlie et al., 2001). TNBC, lacking actionable ER, PR and HER2 expression, remains particularly recalcitrant, with limited targeted options and poor prognosis.

BC progression is driven not by isolated oncogenic events but by interconnected signalling networks involving receptor tyrosine kinases (EGFR, HER2), the PI3K/AKT/mTOR axis, cell-cycle regulators (CDK4/6), DNA-repair machinery (PARP1, Topo IIα), endocrine signalling (ERα, PR, Aromatase), apoptotic suppressors (BCL-2) and immune checkpoints (PD-L1) (Yardley, 2013). Single-target inhibitors frequently fail because compensatory pathway re-activation restores tumour viability; consequently, multi-target ('polypharmacological') strategies that simultaneously engage several oncogenic nodes have emerged as a rational counter-measure (Hu et al., 2017; Bolognesi & Cavalli, 2016).

Natural products have historically contributed disproportionately to anticancer drug discovery, with several clinically established chemotherapeutics including paclitaxel, vincristine, etoposide and topotecan originating from plant-derived secondary metabolites (Newman & Cragg, 2020). Unlike highly selective synthetic inhibitors, phytochemicals frequently exhibit polypharmacological behaviour through structurally adaptable multi-target interactions, an attribute particularly relevant for biologically heterogeneous malignancies such as breast cancer. *Heliotropium indicum* L. (Boraginaceae) is a medicinal herb widely used in traditional systems of medicine across Asia, Africa and Latin America for inflammatory disorders, wound healing and tumour-associated conditions. Phytochemical investigations have identified diverse bioactive constituents including pyrrolizidine alkaloids (lasiocarpine, lycopsamine, indicine and heliotrine derivatives), flavonoids (kaempferol, apigenin, isorhamnetin and dihydroxyisoflavone), phenolics, coumarins and terpenoid-associated metabolites (Roeder, 2000; Mandal et al., 2014). Several individual compounds have demonstrated anti-proliferative or apoptosis-associated activity in experimental cancer models (Bhattacharyya & Saha, 2013; Reddy et al., 2008); however, a systems-level computational evaluation of the broader *H. indicum* phytochemical space against clinically relevant breast-cancer-associated targets has not previously been reported.

Contemporary computational drug-discovery workflows increasingly emphasise integration of multiple orthogonal analytical layers rather than reliance on isolated docking-based screening. Transcriptomic target validation, network pharmacology, consensus docking, pharmacokinetic prioritisation, molecular-dynamics simulation and machine-learning-assisted prediction each address distinct methodological limitations including target-selection bias, scoring-function dependence, structural instability and translational infeasibility (Pinzi & Rastelli, 2019; Vamathevan et al., 2019). Integration of these complementary approaches therefore enables more biologically informed and mechanistically interpretable phytochemical prioritisation compared with conventional single-layer computational screening strategies.

In the present study, we implemented an integrated systems-level computational framework to investigate the anti-breast-cancer potential of *H. indicum* phytochemicals. A curated library of 95 drug-like compounds was generated from 1,365 phytochemical records compiled across multiple natural-product databases and

literature resources. A subtype-relevant 13-target breast-cancer panel encompassing endocrine, receptor tyrosine kinase, immune-checkpoint, apoptosis and PI3K–AKT-associated signalling proteins was subsequently evaluated using TCGA-BRCA transcriptomic analysis, survival profiling and network-pharmacology prioritisation. Multi-engine consensus docking integrating AutoDock Vina, AutoDock4 and GNINA-CNN was combined with pharmacophore extraction, multi-target enrichment analysis and translational ADMET filtering to identify structurally and pharmacologically prioritised phytochemicals. Selected complexes subsequently underwent extended molecular-dynamics simulation and structural stability analysis, while orthogonal machine-learning classifiers trained on ChEMBL-derived datasets were employed as an independent validation layer for compound–target prioritisation. Through integration of these complementary computational approaches, the study identifies a focused set of biologically and translationally prioritised *H. indicum* phytochemicals that warrant further experimental evaluation against breast cancer.

2. Materials and Methods

2.1 Compound library construction

Phytochemical mining was performed across PubMed, Google Scholar and Scopus (search terms: "Heliotropium indicum" AND (phytochemical OR alkaloid OR flavonoid OR terpenoid OR phenolic), 2000–2025) and the phytochemical databases Dr. Duke's Phytochemical and Ethnobotanical, IMPPAT 2.0, NPACT, KNApSAcK, FooDB and COCONUT. Chemical metadata, SMILES, InChI keys and 3D SDF structures were retrieved from PubChem, ChEBI and ChemSpider. The initial unfiltered dataset comprised 1,365 records. Preprocessing in RDKit v2024.03 within Python performed canonical-SMILES standardisation and InChI-key deduplication, removed entries with invalid stereochemistry or ambiguous structures, and validated valence states. Drug-likeness filtering applied Lipinski's Rule of Five (MW < 500, LogP < 5, HBD ≤ 5, HBA ≤ 10), Veber (rotatable bonds ≤ 10, TPSA ≤ 140), Ghose and Egan rules; compounds passing ≥ 3 of 4 rule sets were retained. PAINS alerts were filtered using the RDKit FilterCatalog. After all filters, 95 drug-like phytochemicals (≈7 % of the raw set) were carried forward, comprising terpenoids and steroid-like compounds (~41 %), pyrrolizidine alkaloids (~29 %), phenolics (~11 %), miscellaneous metabolites (~9 %), flavonoids (~6 %) and fatty acids/esters (~3 %). Three-dimensional coordinates were generated by ETKDG (Riniker & Landrum, 2015) and energy-minimised in MMFF94 with 200 iterations. The final library is provided in Supplementary Table S1.

2.2 Target panel and transcriptomic validation

A 13-target panel was assembled to span all major BC subtypes: ERα (PDB 3ERT), PR (1A28), HER2/ErbB2 (3PP0), EGFR (1M17), PARP1 (4ZZZ), PD-L1 (5JDS), Topo IIα (4FM9), PI3Kα (4JPS), AKT1 (4EKL), CDK4/6 (5FGK), BCL-2 (6GL8), mTOR (4DRH) and Aromatase / CYP19A1 (3S79). Crystal structures were prepared in UCSF Chimera and PyMOL: water and co-crystallised ligands removed (catalytic cofactors retained), missing residues modelled in MODELLER (gaps < 15 residues) or AlphaFold2 hybrid models for longer gaps; protonation assigned at pH 7.4 with PROPKA 3.0; and brief minimisation in GROMACS to relieve clashes. Binding sites were validated using DoGSiteScorer and fpocket.

TCGA-BRCA RNA-seq Level-3 RSEM-normalised counts (1,100 tumour vs 113 normal samples) were retrieved from the GDC portal. Differential expression of the 13 target genes was computed in DESeq2 v1.42 with Benjamini–Hochberg FDR control. Subtype-stratified expression analyses were performed for luminal A, luminal B, HER2-enriched and basal-like / TNBC tumours. Kaplan–Meier survival analysis for each target gene used GEPIA2, with log-rank p-values and hazard ratios reported. Protein-level expression was cross-checked against Human Protein Atlas IHC stainings.

2.3 Network pharmacology

Compound targets were predicted in Swiss Target Prediction, SEA, BindingDB and PharmMapper (probability cutoff > 0.7). Disease targets were retrieved from GeneCards (relevance > 5), DisGeNET (score > 0.3), OMIM, TTD and DrugBank. The compound-predicted set was intersected with the disease-associated set; the overlap defined the prioritised target list. A protein–protein interaction (PPI) network was constructed in STRING (confidence > 0.7) and analysed in Cytoscape 3.10 with NetworkAnalyzer (degree, betweenness, closeness centrality). Hub genes were identified by CytoHubba using five algorithms (MCC, DMNC, MNC, Degree, BottleNeck); genes ranked highly in ≥ 3 algorithms were designated core hubs. Gene Ontology biological-process and KEGG pathway enrichments were computed in DAVID and ShinyGO with adjusted p-values. A multi-layer compound-target-pathway-disease network was constructed in Cytoscape (compounds green, targets red, pathways blue, disease yellow).

2.4 Consensus molecular docking

Each of the 95 phytochemicals was docked into all 13 targets using four orthogonal engines: AutoDock Vina 1.2.5 (empirical scoring, iterated local search), AutoDock4 (Lamarckian genetic algorithm, semi-empirical free-energy scoring), GNINA 1.0 (CNN-based scoring trained on experimentally validated complexes) and PLANTS (ant colony optimisation with ChemPLP). Receptors, ligands were prepared using our in house software MolCura (BioCogniz). Grid boxes (20 × 20 × 20 Å) were centred on experimentally validated active-site residues; AutoDock4 grid spacing was 0.375 Å. Z-score normalisation across the four engines preceded rank averaging; compounds in the top 10 % of ≥ 2 of 4 programs were classified consensus hits. The four-engine design - exceeding the typical 2–3-tool standard - was specifically chosen to neutralise scoring-function bias.

Validation employed three orthogonal procedures. (i) Re-docking RMSD: for each target, the co-crystallised native ligand was extracted and re-docked into its original binding pocket under identical docking conditions. Structural agreement between experimental and predicted poses was quantified using RMSD, with values below 2.0 Å considered acceptable for reliable pose recovery.

(ii) Decoy enrichment validation: property-matched inactive decoys inspired by DUD-E benchmarking principles were docked alongside candidate phytochemicals and reference ligands to evaluate discriminatory performance of the consensus docking framework. Comparative enrichment behaviour within predefined high-affinity docking ranges was subsequently analysed.

(iii) FDA-approved drug benchmarking: tamoxifen (ERα), lapatinib (HER2), olaparib (PARP1), alpelisib (PI3Kα), palbociclib (CDK4/6), venetoclax (BCL-2), capivasertib (AKT1), letrozole (Aromatase), megestrol (PR), osimertinib (EGFR), torin1 (mTOR), BMS-202 (PD-L1) and etoposide (TOP2A) were docked under identical conditions as positive-control reference ligands .

Protein–ligand interactions were profiled using PLIP and Discovery Studio Visualizer to identify hydrogen bonds, hydrophobic interactions, π-stacking interactions and salt bridges. Three-dimensional pharmacophore models were generated from top-ranked consensus poses using PharmaGist to identify conserved hydrogen-bond donor, hydrogen-bond acceptor, hydrophobic and aromatic pharmacophoric features associated with multi-target binding behaviour.

2.5 ADMET consensus prediction

Lead compounds were profiled across four complementary ADMET platforms: SwissADME for physicochemical descriptors, drug-likeness filters, gastrointestinal absorption, BBB permeability, P-glycoprotein interaction, CYP inhibition and bioavailability radar assessment (Daina et al., 2017); pkCSM for

absorption, distribution, metabolism, excretion and toxicity prediction including intestinal absorption, Caco-2 permeability, VDss, CYP substrate/inhibitor prediction, total clearance, AMES mutagenicity, hERG inhibition, hepatotoxicity, oral rat LD50 and LOAEL estimation (Pires et al., 2015); ADMETlab 3.0 for large-scale multi-endpoint pharmacokinetic and toxicity profiling with confidence scoring (Fu et al., 2024); and ProTox for organ toxicity, mutagenicity, carcinogenicity, Tox21 pathway toxicity and LD50 class prediction (Banerjee et al., 2018).

A consensus ADMET matrix integrating key pharmacokinetic and toxicity endpoints was subsequently constructed for prioritised compounds. For each endpoint, predictions were comparatively evaluated across the available platforms, and concordant favourable predictions across multiple tools were considered indicative of increased confidence. Compounds demonstrating broad agreement across major safety and pharmacokinetic parameters were prioritised as pharmacokinetically favourable leads for downstream molecular-dynamics evaluation.

2.6 Molecular dynamics simulation and trajectory analysis

Three translationally prioritised complexes - CDK4–Kaempferol, AKT1–Pestalamide B and Aromatase–Apigenin - were subjected to all-atom molecular-dynamics (MD) simulation using GROMACS 2019.4 following consensus docking and integrated ADMET filtering. Protein topologies for CDK4 and AKT1 were generated using the AMBER99SB-ILDN force field (Lindorff-Larsen et al., 2010), whereas Aromatase employed the CHARMM27 force field to enable validated all-atom HEME residue parameterisation with explicit Fe–N coordination suitable for haem-containing metalloenzymes (MacKerell et al., 1998; Autenrieth et al., 2004). Ligand topologies were generated using GAFF2 through ACPYPE 2023.10 with AM1-BCC partial charges (Sousa da Silva & Vranken, 2012). The native haem cofactor was retained within the Aromatase catalytic cavity throughout simulation setup and parameterisation.

Each protein–ligand complex was solvated in a TIP3P water-filled rhombic dodecahedron simulation box with 1.0 nm boundary padding and neutralised with 0.15 M NaCl using gmx genion. Final solvated systems contained approximately 40,000–68,000 atoms depending on protein size and cofactor composition. Energy minimisation was performed using the steepest-descent algorithm until the maximum force converged below 1,000 kJ mol^{-1} nm^{-1} or 50,000 steps were completed. Long-range electrostatics were treated using the Particle Mesh Ewald (PME) method with a 1.0 nm cutoff and 0.16 nm PME grid spacing.

Systems subsequently underwent 100 ps NVT equilibration at 300 K using the velocity-rescale thermostat (Bussi et al., 2007), followed by 100 ps NPT equilibration at 1 bar using the Berendsen barostat with positional restraints retained on heavy atoms. Production simulations were performed for 200 ns under NPT conditions using the Parrinello–Rahman barostat (Parrinello & Rahman, 1981), leap-frog integrator with a 2 fs timestep, LINCS bond constraints and PME electrostatics. Trajectory coordinates were recorded every 10 ps. Simulations were executed on NVIDIA RTX 3090 and RTX 2070 SUPER GPU platforms using GROMACS 2019.4 (Abraham et al., 2015). Periodic-boundary-condition corrections and trajectory centring were performed using gmx trjconv -center -pbc mol -ur compact.

Post-simulation analyses were performed on periodic-boundary-condition (PBC)-corrected trajectories using standard GROMACS analysis utilities. Backbone root-mean-square deviation (RMSD) was calculated relative to the energy-minimised starting structure following least-squares backbone fitting to assess overall structural stability throughout the 200-ns production trajectories. Residue-level flexibility was evaluated using per-residue Cα root-mean-square fluctuation (RMSF) analysis, while global protein compactness was monitored through radius of gyration (Rg) calculations. Protein–ligand hydrogen-bond occupancy was

determined using a donor–acceptor distance cutoff of 3.5 Å and donor–H–acceptor angle cutoff of 30°. Solvent-accessible surface area (SASA) calculations were additionally performed to evaluate changes in conformational exposure and structural packing during simulation. Binding free energies were subsequently estimated using MM-GBSA and MM-PBSA calculations implemented through the gmx_MMPBSA package on equilibrated trajectory windows.

2.7 Machine-learning bioactivity prediction

Target-specific binary classifiers were trained for each of the 13 targets on ChEMBL (release 33) bioactivity data: IC50 / Ki / EC50 / Kd records were retrieved by ChEMBL target ID and binarised with a 10-μM activity threshold. Class imbalance was addressed with SMOTE oversampling. Molecular features comprised Morgan fingerprints (radius 2, 2,048 bits) plus eight RDKit descriptors (MW, LogP, HBA, HBD, TPSA, rotatable bonds, ring count, fraction Csp3) for 2,056 features. Three algorithms were trained per target - Random Forest (200 estimators, balanced class weights), XGBoost (200 trees, lr = 0.1, max-depth 6) and SVM (RBF kernel, balanced weights, calibrated probabilities) - within scikit-learn pipelines (SMOTE → StandardScaler → classifier). An 80/20 stratified train/test split with 5-fold stratified cross-validation reported accuracy, F1, precision, recall and AUROC. SHAP (TreeExplainer for RF/XGB) provided interpretability. The 39 trained classifiers were applied to the 95-compound H. indicum library to predict per-target activity; compounds with predicted probability ≥ 0.7 were classified high-confidence ML hits. The intersection of consensus docking top-10s and ML high-confidence hits per target identified the joint shortlist.

2.8 Statistical analysis and data availability

Continuous values are reported as mean ± standard deviation unless stated. Differential-expression p-values were FDR-corrected by Benjamini–Hochberg. Plot generation used matplotlib 3.10 and seaborn. All raw data and analysis scripts are provided in the supplementary repository.

3. Results

3.1 Compound library and physicochemical landscape

Comprehensive phytochemical mining of *H. indicum* yielded 1,365 raw records - one of the most comprehensive computationally curated libraries reported for this medicinal species. Sequential canonicalisation, deduplication, drug-likeness filtering (Lipinski / Veber / Ghose / Egan, ≥ 3-of-4 pass) and PAINS removal retained 95 drug-like compounds (~6.96 %) spanning seven major chemical classes, dominated by terpenoids and steroid-like scaffolds (~41 %) and pyrrolizidine alkaloids (~29 %), with flavonoids, phenolics, coumarins, fatty acids and miscellaneous secondary metabolites comprising the remainder (Supplementary Table S1). Mean physicochemical properties were within Lipinski-acceptable ranges (mean MW 318 Da, mean LogP 2.1, mean TPSA 95 Å², mean rotatable bonds 4) (Figure 1). The retention of both rigid hydrophobic terpenoid frameworks and hydrogen-bond-rich aromatic flavonoid scaffolds is favourable for polypharmacology because complementary interaction modes increase the probability of multi-target binding (Hu et al., 2017).

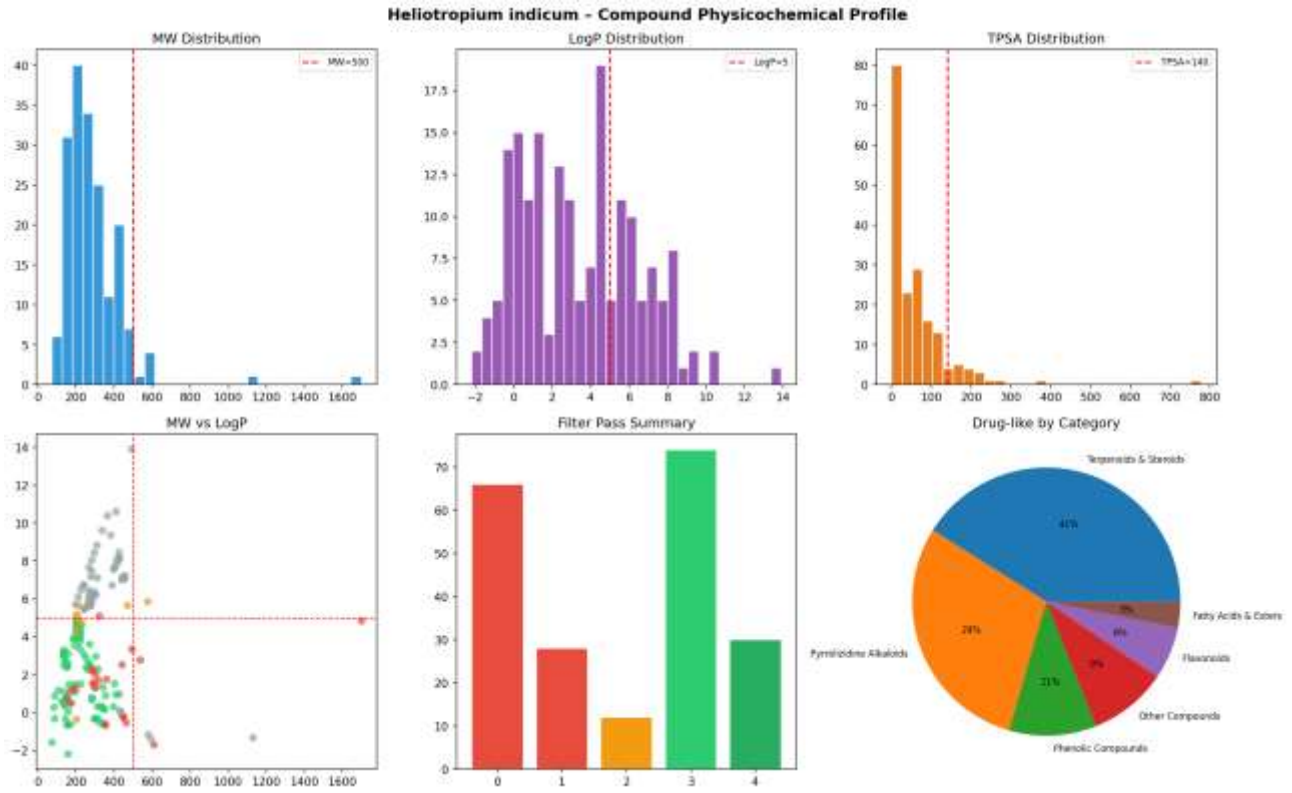

Figure 1. Physicochemical and chemical-space profiling of the curated *Heliotropium indicum* phytochemical library.

Distribution of molecular weight (MW), LogP and topological polar surface area (TPSA) of retained compounds following drug-likeness filtering. Scatter analysis of MW versus LogP illustrates compound occupancy within acceptable drug-like chemical space. Filter-pass summary represents the number of compounds satisfying Lipinski, Veber, Ghose and Egan criteria. Pie-chart analysis shows chemical-class composition of the final 95-compound library, dominated by terpenoid/steroid-like scaffolds and pyrrolizidine alkaloids.

3.2 Target validation: TCGA-BRCA expression and survival

Differential-expression profiling across TCGA-BRCA samples demonstrated heterogeneous dysregulation patterns among the selected therapeutic targets (Figure 2; Supplementary Table S2). ESR1, TOP2A, PARP1 and ERBB2 exhibited prominent tumour-associated expression changes, whereas EGFR, BCL2, CYP19A1 and MTOR displayed comparatively reduced or subtype-dependent expression patterns. Hierarchical clustering further revealed clear segregation of luminal-associated markers (ESR1, PGR, BCL2) from proliferative and DNA-repair-associated targets such as TOP2A and PARP1, supporting the biological diversity of the selected panel. Subtype annotation indicated representation across luminal, HER2-positive and TNBC-associated contexts, enabling broader therapeutic coverage rather than restriction to a single breast-cancer subtype.

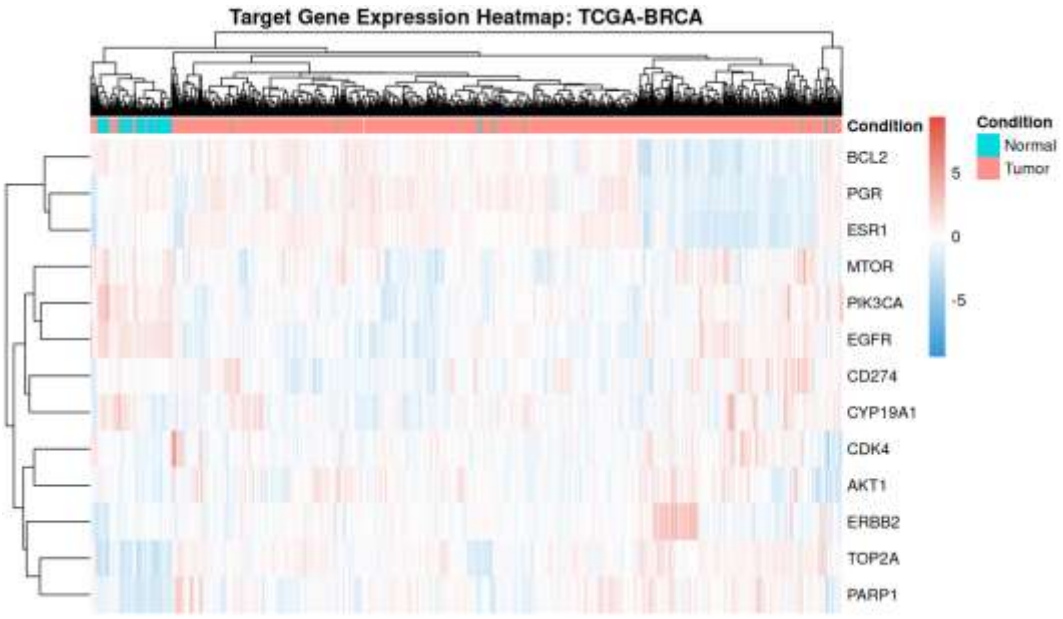

Figure 2. Hierarchical clustering heatmap of TCGA-BRCA expression profiles for the 13 prioritised therapeutic targets across tumour and matched normal breast-tissue samples. Red indicates relative upregulation and blue indicates relative downregulation. Luminal-associated markers (ESR1, PGR, BCL2) clustered separately from proliferative and DNA-repair-associated targets including TOP2A and PARP1, reflecting subtype-dependent transcriptional organisation within breast cancer.

Kaplan–Meier survival analysis identified significant prognostic association for PIK3CA expression (log-rank $p = 0.0082$; Figure 3), where elevated expression correlated with reduced overall survival probability. Several additional targets, including CYP19A1 and MTOR, demonstrated borderline survival trends, whereas the remaining targets showed limited or non-significant prognostic separation within the analysed cohort. Collectively, these findings support the clinical and biological relevance of the selected target panel despite heterogeneous prognostic contributions across individual genes (Supplementary Figure S1).

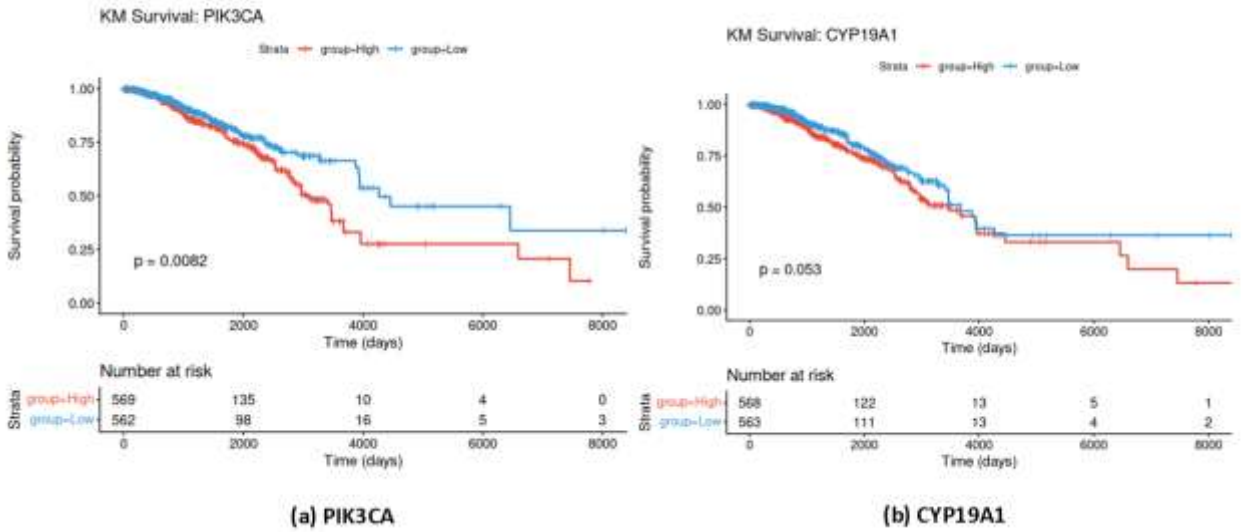

Figure 3. Kaplan–Meier overall-survival analysis of representative therapeutic targets across the TCGA-BRCA cohort following stratification into high- and low-expression patient groups based on median transcript abundance. Survival probabilities were compared using the log-rank test. (a) PIK3CA expression demonstrated a statistically significant association with overall survival (log-rank p = 0.0082), with the high-expression group exhibiting reduced long-term survival probability relative to the low-expression cohort. (b) CYP19A1 (Aromatase) showed a near-significant or borderline prognostic trend (log-rank p = 0.053), suggesting potential subtype-dependent clinical relevance. Red curves represent the high-expression group, whereas blue curves represent the low-expression group. Corresponding patient numbers at risk are shown below each survival curve. These analyses provided transcriptomic and prognostic support for inclusion of PIK3CA and CYP19A1 within the downstream multi-target computational framework.

3.3 Network-pharmacology landscape

Compound-target prediction integrating SwissTargetPrediction, SEA, PharmMapper and BindingDB identified multiple putative human protein targets associated with the *H. indicum* phytochemical library. Intersection analysis with curated breast-cancer-associated genes retrieved 19 overlapping therapeutic targets shared between compound-predicted and disease-associated datasets (Figure 4). Further prioritisation against Phase IV / clinically validated breast-cancer targets identified a core subset of 7 therapeutically relevant proteins, supporting translational relevance of the network-pharmacology framework.

STRING protein–protein interaction analysis (confidence score > 0.7) revealed a densely interconnected signalling network enriched in oncogenic and survival-associated pathways (Figure 5). CytoHubba consensus ranking across five independent topological algorithms (Degree, Betweenness, Closeness, MCC and DMNC) consistently identified AKT1, CDK4, PIK3CA, ERBB2, MTOR, ESR1, BCL2 and PARP1 as dominant hub genes, with AKT1 exhibiting the highest interaction degree (n = 9), followed by PIK3CA and CDK4 (n = 8 each). EGFR and CD274 also retained high multi-algorithm support, whereas TOP2A and PGR demonstrated comparatively lower network centrality (Supplementary Figure S2).

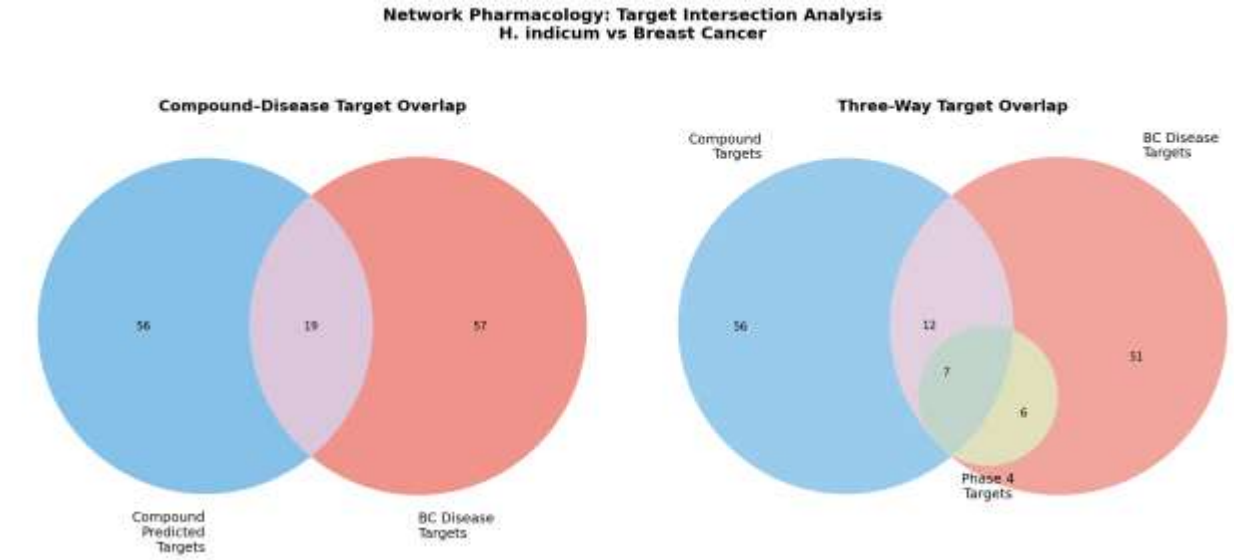

Figure 4. Network-pharmacology target intersection analysis between predicted phytochemical targets and curated breast-cancer-associated genes. Left: overlap between compound-predicted and breast-cancer-associated target datasets. Right: three-way overlap highlighting clinically validated Phase IV therapeutic targets retained after prioritisation.

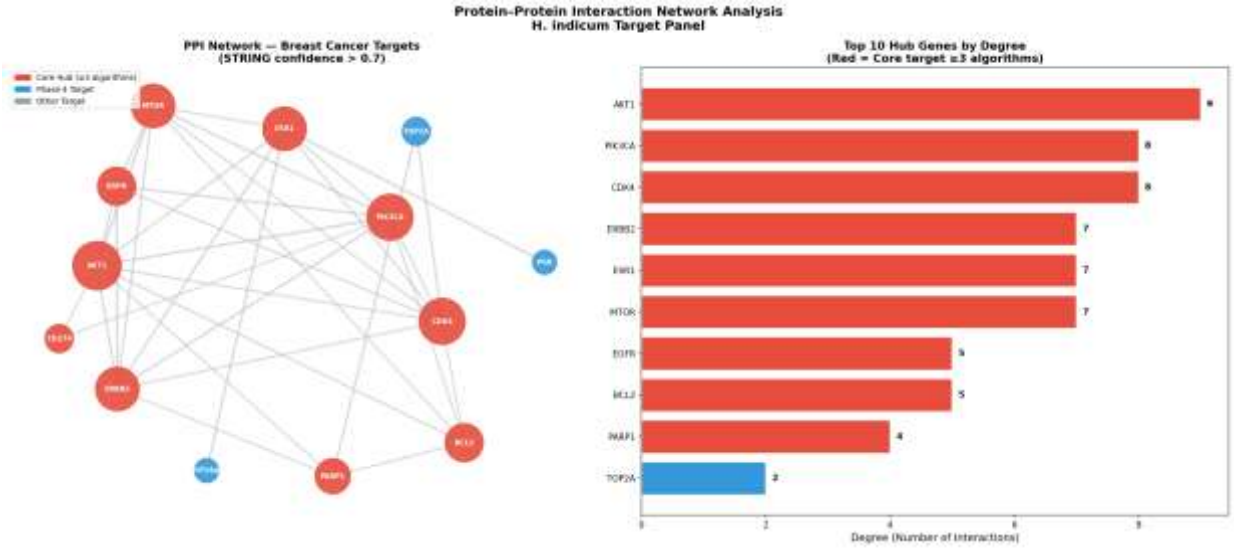

Figure 5. Protein–protein interaction (PPI) network and hub-gene prioritisation of breast-cancer-associated targets identified from *H. indicum* phytochemicals. STRING network analysis (confidence > 0.7) revealed extensive functional connectivity among oncogenic signalling proteins. CytoHubba consensus ranking identified AKT1, PIK3CA, CDK4, ERBB2 and MTOR as dominant hub genes based on multi-algorithm topological centrality analysis.

Functional enrichment analysis demonstrated highly significant overrepresentation of cancer-associated signalling pathways (Figure 6). KEGG enrichment prominently identified PI3K–AKT signalling, breast-cancer signalling, oestrogen signalling, proteoglycans in cancer, AGE–RAGE signalling, ErbB signalling and apoptosis-related pathways among the most enriched biological processes. Gene Ontology analysis further revealed strong enrichment for regulation of apoptotic processes, receptor tyrosine kinase signalling, oxidative-stress response, protein phosphorylation and intracellular signal transduction. Molecular-function enrichment highlighted kinase activity, protein tyrosine kinase activity, ubiquitin-protein ligase binding and ATPase binding, whereas cellular-component analysis associated the targets with membrane rafts, nuclear membrane structures, intracellular organelle lumen and endosomal compartments.

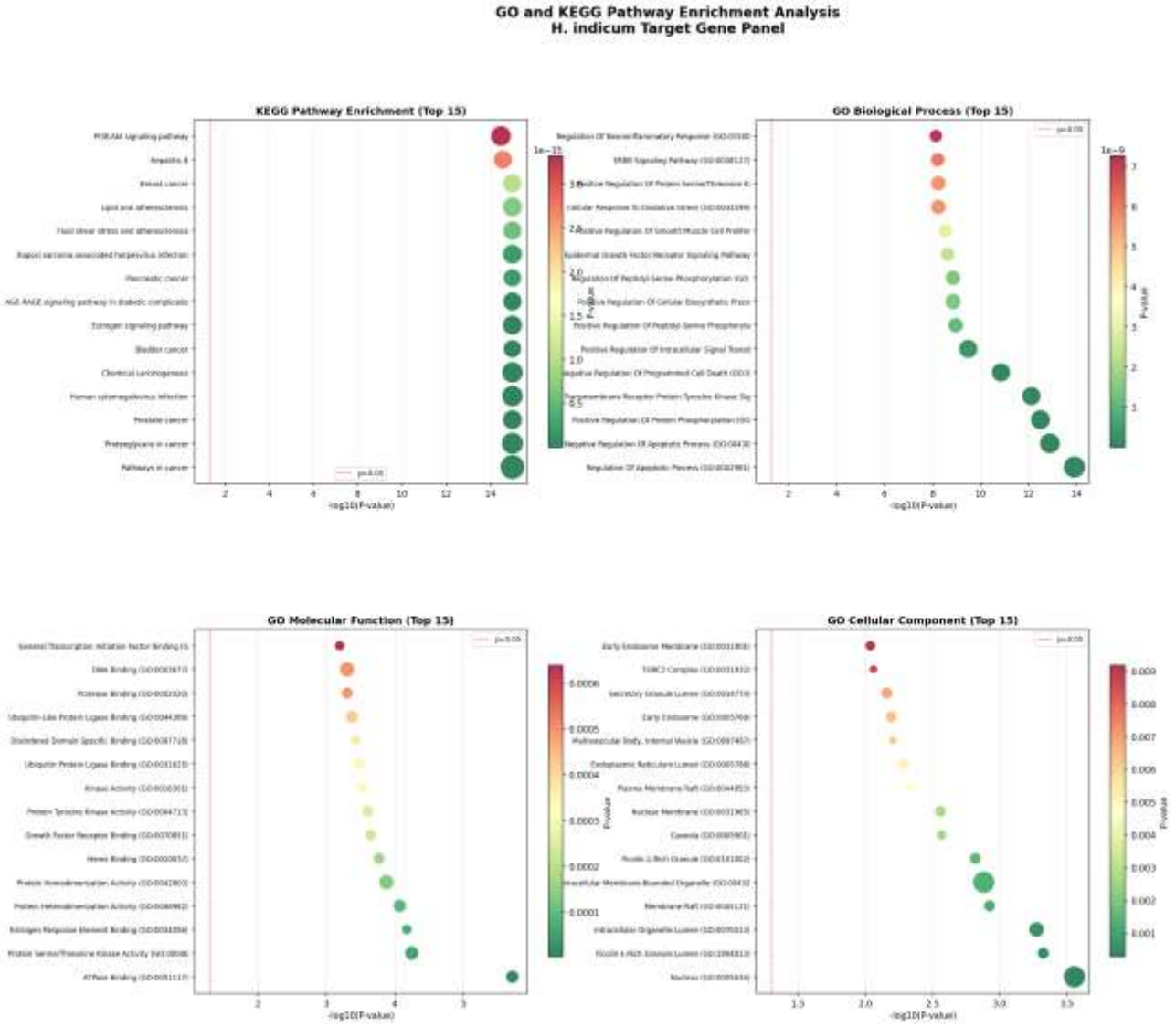

Figure 6. GO and KEGG pathway enrichment analysis of prioritised breast-cancer-associated targets. KEGG analysis identified significant enrichment of PI3K–AKT, oestrogen, ErbB and cancer-associated signalling pathways. Gene Ontology analysis highlighted enrichment for apoptotic regulation, receptor tyrosine kinase signalling, oxidative-stress response, kinase activity and intracellular signalling-associated cellular compartments.

The multilayer compound–target–pathway–disease interaction network demonstrated extensive polypharmacological connectivity between major phytochemicals and breast-cancer-associated pathways

(Figure 7). Kaempferol and apigenin occupied high-degree positions linking multiple oncogenic targets including AKT1, CDK4, PIK3CA, ERBB2 and MTOR, whereas allantoin displayed broader pathway connectivity through inflammatory and stress-associated signalling components. The integrated network architecture supports the mechanistic basis for the multi-target docking behaviour observed in subsequent structure-based analyses.

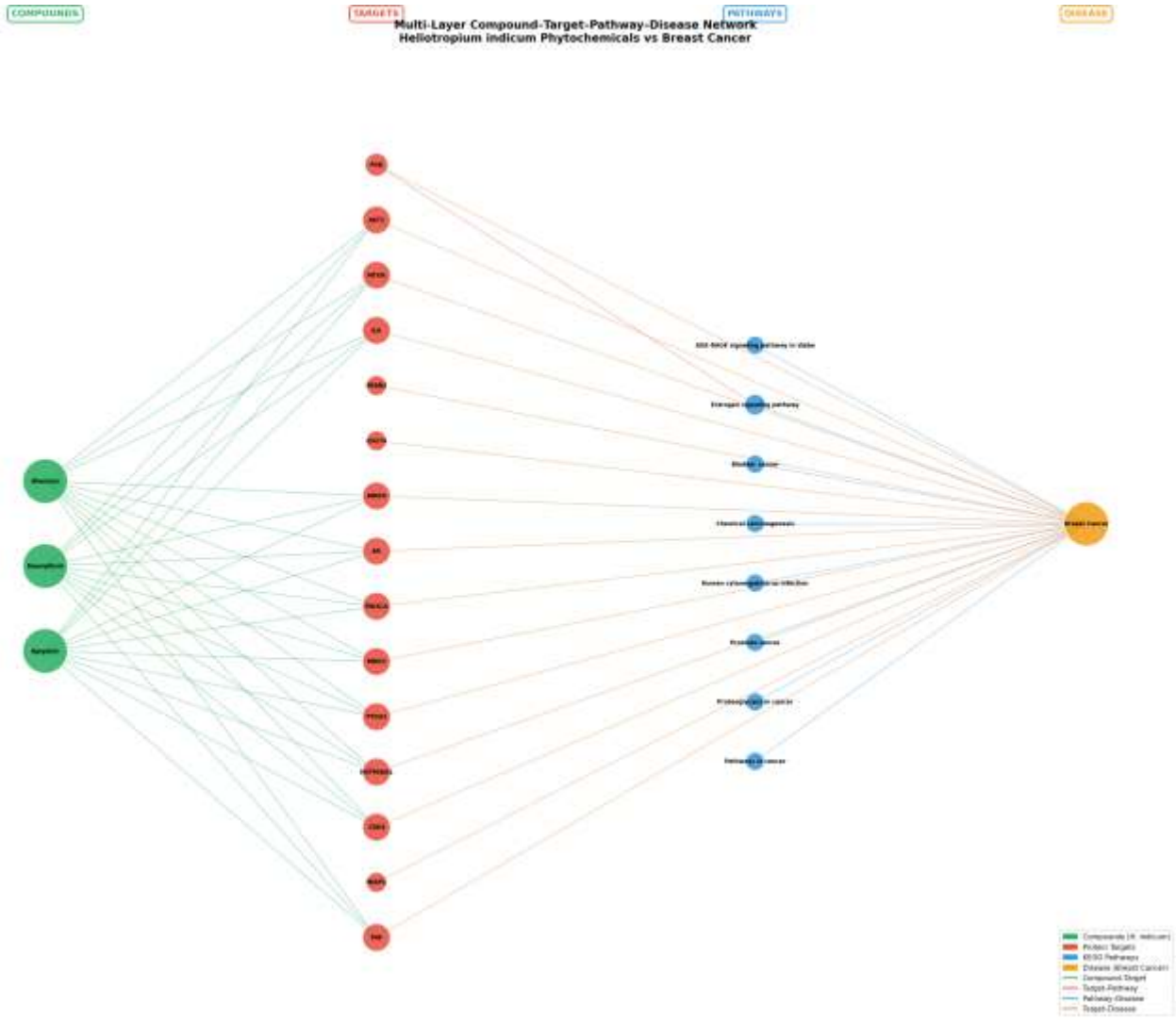

Figure 7. Multi-layer compound–target–pathway–disease interaction network illustrating relationships between major *H. indicum* phytochemicals, breast-cancer-associated protein targets, enriched KEGG pathways and breast cancer. Kaempferol and apigenin demonstrated extensive multi-target connectivity across oncogenic signalling pathways.

3.4 Consensus docking, decoy validation and FDA benchmarking

Re-docking RMSD validation across all 13 targets achieved sub-2 Å recovery of crystallographic poses across the evaluated docking engines, with a mean best-pose RMSD of 0.205 ± 0.221 Å (best pose across AutoDock Vina, AutoDock4 and GNINA, averaged across the 13 targets) and 100 % of targets remaining below the 2 Å acceptance threshold across all three engines, confirming highly reliable reproduction of experimentally observed binding modes (Supplementary Figure S3). Consensus-based validation further demonstrated favourable separation between FDA-reference ligands and target-specific decoys across the

majority of therapeutic targets, supporting the discriminatory capability of the multi-engine docking framework (Supplementary Table S3). FDA-approved reference ligands consistently achieved competitive consensus docking scores and served as calibration anchors for phytochemical prioritization (Supplementary Table S4). Comparative enrichment analysis further demonstrated that the curated phytochemical candidates were substantially overrepresented within the high-affinity binding range (≤ −7.0 kcal/mol), with several compounds extending into stronger affinity regions below −7.5 and −8.0 kcal/mol relative to property-matched decoys (Supplementary Figure S4). The enrichment profiles confirmed effective discrimination between candidate phytochemicals and inactive decoy sets, supporting the robustness and selectivity of the multi-engine consensus docking framework (Supplementary Figure S4).

Consensus-ranking analysis of the top-scoring compound–target interaction pairs integrating GNINA, AutoDock Vina, AutoDock4 and PLANTS scores identified recurrent enrichment of Pestalamide B, Dihydroxyisoflavone, Kaempferol, Apigenin and Lasiocarpine across multiple therapeutic targets (Figure 8,9). Complementary multi-target enrichment analysis further demonstrated that several phytochemicals repeatedly appeared among the strongest consensus binders across distinct protein classes, supporting the presence of broad-spectrum target engagement patterns within the H. indicum phytochemical library. Final MD candidates were therefore selected not solely on docking rank, but through combined consideration of chemical diversity, multi-target representation, mechanistic relevance and downstream translational prioritization.

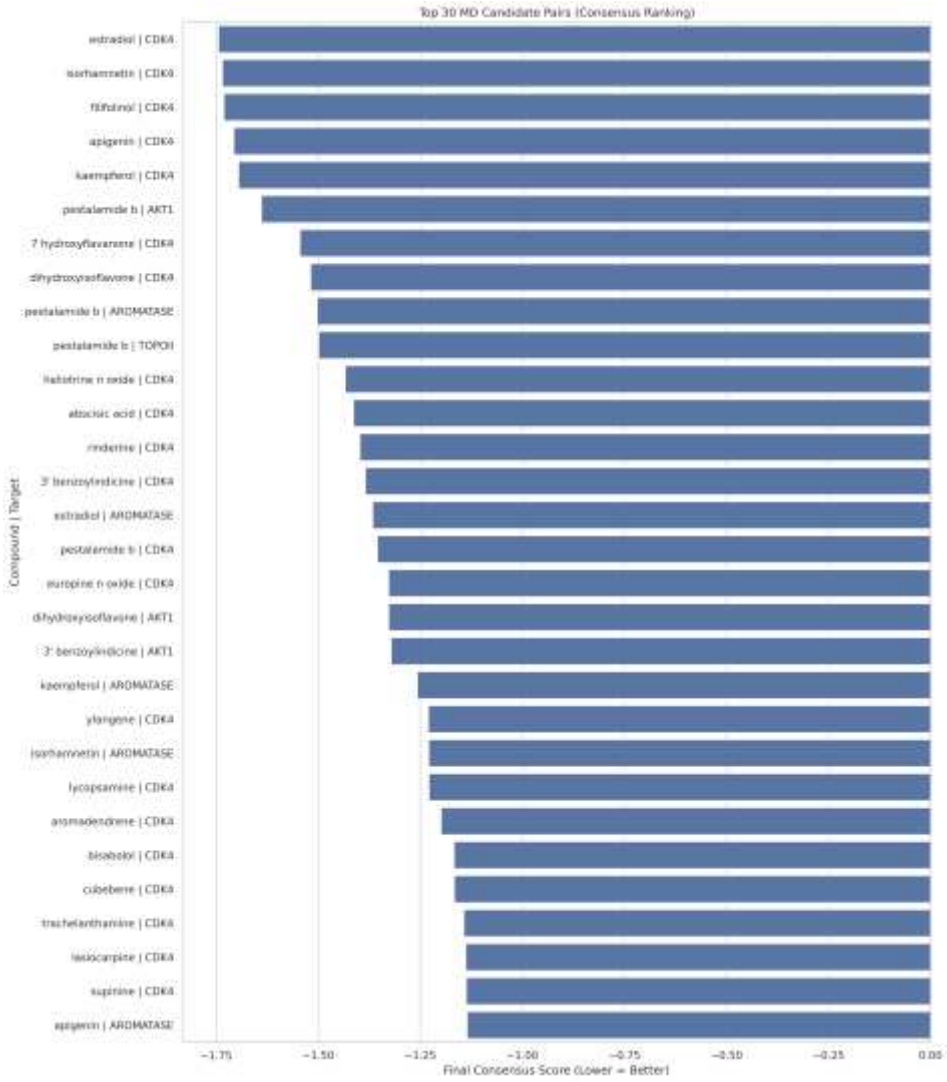

Figure 8. Top 30 consensus-ranked compound–target interaction pairs identified through integrated GNINA, AutoDock Vina, AutoDock4 and PLANTS consensus docking analysis.

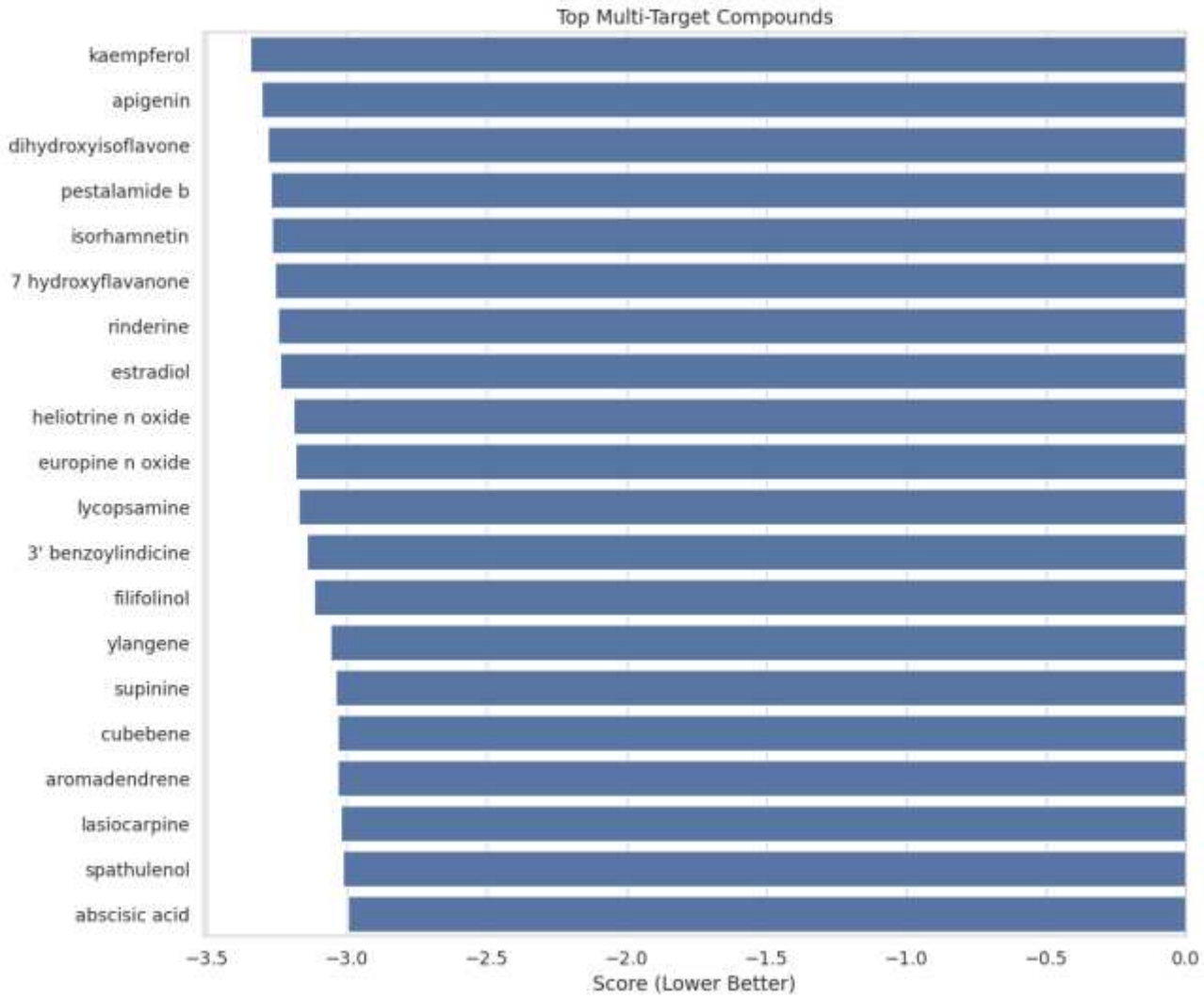

Figure 9. Multi-target enrichment analysis of the top-performing phytochemicals across the evaluated therapeutic targets. Compounds were ranked according to their cumulative multi-target consensus-docking performance derived from integrated GNINA, AutoDock Vina, AutoDock4 and PLANTS scoring.

PLIP interaction analysis idemonstrated biologically meaningful binding interactions across all selected protein–ligand complexes (Figure 10). In the AKT1–Pestalamide B complex, the selected Pose 3 established hinge-associated hydrogen bonding with Ala230 along with stabilizing contacts involving Glu234, Asp292, Phe161, Val164, and Ala177, supporting a plausible ATP-competitive kinase inhibition mechanism. The CDK4–Kaempferol and CDK4–Lasiocarpine complexes (Pose 2) occupied the ATP-binding cleft and interacted with catalytically relevant residues including Lys35, Thr19, Val20, Gly47, Leu49, Thr83, Val96, Gln98, and Tyr151. Among these, Lys35 and the Val96/Gln98 region are particularly important for ATP positioning, hinge-associated stabilization, and kinase pocket occupancy, indicating mechanistically relevant CDK4 inhibitory binding. In the Aromatase–Apigenin complex, Pose 2 occupied the catalytic steroid-binding cavity above the haem group and formed interactions with Arg115, Ile133, Phe134, Trp224, Thr310, Val370,

and Met374. These residues are known components of the substrate-access channel and catalytic pocket of Aromatase, supporting stable ligand accommodation and potential enzyme inhibition.

Among the nine docking poses generated for each complex, final pose selection was performed using combined assessment of GNINA binding affinity, CNN pose confidence score, RMSD clustering consistency, pocket occupancy, and biologically relevant interaction patterns. GNINA was selected as the primary docking environment due to its improved pose-ranking reliability through convolutional neural network (CNN)-based scoring integrated with conventional docking energy evaluation. Poses showing stable catalytic-pocket occupancy, hinge-region interactions, and consistent clustering with low RMSD deviation from top-ranked conformations were prioritized for downstream molecular dynamics simulations.

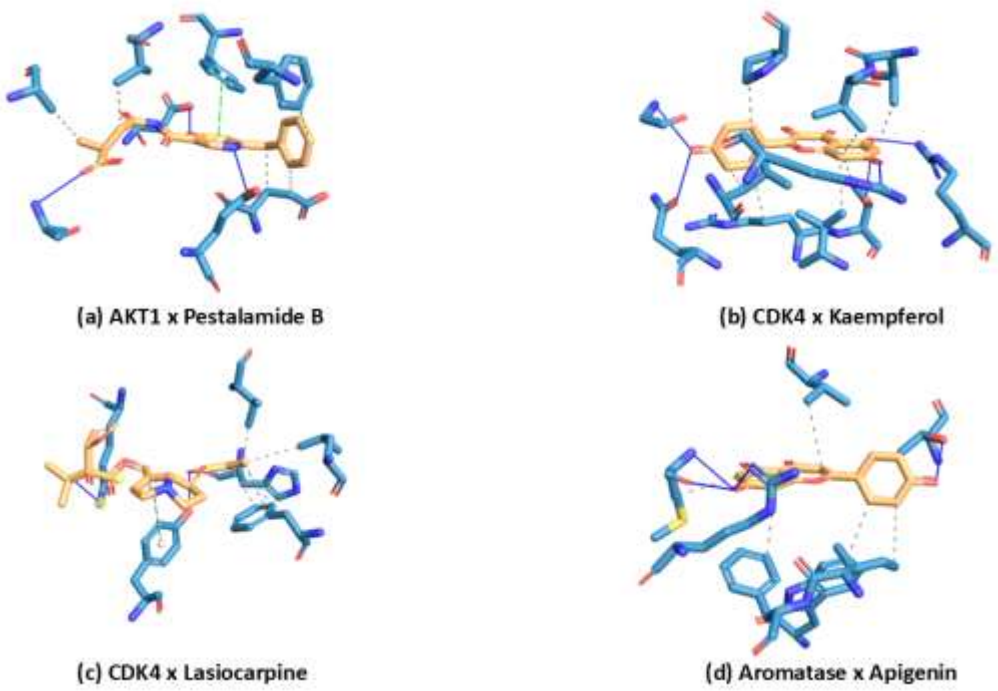

Figure 10. Final consensus-docking poses and PLIP-derived interaction profiles for the shortlisted protein–ligand complexes selected for downstream molecular dynamics evaluation. Protein binding-pocket residues are shown as stick representations (cyan), while ligands are shown in orange. Dashed lines indicate hydrogen-bonding and other non-covalent interactions identified by PLIP. Complexes include (a) AKT1–Pestalamide B, demonstrating stable ATP-pocket engagement within the kinase active site; (b) CDK4–Kaempferol, showing interactions within the canonical CDK4 ATP-binding cleft; (c) CDK4–Lasiocarpine, included as a comparative high-affinity docked complex prior to ADMET-based de-prioritisation; and (d) Aromatase–Apigenin, illustrating ligand accommodation within the haem-containing catalytic cavity. These interaction profiles formed the structural basis for selecting representative complexes for subsequent molecular dynamics simulations and MM-GBSA binding-energy analyses.

Three-dimensional pharmacophore models extracted from the top-ranked docking conformations identified conserved spatial arrangements of hydrogen-bond donor/acceptor centres, hydrophobic moieties and aromatic features across the prioritised phytochemicals (Figure 11). Pestalamide B exhibited a distributed donor–acceptor architecture linked with extended hydrophobic and aromatic regions, whereas Kaempferol and Apigenin displayed classical flavonoid-associated planar aromatic pharmacophore patterns enriched in hydrogen-bond acceptors. Lasiocarpine, despite its structurally distinct pyrrolizidine alkaloid scaffold, retained comparable hydrophobic and hydrogen-bonding feature organisation. The recurrent donor–acceptor–hydrophobic feature combinations observed across chemically diverse compounds support the structural consistency of the identified docking poses and provide a pharmacophore-level framework for subsequent interaction analysis and lead prioritisation.

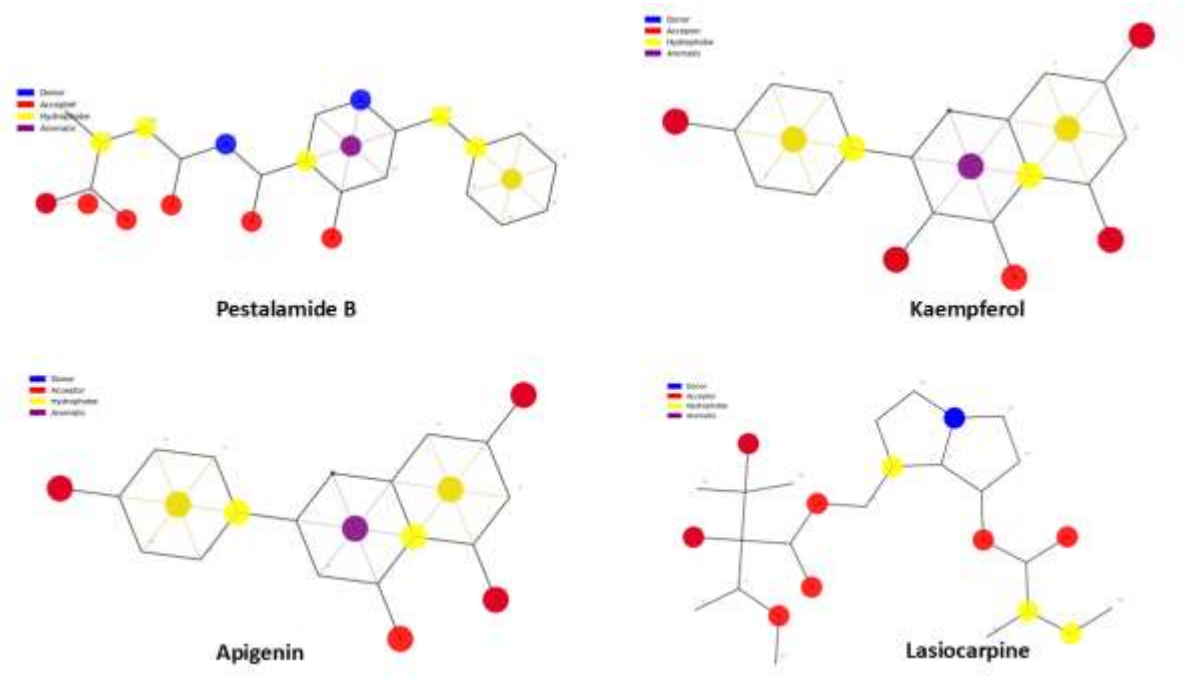

Figure 11. Three-dimensional pharmacophore models derived from the top-ranked docking conformations of prioritised phytochemicals. Pharmacophore extraction identified conserved spatial arrangements of hydrogen-bond donor, hydrogen-bond acceptor, hydrophobic and aromatic features across chemically diverse compounds including Pestalamide B, Kaempferol, Apigenin and Lasiocarpine.

3.5 Multi-tool ADMET consensus

Multi-platform ADMET assessment integrating SwissADME, pkCSM, ADMETlab 3.0 and ProTox predictions was performed for the four MD-prioritised phytochemicals (Figure 12). Comparative consensus analysis indicated that Pestalamide B exhibited the most balanced overall pharmacokinetic and toxicity profile across the evaluated computational platforms, with comparatively fewer predicted toxicity liabilities relative to the remaining compounds. Apigenin and Kaempferol also demonstrated generally favourable ADMET behaviour, although both compounds were associated with predicted AMES mutagenicity alerts within pkCSM-based assessment, consistent with previously reported flavonoid-associated mutagenicity warnings observed in computational toxicology frameworks.

Consensus ADMET scoring across the 19 evaluated endpoints with a ≥ 2/3 tool-agreement threshold yielded overall pass rates of 78.9 % for Pestalamide B (with 8 of 9 toxicity-specific endpoints classified favourable), 73.7 % for both Apigenin and Kaempferol (each flagged for AMES mutagenicity by ≥ 2 tools) and 73.7 % for Lasiocarpine, which was additionally flagged for carcinogenicity by ProTox-3 (Active, probability 0.92) and identified as a P-glycoprotein substrate by both SwissADME and pkCSM (Figure 12). Pestalamide B therefore emerged as the most pharmacokinetically favourable lead across the four-tool consensus.

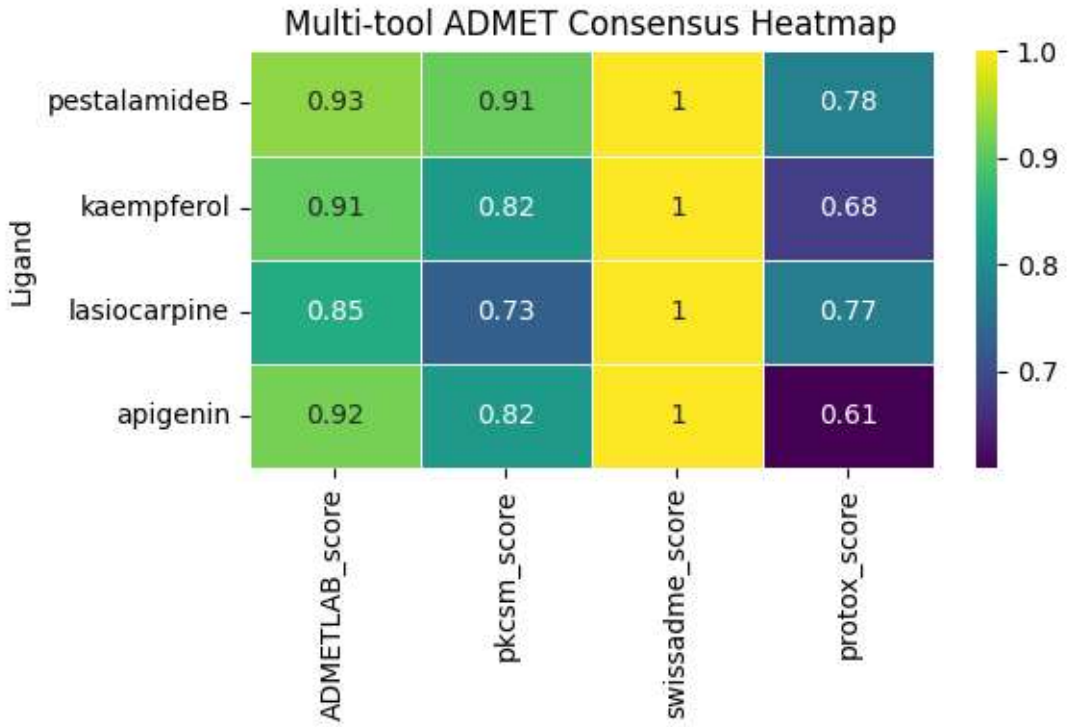

Figure 12. Multi-tool ADMET consensus heatmap for the four MD-prioritised phytochemicals. Heatmap values represent relative consensus favourability scores derived from SwissADME, pkCSM, ADMETlab 3.0 and ProTox predictions.

In contrast, Lasiocarpine displayed multiple predicted toxicity liabilities despite favourable docking and machine-learning prioritisation. Specifically, pkCSM analysis identified predicted hepatotoxicity and hERG-II inhibition risk, while complementary toxicity-platform predictions further supported the presence of broader toxicological concerns. These observations are consistent with the established toxicological profile of pyrrolizidine alkaloids and support the deprioritisation of Lasiocarpine for translational development despite its favourable target-binding characteristics.

A compact summary of selected translationally relevant ADMET endpoints is provided in Table 1, highlighting the major medicinal-chemistry liabilities and pharmacokinetic characteristics used during lead prioritisation. Collectively, the integrated ADMET framework reinforced Pestalamide B, Kaempferol and Apigenin as the most translationally favourable candidates emerging from the multi-layer computational workflow.

Table 1. Curated summary of selected translationally relevant ADMET endpoints for the four MD-prioritised phytochemicals.

Compound	Intestinal Absorption (%)	AMES	Hepatotoxicity	hERG-II Risk
Apigenin	90.14	Yes	No	No
Kaempferol	79.391	Yes	No	No
Lasiocarpine	64.526	No	Yes	Yes
Pestalamide B	52.241	No	No	No

3.6 MD: Structural stability and conformational dynamics

Four protein–ligand complexes were initially prioritised for molecular-dynamics evaluation, including CDK4–Lasiocarpine. However, subsequent integrated ADMET assessment identified substantial toxicity liabilities associated with Lasiocarpine, particularly hepatotoxicity-related concerns consistent with the established toxicological profile of pyrrolizidine alkaloids. Consequently, detailed downstream MD interpretation and mechanistic analysis were focused on the three translationally favourable complexes showing the strongest combined pharmacological and safety profiles (Supplementary Table S5).

All three protein–ligand systems successfully completed 200 ns production molecular dynamics simulations without trajectory interruption, maintaining stable structural behaviour throughout the simulation period. Backbone RMSD analysis demonstrated distinct dynamic characteristics among the three complexes. AKT1–Pestalamide B exhibited the highest structural stability with a mean backbone RMSD of 0.189 ± 0.022 nm, indicating highly stable maintenance of the kinase-domain architecture throughout the trajectory. Aromatase–Apigenin similarly displayed excellent conformational stability with a backbone RMSD of 0.195 ± 0.010 nm, confirming preservation of the CYP19A1 catalytic-channel structure and haem-associated active-site integrity. In contrast, CDK4–Kaempferol showed comparatively elevated backbone fluctuations (0.818 ± 0.080 nm), consistent with the documented intrinsic flexibility of cyclin-unbound CDK4 kinase domains. Despite this increased mobility, no evidence of structural unfolding or instability was observed during the simulation.

Residue-level RMSF analysis further supported these observations. AKT1 displayed the lowest mean Cα fluctuation (0.093 ± 0.045 nm), reflecting a highly rigid and stabilised protein fold upon Pestalamide B binding. Aromatase maintained similarly low residue-level flexibility (0.098 ± 0.033 nm), whereas CDK4 exhibited greater localised flexibility within functionally dynamic loop regions, including the glycine-rich P-loop and activation-loop regions characteristic of CDK-family kinases. Collectively, these analyses indicate that the selected phytocompounds maintain stable binding-associated conformations throughout the 200 ns simulation period.

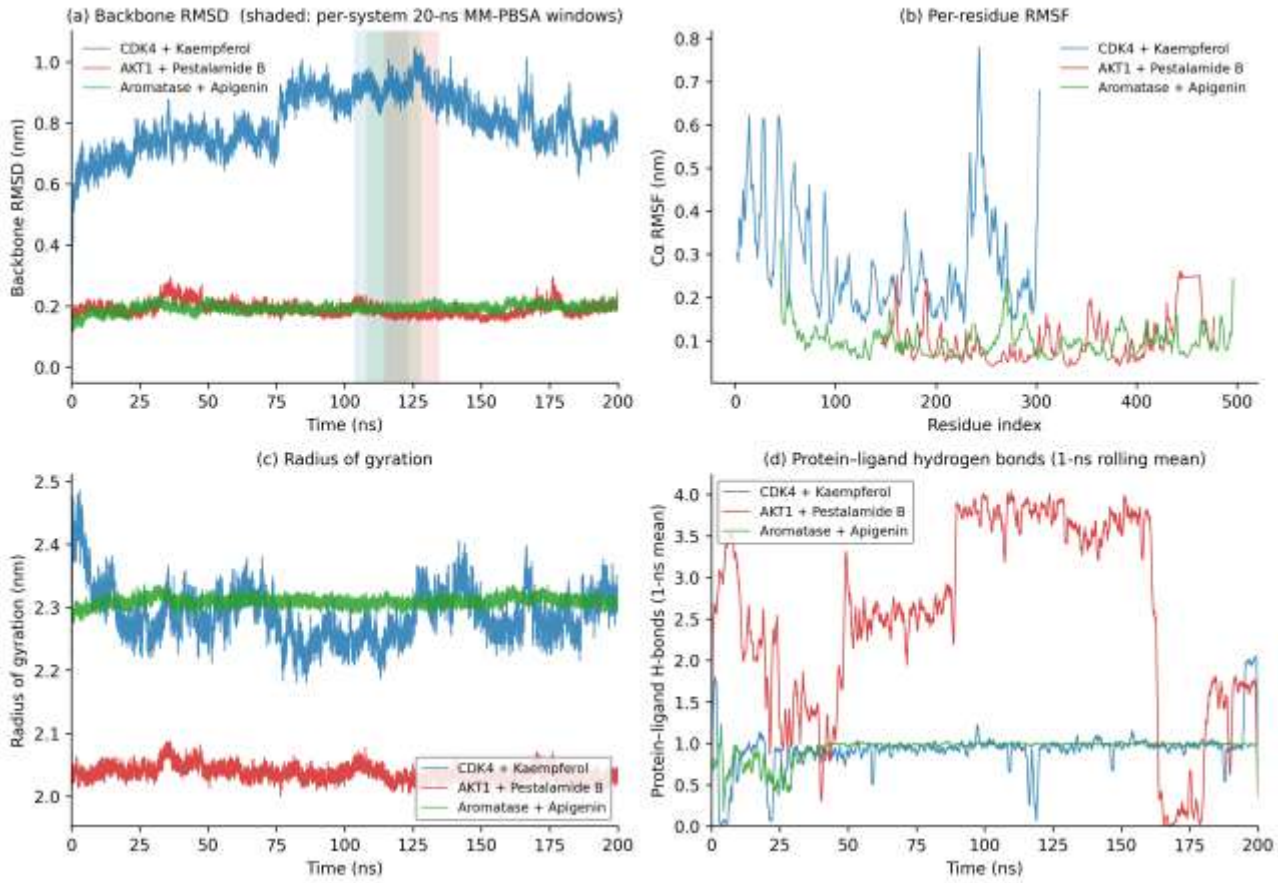

Figure 13. Molecular-dynamics stability of the three prioritised complexes over the 200 ns production trajectories. (a) Backbone RMSD over 200 ns with the per-system 20-ns MM-PBSA equilibrated windows highlighted. (b) Per-residue Cα RMSF showing the flexibility hotspots at the CDK4 P-loop / activation-loop, the AKT1 hinge region (residues 159–161, 188–191) and the Aromatase substrate-channel entry loops. (c) Radius of gyration. (d) Protein–ligand H-bonds (1-ns rolling mean).

3.7 Protein compactness and ligand interaction stability

Radius of gyration (Rg) analysis demonstrated stable protein compactness across all three simulated systems. AKT1–Pestalamide B exhibited the lowest and most stable Rg value (2.039 ± 0.012 nm), consistent with its highly rigid structural behaviour observed in RMSD and RMSF analyses. Aromatase–Apigenin maintained a stable compact fold throughout the simulation with an Rg of 2.310 ± 0.007 nm, confirming structural preservation of the large CYP19A1 catalytic framework. CDK4–Kaempferol showed moderate fluctuation in compactness (2.292 ± 0.031 nm), corresponding to the greater conformational flexibility characteristic of cyclin-free CDK4 while still maintaining an overall folded state.

Protein–ligand hydrogen-bond analysis revealed persistent intermolecular interactions throughout the simulation trajectories. Pestalamide B formed the highest average number of hydrogen bonds with AKT1 (2.49 ± 1.36), suggesting stable anchoring within the binding pocket through amide- and carbonyl-mediated interactions. Kaempferol maintained consistent hydrogen bonding within the CDK4 ATP-binding cleft despite increased conformational mobility, whereas Apigenin sustained stable polar interactions within the Aromatase catalytic cavity above the haem iron centre. Together, the compactness and hydrogen-bond analyses support stable ligand retention and persistent binding-pocket interactions across all three complexes.

To probe the essential dynamics governing each complex, principal component analysis (PCA) of the backbone trajectory was carried out using gmx covar and gmx anaeig. The first two eigenvectors captured 65–69 % of the total conformational variance across all three systems (CDK4–Kaempferol 69.2 %, AKT1–Pestalamide B 65.5 %, Aromatase–Apigenin 68.3 %), consistent with low-dimensional collective dynamics. Projection onto the PC1–PC2 subspace revealed compact and well-defined sampling for AKT1–Pestalamide B and Aromatase–Apigenin, whereas CDK4–Kaempferol exhibited a more distributed conformational envelope, reflecting the documented intrinsic flexibility of cyclin-free CDK4. The corresponding Gibbs free-energy landscape constructed on the PC1–PC2 subspace (gmx sham) displayed a single deep basin for AKT1–Pestalamide B, a clearly defined catalytic-channel basin for Aromatase–Apigenin and a broader basin for CDK4–Kaempferol - features that closely mirror the RMSD profiles. Solvent-accessible surface area (SASA) analysis yielded mean total SASA values of 161.8 ± 2.7 nm² (AKT1–Pestalamide B), 185.8 ± 5.5 nm² (CDK4–Kaempferol) and 221.6 ± 2.5 nm² (Aromatase–Apigenin, the largest protein), with AKT1 displaying the smallest fluctuation in surface exposure, consistent with its overall structural rigidity (Figure 14).

Figure 8. Phase V: Essential dynamics, free energy landscape, and solvent accessibility

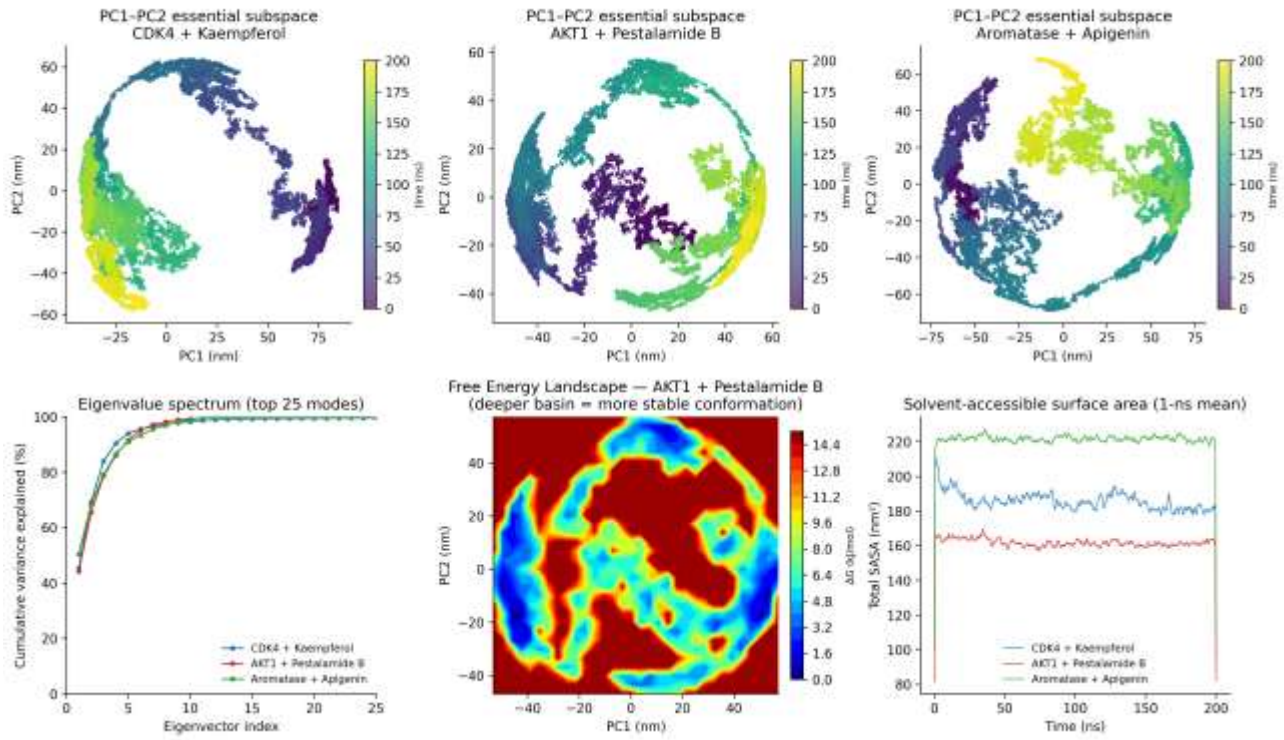

Figure 14. Essential dynamics, free-energy landscape and solvent accessibility of the three prioritised complexes. (a) PC1–PC2 projection of the backbone trajectory (PCA on Cα atoms) coloured by simulation time, showing compact sampling for AKT1–Pestalamide B and Aromatase–Apigenin and broader exploration for CDK4–Kaempferol. (b) Cumulative variance explained by the top 25 eigenvectors (top-2 modes account for 65–69 %). (c) Gibbs free-energy landscape on the PC1–PC2 subspace for AKT1–Pestalamide B, showing a single deep basin. (d) Total SASA over 200 ns (1-ns rolling mean); AKT1–Pestalamide B has the smallest SASA fluctuation (σ = 2.7 nm²).

3.8 Comparative molecular dynamics interpretation

Comparative analysis of the molecular dynamics trajectories identified AKT1–Pestalamide B as the most structurally stable complex among the three investigated systems, supported by the lowest backbone RMSD,

lowest residue-level fluctuation, smallest radius of gyration variation and highest average hydrogen-bond occupancy. Aromatase–Apigenin likewise demonstrated stable catalytic-channel binding behaviour with highly preserved structural compactness throughout the simulation. Although CDK4–Kaempferol exhibited comparatively greater conformational flexibility, the complex retained stable ligand-associated interactions and maintained overall protein fold integrity throughout the 200 ns trajectory. These findings collectively support the dynamic stability and persistent binding behaviour of the selected phytocompounds under physiologically relevant simulation conditions. The principal molecular dynamics stability metrics for all three protein–ligand systems are summarised in Table 2.

To quantify the binding energetics of the three prioritised complexes, MM-GBSA and MM-PBSA binding free energies were computed using the gmx_MMPBSA package on a 20-ns equilibrated trajectory window per system (200 frames, stride 10 ps) selected by the lowest sliding-window RMSD standard deviation in the second half of each trajectory. AKT1–Pestalamide B exhibited the strongest binding with MM-GBSA $\Delta G = -33.79 \pm 2.56$ kcal/mol and MM-PBSA $\Delta G = -3.92 \pm 2.57$ kcal/mol, driven by a balanced combination of van der Waals ($\Delta E\_VDW = -33.22$ kcal/mol) and strong electrostatic contributions ($\Delta E\_EEL = -87.63$ kcal/mol). CDK4–Kaempferol followed with MM-GBSA $\Delta G = -30.30 \pm 1.18$ kcal/mol and MM-PBSA $\Delta G = -4.44 \pm 1.17$ kcal/mol, with binding dominated by van der Waals interactions ($\Delta E\_VDW = -36.92$ kcal/mol) - consistent with the polyphenolic flavonol scaffold of Kaempferol fitting tightly into the CDK4 ATP-binding cleft. Aromatase–Apigenin yielded a moderate MM-GBSA ΔG of -9.45 ± 1.63 kcal/mol; the formally positive MM-PBSA estimate ($+19.59 \pm 1.55$ kcal/mol) reflects the known limitation of implicit-solvent treatment of haem-iron Fe(II) and porphyrin charge states (Genheden & Ryde, 2015), and the MM-GBSA value is therefore the biologically interpretable estimate for this metalloenzyme complex.

Per-residue decomposition of the MM-GBSA binding energy identified the dominant binding-pocket contributors for each complex. For CDK4–Kaempferol, the top contributors were Ile51 ($\Delta G = -1.98$ kcal/mol), Val96 (−1.78), Arg85 (−1.67), Leu49 (−1.58) and Arg38 (−1.37), mapping onto the classical CDK4 ATP-binding cleft and the catalytic hinge region. For AKT1–Pestalamide B, Val164 (−1.69), Gly157 (−1.65), Thr291 (−1.44), Lys158 (−1.29) and Met227 (−0.89) dominated, encompassing the kinase hinge and the hydrophobic ATP pocket and consistent with the previously described hinge anchoring of Pestalamide B via Ala230 and Thr291. For Aromatase–Apigenin, Thr310 (−1.47, electrostatic-dominated; $\Delta E\_EEL = -2.53$ kcal/mol), Ile133 (−1.37), Ala306 (−1.33), Val370 (−0.43) and Val369 (−0.20) were the principal contributors, mapping onto residues lining the catalytic substrate channel above the haem iron and matching the binding signature of letrozole and other clinical aromatase inhibitors (Figure 15).

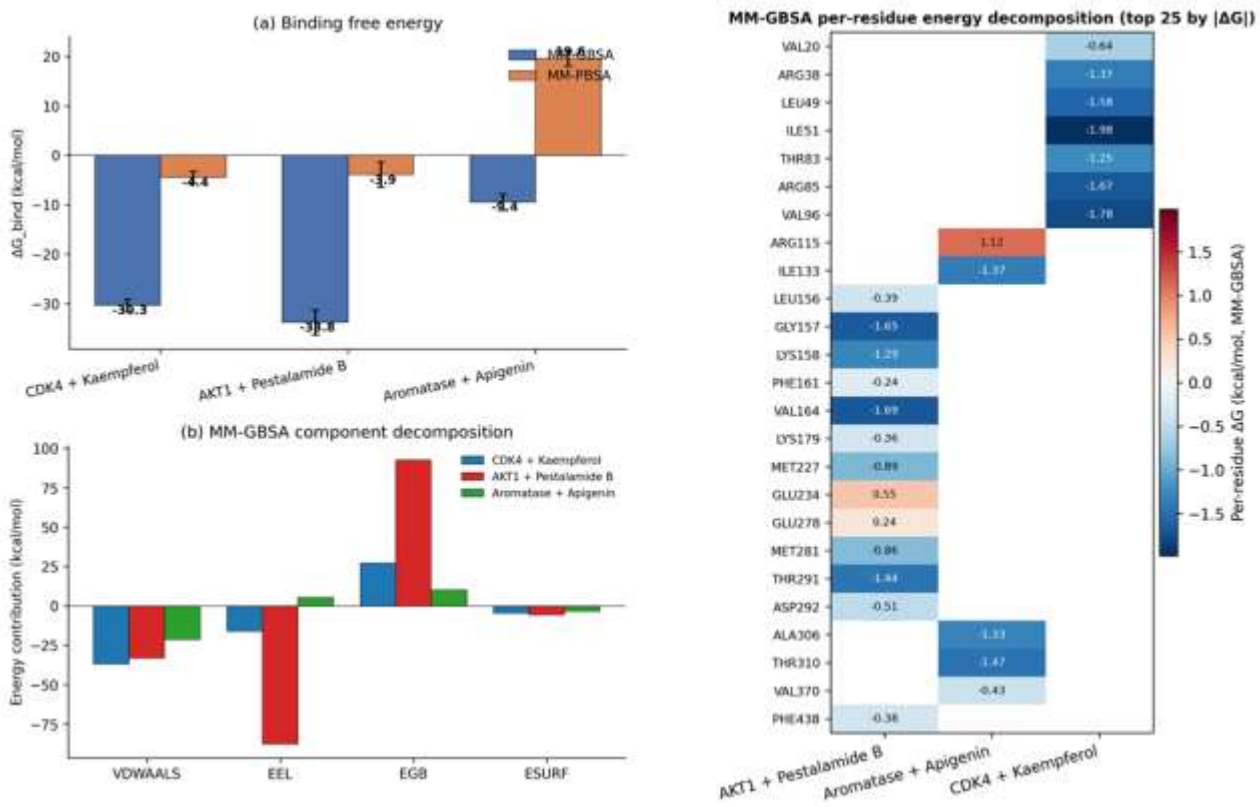

Figure 15. MM-GBSA / MM-PBSA binding free energy of the three prioritised complexes computed over a 20-ns equilibrated trajectory window (200 frames per system). (a) ΔG_bind bar chart by Generalised-Born and Poisson–Boltzmann solvation schemes (mean ± SD). (b) Energy-component decomposition (ΔE_VDW, ΔE_EEL, ΔE_GB, ΔE_SURF) for MM-GBSA showing the balanced VDW + EEL profile of AKT1–Pestalamide B and the strongly VDW-driven binding of CDK4–Kaempferol. (c) Per-residue energy decomposition heatmap (top-25 residues by absolute |ΔG|, MM-GBSA) identifying the dominant binding-pocket residues for each complex.

Table 2. Comprehensive molecular dynamics analysis summary for the three phytocompound–protein complexes over the 200 ns simulation period.

Parameter	**CDK4 + Kaempferol**	**AKT1 + Pestalamide B**	**Aromatase + Apigenin**
Backbone RMSD mean (nm)	0.818 +/- 0.080	0.189 +/- 0.022	0.195 +/- 0.010
Ligand RMSD mean (nm)	0.918 +/- 0.058	0.428 +/- 0.138	0.606 +/- 0.032
Ca RMSF mean (nm)	0.286	0.093	0.098
Ca RMSF max (nm)	0.780	0.261	0.333
Radius of gyration (nm)	2.292 +/- 0.031	2.039 +/- 0.012	2.310 +/- 0.007
H-bonds (protein-ligand)	0.94 +/- 0.40	2.49 +/- 1.36	0.96 +/- 0.22

3.9 Machine-learning bioactivity prediction and orthogonal validation

Target-specific Random Forest, XGBoost and Support Vector Machine (SVM) classifiers were trained using ChEMBL-derived bioactivity datasets for the selected therapeutic targets, generating 39 target-specific machine-learning models in total. Rather than serving as standalone predictive endpoints, the ML classifiers were implemented as an orthogonal prioritisation layer to evaluate whether consensus docking-supported H. indicum phytocompounds also resembled known target-associated bioactive chemotypes. Model-performance summaries derived from classification-report metrics demonstrated consistently strong predictive behaviour across the evaluated targets, with Random Forest classifiers showing the highest overall mean AUROC of 0.95 (range 0.87 for Topo IIα to 0.99 for PD-L1 and PI3Kα), closely followed by XGBoost (mean AUROC 0.95) and Support Vector Machines (mean AUROC 0.91); the corresponding weighted F1-scores were 0.929 (Random Forest), 0.922 (XGBoost) and 0.903 (SVM) (Supplementary Table S6).

Application of the trained classifiers to the H. indicum phytochemical library generated target-wise activity predictions that were subsequently compared with consensus docking rankings. Integrated ML–docking convergence analysis revealed selective agreement patterns across therapeutic targets rather than uniform convergence, supporting the independence of the orthogonal validation layers (Figure 16). Strong ML–docking convergence was particularly evident for PD-L1, PI3K, PR and PARP1, where multiple phytochemicals simultaneously achieved high docking support and ML-predicted activity probabilities. In contrast, kinase-focused targets including AKT1, CDK4, EGFR and HER2 demonstrated comparatively weaker ML–docking convergence despite strong docking rankings for several ligands. This behaviour likely reflects the dominance of synthetic kinase-focused chemotypes within ChEMBL training datasets relative to the structurally distinct flavonoid- and pyrrolizidine alkaloid-rich phytochemical space of H. indicum.

Importantly, the selective convergence pattern demonstrates that the workflow did not artificially force agreement between docking and machine-learning prioritisation layers. Instead, the orthogonal framework preserved biologically informative disagreement patterns while simultaneously identifying a subset of higher-confidence multi-target phytochemical candidates. Compounds including dihydroxyisoflavone, kaempferol, rinderine, lycopsamine, supinine and apigenin demonstrated broad convergence across multiple targets, supporting their prioritisation for downstream investigation. Notably, the MD-selected complexes AKT1–Pestalamide B, CDK4–Kaempferol and Aromatase–Apigenin remained strongly supported by the integrated validation framework despite varying ML convergence profiles, indicating that docking-driven molecular dynamics prioritisation and ML-supported convergence analysis represent complementary rather than redundant pharmacological evidence layers.

Quantitatively, the intersection of consensus docking top-10 hits with ML high-confidence predictions (probability ≥ 0.7) yielded 28 compound–target pairs predicted active by both signals across the 13 targets. The strongest joint convergence was observed for PI3Kα (n = 10 high-confidence pairs), PD-L1 (n = 8) and PR (n = 6), with smaller overlaps at Aromatase and PARP1 (n = 2 each). In contrast, AKT1, BCL-2, CDK4, EGFR, HER2 and Topo IIα yielded no joint overlap, because the ChEMBL-trained ML models - dominated by synthetic kinase chemotypes - conservatively rejected the H. indicum pyrrolizidine alkaloids and flavonoids that scored well by structure-based docking on these same targets. This selective rather than uniform convergence indicates that the structure-based (docking + MM-GBSA) and ligand-based (ML) layers captured complementary rather than redundant aspects of phytochemical behaviour (Figure 16).

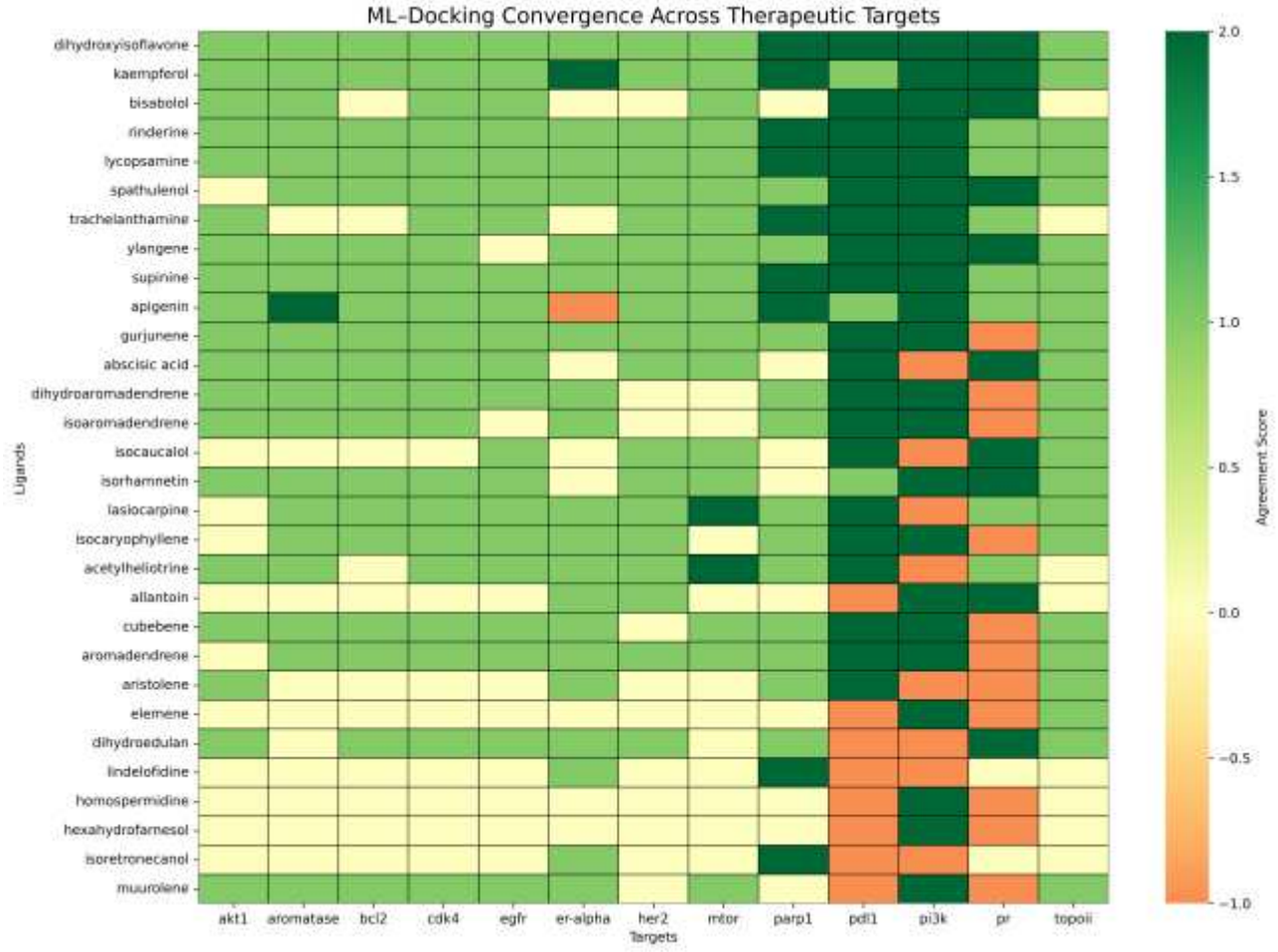

Figure 16. Machine-learning bioactivity prediction and orthogonal validation. (a) ROC curves for the 39 ML classifiers (13 targets × {Random Forest, XGBoost, SVM}) on 20 % hold-out test sets; mean Random-Forest AUROC = 0.95 with PD-L1 and PI3Kα reaching 0.99. (b) SHAP feature-importance summary for the AKT1 Random-Forest model (top 20 features) - Morgan-fingerprint substructures encoding the kinase ATP-pocket pharmacophore dominate the model's decisions. (c) Heatmap of overlap patterns between top-ranked consensus docking phytocompounds and ML activity predictions across the 13 targets, with 28 high-confidence compound–target pairs identified.

3.10 Lead-compound prioritisation

Integration of consensus docking, molecular-dynamics stability analysis, multi-tool ADMET profiling and machine-learning-assisted activity prediction identified Pestalamide B, Kaempferol and Apigenin as the most promising multi-target phytochemical candidates from *H. indicum* (Table 1). Pestalamide B demonstrated consistent prioritisation across multiple analytical layers, including strong consensus docking performance against AKT1 and EGFR, stable AKT1 binding behaviour during 200 ns molecular dynamics simulation, favourable ADMET characteristics and predicted activity across multiple kinase-associated targets. Kaempferol showed strong mechanistic relevance toward CDK4, supported by stable kinase-domain interactions and conserved hinge-region binding features, although mild AMES-related toxicity alerts were observed in pkCSM and ADMETlab predictions. Apigenin displayed favourable Aromatase binding geometry with stable active-site occupancy and acceptable pharmacokinetic properties. Dihydroxyisoflavone additionally demonstrated broad multi-target docking behaviour and ML-predicted activity across several

breast-cancer-associated targets, supporting its relevance as a secondary multi-target scaffold. Although Lasiocarpine exhibited acceptable docking performance and appeared among the top-ranked multi-target candidates, its predicted hepatotoxicity and carcinogenicity liabilities resulted in de-prioritisation during downstream candidate refinement.

4. Discussion

The present study implemented an integrated multi-layer computational framework to systematically prioritise anti-breast-cancer phytochemicals from *Heliotropium indicum*. Unlike conventional natural-product screening workflows that rely primarily on docking-score ranking, the current pipeline incorporated transcriptomic target validation, network-pharmacology analysis, multi-engine consensus docking, pharmacokinetic filtering, molecular-dynamics simulation, pharmacophore modelling and machine-learning-assisted orthogonal validation to establish a biologically and translationally informed prioritisation strategy. Importantly, the workflow was designed sequentially such that consensus docking first identified structurally promising compound–target pairs, subsequent multi-tool ADMET assessment filtered translationally unsuitable candidates prior to detailed MD interpretation, and machine-learning analysis was employed independently as a final orthogonal validation layer rather than as a primary screening filter. This layered architecture reduced dependence on any single computational methodology and enabled cross-validation of compound behaviour through mechanistically distinct analytical approaches.

Transcriptomic and network-level analyses demonstrated that the selected therapeutic panel represented biologically relevant breast-cancer signalling architecture rather than an arbitrarily assembled docking set. Differential-expression profiling across TCGA-BRCA datasets identified heterogeneous dysregulation patterns among the prioritised targets, with prominent alterations observed in ESR1, TOP2A, PARP1 and ERBB2, while subtype-associated expression patterns further highlighted relevance across luminal, HER2-positive and basal-like/TNBC-associated contexts. Protein–protein interaction and CytoHubba analyses consistently identified AKT1, PIK3CA, CDK4, ERBB2 and MTOR as dominant network hubs with high topological centrality, whereas KEGG enrichment strongly implicated PI3K–AKT, ErbB, oestrogen-signalling and apoptosis-associated pathways. Collectively, these findings support the biological validity of the selected target space and reinforce the relevance of multi-target therapeutic strategies for breast-cancer systems characterised by pathway redundancy and adaptive resistance.

Consensus docking analysis integrating AutoDock Vina, AutoDock4, GNINA and PLANTS identified recurrent multi-target enrichment of Pestalamide B, Kaempferol, Apigenin, Dihydroxyisoflavone and Lasiocarpine across multiple oncogenic proteins. Importantly, compound prioritisation was not based solely on docking rank, but additionally considered chemical diversity, multi-target recurrence, pharmacophore consistency and downstream translational suitability. Three-dimensional pharmacophore extraction revealed conserved donor–acceptor–hydrophobic feature arrangements shared across structurally diverse phytochemicals, supporting the structural consistency of the identified binding poses and reinforcing the mechanistic plausibility of the consensus docking framework. GNINA CNN rescoring further contributed to prioritisation robustness by penalising sterically unrealistic poses that would otherwise receive artificially favourable empirical docking scores, thereby improving confidence in the final consensus-ranking landscape.

Pestalamide B emerged as the most consistently prioritised phytochemical across the integrated analytical layers. Although *H. indicum* is traditionally associated with pyrrolizidine alkaloids including lasiocarpine, lycopsamine and indicine, Pestalamide B belongs to a chemically distinct scaffold class lacking the classical

unsaturated necine-associated structural features commonly implicated in pyrrolizidine alkaloid toxicity. Structurally, Pestalamide B demonstrated stable hinge-associated interactions within the AKT1 binding cavity involving Ala230 and Thr291, accompanied by strong multi-target consensus docking behaviour across kinase-associated targets. Its favourable ADMET profile relative to other prioritised phytochemicals further strengthened its translational suitability. Dihydroxyisoflavone additionally demonstrated recurrent multi-target enrichment and broad ML-supported activity patterns, supporting its relevance as a secondary multi-target scaffold candidate.

Kaempferol and Apigenin represented mechanistically distinct but biologically complementary phytochemical leads. Kaempferol demonstrated stable occupancy within the CDK4 ATP-binding cleft with conserved kinase hinge-region interactions involving Val96 and Lys35, consistent with established CDK4/6 inhibitor binding geometry. Apigenin occupied the Aromatase catalytic channel with stable interactions involving Arg115, Trp224 and Met374, supporting its relevance toward endocrine-associated breast-cancer pathways. The pharmacophore architectures of both flavonoids displayed conserved planar aromatic and hydrogen-bond acceptor arrangements characteristic of kinase- and catalytic-pocket-associated phytochemical binding motifs.

Multi-tool ADMET analysis proved particularly important for translational refinement of the docking-prioritised compounds. While Lasiocarpine repeatedly appeared among the strongest multi-target consensus binders and was initially retained because of its chemically distinct pyrrolizidine alkaloid scaffold, subsequent ADMET assessment identified predicted hepatotoxicity and hERG-associated liabilities consistent with the established toxicological concerns associated with pyrrolizidine alkaloids. The sequential placement of ADMET filtering prior to detailed MD interpretation therefore enabled rational deprioritisation of translationally unsuitable candidates despite favourable docking behaviour. This represents an important strength of the workflow because it prevented overinterpretation of structurally favourable but pharmacologically problematic compounds.

Molecular-dynamics simulations further supported the structural stability of the prioritised complexes. AKT1–Pestalamide B and Aromatase–Apigenin exhibited highly stable backbone behaviour throughout the 200 ns simulation period, whereas CDK4–Kaempferol displayed comparatively elevated fluctuations consistent with the known intrinsic flexibility of cyclin-unbound CDK4 kinase domains. PCA and free-energy landscape analyses indicated relatively confined conformational sampling for AKT1–Pestalamide B and Aromatase–Apigenin, supporting stable ligand occupancy within their respective binding environments. Importantly, the explicit treatment of the Aromatase haem cofactor using validated CHARMM-compatible topology parameters was critical for maintaining catalytically relevant cavity geometry during simulation, highlighting the importance of accurate cofactor parameterisation when studying haem-associated targets.

The machine-learning layer provided an independent orthogonal validation framework rather than serving as a direct prioritisation filter. Selective convergence between ML predictions and docking-derived prioritisation was observed for PI3K, PD-L1, PR and PARP1, whereas kinase-associated targets including AKT1, CDK4 and EGFR demonstrated comparatively limited ML–docking overlap. Rather than weakening the workflow, this divergence highlights the distinct information captured by ligand-based ML models trained predominantly on synthetic ChEMBL chemotypes relative to structure-based docking approaches applied to chemically diverse phytochemicals. The observation that orthogonal evidence layers converged selectively rather than uniformly strengthens confidence in the overall prioritisation strategy by demonstrating that agreement was not artificially forced across computational methods.

Collectively, the present study demonstrates the value of integrating transcriptomics, network pharmacology, consensus docking, translational ADMET filtering, molecular dynamics and orthogonal

machine-learning validation into a unified phytochemical prioritisation framework. The combination of multi-target systems-level analysis with sequential translational refinement enabled identification of structurally diverse phytochemicals exhibiting biologically plausible target engagement, favourable pharmacokinetic characteristics and stable dynamic behaviour across clinically relevant breast-cancer-associated pathways. Although experimental validation remains necessary, the computational evidence generated here provides a robust foundation for future biochemical, cellular and preclinical investigation of *H. indicum*-derived anti-breast-cancer leads.

5. Limitations and future perspectives

Several limitations of the present study should be acknowledged. First, the entire framework is computational and therefore requires experimental validation to confirm biological activity, target engagement and translational relevance in breast-cancer systems. Validation in representative subtype-specific cell models including MCF-7 (luminal), MDA-MB-231 (TNBC) and BT-474 (HER2-positive), together with orthogonal binding assays such as surface plasmon resonance (SPR) or isothermal titration calorimetry (ITC), represents an essential next step. Second, molecular-dynamics simulations were performed as single 200 ns production trajectories for each prioritised complex. Although stable behaviour was consistently observed, additional independent replicate simulations and apo-protein control trajectories would further strengthen confidence in the observed conformational trends and ligand-induced stabilisation patterns. Third, the ADMET layer was entirely prediction-based and should therefore be interpreted as translational prioritisation rather than experimentally validated pharmacokinetic evidence. Parameters including bioavailability, plasma-protein binding, tissue distribution, metabolic stability and long-term toxicity remain to be experimentally characterised.

Additional methodological limitations should also be considered. The machine-learning classifiers were trained primarily on ChEMBL-derived datasets enriched for synthetic kinase-oriented chemotypes, potentially limiting generalisation toward structurally distinct phytochemical scaffolds. This likely contributed to the selective divergence observed between docking-derived and ML-derived prioritisation for several kinase-associated targets. Importantly, this divergence does not invalidate the orthogonal framework but instead reflects differences between ligand-based chemotype recognition and structure-based interaction modelling. Furthermore, explicit quantum-mechanical characterisation of electronic properties, including frontier molecular orbitals and electrostatic-potential surfaces, was beyond the scope of the present work and may provide additional structure–activity insight in future studies.

Despite these limitations, the study possesses several notable strengths. The workflow integrated transcriptomic validation, network pharmacology, multi-engine consensus docking, pharmacophore modelling, translational ADMET filtering, extended molecular-dynamics simulation and orthogonal machine-learning validation into a unified prioritisation framework. Importantly, ADMET filtering was strategically positioned after consensus docking but prior to detailed MD interpretation, enabling early elimination of translationally unsuitable candidates such as Lasiocarpine despite favourable docking behaviour. The selective rather than forced convergence observed between docking and ML layers further strengthens confidence in the robustness of the prioritisation strategy by demonstrating that independent computational approaches captured complementary aspects of phytochemical behaviour. The incorporation of multiple orthogonal

validation layers therefore reduced dependence on any single scoring methodology and improved overall interpretability of the final prioritised leads.

Future work should focus on experimental validation and methodological expansion of the current framework. Immediate priorities include replicate MD simulations with apo-protein controls, biochemical target-engagement assays for AKT1, CDK4 and Aromatase, and subtype-specific cellular validation of Pestalamide B, Kaempferol and Apigenin using viability, apoptosis and kinase-inhibition assays. Extension of the machine-learning framework toward graph-neural-network architectures and phytochemical-enriched training datasets may additionally improve predictive generalisation for structurally diverse natural-product scaffolds. Collectively, the present computational framework provides a robust systems-level foundation for future translational investigation of *H. indicum*-derived anti-breast-cancer phytochemicals.

6. Conclusion

This study presents the first comprehensive multi-layer computational evaluation of *Heliotropium indicum* phytochemicals against a subtype-relevant breast-cancer therapeutic landscape integrating transcriptomic target validation, network pharmacology, multi-engine consensus docking, pharmacophore modelling, translational ADMET filtering, molecular-dynamics simulation and orthogonal machine-learning-assisted bioactivity prediction. Systematic phytochemical mining and prioritisation identified structurally diverse compounds capable of engaging multiple clinically relevant breast-cancer-associated targets involved in PI3K–AKT, ErbB, endocrine and apoptosis-associated signalling pathways.

Integrated evidence across the orthogonal computational layers consistently prioritised Pestalamide B, Kaempferol, Apigenin and Dihydroxyisoflavone as the most promising phytochemical leads. Pestalamide B emerged as the strongest overall multi-target candidate because of its recurrent consensus docking enrichment, stable AKT1 interaction profile and comparatively favourable ADMET characteristics. Kaempferol demonstrated strong mechanistic relevance toward CDK4-associated cell-cycle regulation, whereas Apigenin displayed stable Aromatase-associated binding geometry consistent with endocrine-pathway targeting. Dihydroxyisoflavone additionally exhibited broad multi-target enrichment and ML-supported activity patterns across several breast-cancer-associated targets. In contrast, compounds such as Lasiocarpine, despite favourable docking behaviour, were translationally deprioritised following integrated ADMET assessment, highlighting the importance of sequential pharmacokinetic filtering within multi-stage phytochemical prioritisation workflows.

Importantly, selective rather than universal convergence between docking-derived and machine-learning-derived prioritisation strengthened confidence in the orthogonal validation strategy by demonstrating that independent computational methodologies captured complementary aspects of phytochemical behaviour rather than artificially forcing agreement. Collectively, the study establishes a robust systems-level computational framework for identifying biologically relevant and translationally prioritised natural-product leads against heterogeneous breast-cancer signalling networks. The prioritised *H. indicum* phytochemicals therefore warrant further experimental validation through biochemical, cellular and preclinical investigation.

References

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020. CA: A Cancer Journal for Clinicians, 71(3), 209–249.

Perou, C. M., Sørlie, T., Eisen, M. B., et al. (2000). Molecular portraits of human breast tumours. Nature, 406(6797), 747–752.

Sørlie, T., Perou, C. M., Tibshirani, R., et al. (2001). Gene expression patterns of breast carcinomas distinguish tumour subclasses with clinical implications. PNAS, 98(19), 10869–10874.

Hu, Y., Gupta-Ostermann, D., & Bajorath, J. (2017). Exploring compound promiscuity patterns and multi-target activity spaces. Computational and Structural Biotechnology Journal, 15, 33–45.

Bolognesi, M. L., & Cavalli, A. (2016). Multitarget Drug Discovery and Polypharmacology. ChemMedChem, 11(12), 1190–1192.

Newman, D. J., & Cragg, G. M. (2020). Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. Journal of Natural Products, 83(3), 770–803.

Roeder, E. (2000). Medicinal plants in China containing pyrrolizidine alkaloids. Pharmazie, 55(10), 711–726.

Mandal, S. C., Mandal, V., & Das, A. K. (2014). Phytochemical and pharmacological review of Heliotropium indicum. Pharmacognosy Reviews, 8(15), 24–32.

Bhattacharyya, D., & Saha, A. (2013). Anti-tumour activity of Heliotropium indicum on Ehrlich ascites carcinoma. Indian Journal of Pharmacology, 45(1), 50–53.

Reddy, J. S., Rao, P. R., & Reddy, M. S. (2008). Cytotoxic alkaloids from Heliotropium indicum. Indian Journal of Chemistry B, 47(10), 1572–1577.

Pinzi, L., & Rastelli, G. (2019). Molecular docking: shifting paradigms in drug discovery. International Journal of Molecular Sciences, 20(18), 4331.

Vamathevan, J., Clark, D., Czodrowski, P., et al. (2019). Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery*, 18(6), 463–477.

Riniker, S., & Landrum, G. A. (2015). Better informed distance geometry: using what we know to improve conformation generation. Journal of Chemical Information and Modeling, 55(12), 2562–2574

Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S., & Olson, A. J. (2009). AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. Journal of Computational Chemistry, 30(16), 2785–2791.

Stamford, A., & Banerji, U. (2023). Capivasertib in breast cancer: pharmacology and clinical translation. Nature Reviews Clinical Oncology, 20(10), 707–722.

Daina, A., Michielin, O., & Zoete, V. (2017). SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Scientific Reports*, 7, 42717.

Pires, D. E. V., Blundell, T. L., & Ascher, D. B. (2015). pkCSM: predicting small-molecule pharmacokinetic and toxicity properties using graph-based signatures. *Journal of Medicinal Chemistry*, 58(9), 4066–4072.

Fu, L., Shi, S., Yi, J., Wang, N., He, Y., Wu, Z., Peng, J., Deng, Y., Wang, W., Wu, C., Lyu, A., Zeng, X., Zhao, W., Hou, T., & Cao, D. (2024). ADMETlab 3.0: an updated comprehensive online ADMET prediction platform. Nucleic Acids Research, 52(W1), W422–W431.

Banerjee, P., Eckert, A. O., Schrey, A. K., & Preissner, R. (2018). ProTox-II: a webserver for the prediction of toxicity of chemicals. *Nucleic Acids Research*, 46(W1), W257–W263.

Lindorff-Larsen, K., Piana, S., Palmo, K., et al. (2010). Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Structure, Function, and Bioinformatics*, 78(8), 1950–1958.

Bussi, G., Donadio, D., & Parrinello, M. (2007). Canonical sampling through velocity rescaling. *The Journal of Chemical Physics*, 126(1), 014101.

Parrinello, M., & Rahman, A. (1981). Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics*, 52(12), 7182–7190.

Abraham, M. J., Murtola, T., Schulz, R., et al. (2015). GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1–2, 19–25.

MacKerell, A. D., Bashford, D., Bellott, M., et al. (1998). All-atom empirical potential for molecular modelling and dynamics studies of proteins. Journal of Physical Chemistry B, 102(18), 3586–3616.

Sousa da Silva, A. W., & Vranken, W. F. (2012). ACPYPE-AnteChamber PYthon Parser interfacE. *BMC Research Notes*, 5, 367.

Autenrieth, F., Tajkhorshid, E., Baudry, J., & Luthey-Schulten, Z. (2004). Classical force field parameters for the haem prosthetic group of cytochrome c. Journal of Computational Chemistry, 25(13), 1613–1622.

Genheden, S., & Ryde, U. (2015). The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. Expert Opinion on Drug Discovery, 10(5), 449–461.

Funding

This work was supported by the Vels University Seed Grant Project. The funders had no role in study design, data collection, or analysis.

Author contributions

S S – Overall project work, Initial writing and correction , MN – Docking and Dyanmics support , AL – Figures, ML works , AP – Manuscript structure design and wring, Project design , DS – Supervision, Guidance , RM – Project design, Funding support

Conflict of interest

The authors declare no competing interests.

Data availability

All raw data (compound SMILES, prepared protein PDBs, docking poses, MD trajectories, MM-PBSA per-frame energies, ADMET tool outputs, ML pipelines, ROC data, SHAP values), analysis scripts (Python, GROMACS commands) and generated tables are available in the supplementary repository accompanying this manuscript.