



Integrating machine learning with in silico studies and Quantum Chemistry: Exploring novel compounds through multiscale screening targeting the CDK2 enzyme



Priyanka Solanki ^a , Sk. Abdul Amin ^{b,c} , Anu Manhas ^{a,*}

^a Department of Chemistry, Pandit Deendayal Energy University, Gandhinagar, 382426, India

^b Department of Pharmacy, University of Salerno, Via Giovanni Paolo II 132, Fisciano, 84084, SA, Italy

^c Department of Pharmaceutical Technology, JIS University, 81, Nilgunj Road, Agarpara, Kolkata, 700109, West Bengal, India

ARTICLE INFO

Keywords:

Machine learning
CDK2
Anticancer
Molecular docking
Molecular dynamics simulations
Free energy calculations

ABSTRACT

Cyclin-dependent kinase 2 (CDK2) modulates the progression of the cell cycle, and its dysregulation results in unchecked cellular proliferation, establishing it as a pivotal target in oncological therapies. We implemented a comprehensive screening pipeline to identify potential novel inhibitors for the CDK2 enzyme by integrating advanced machine learning classification methods. The random forest (RF) method shows better performance based on the statistical metrics assessment. This RF model was used to filter a large coconut dataset comprising 477,975 molecules to identify potential candidates. This initial screening process identified 327 candidate molecules. The subsequent application of PAINS (Pan-Assay Interference Structures) filtration refined this list to 309 molecules, which were then selected for molecular docking analysis. Based on the docking score, the top 40 potential candidates from molecular docking analysis were chosen for pharmacokinetics (PK) and pharmacodynamics (PD) studies (ADMET). Three molecules that satisfy the PK/PD criteria were selected for DFT and molecular dynamics simulation studies. The finalized three molecules displayed conserved interactions with the residues Lys33 and Asp145, crucial for enzyme inhibition. Moreover, Molecule 2 possessed an extended fused heterocyclic system, which may enhance its inhibitory potential. The simulation studies indicate that these compounds showed stable behavior within the binding pocket of the CDK2 enzyme. Also, we have developed an open-access online tool named "pCDK2i_v1.0" to help the scientific community efficiently screen the potential CDK2 inhibitors. This work demonstrates the importance of integrating machine learning in drug design to discover novel anti-cancer inhibitors of the CDK2 enzyme. The pCDK2i_v1.0 tool for screening and predicting the CDK2 activity as active (1), and inactive (0) is available at https://github.com/Amincheminfom/pCDK2i_v1.

1. Introduction

Cancer has become one of the most widespread and deadly diseases around the world. The rate of cancer cases is increasing quickly and becoming a challenge for healthcare systems worldwide. The World Health Organisation (WHO) stated that there were 19.3 million new cases and 9.96 million cancer-related deaths reported in 2020. Cancer incidence is expected to rise significantly, with a projection of 28.4 million new cases by 2040. In 2020, breast, lung, and colorectal cancers emerged as the most prevalent diagnoses worldwide. Breast cancer led globally with over 2.3 million new cases in 2020, making it the most commonly diagnosed cancer. It is also the leading cause of cancer-related deaths among women, with approximately 685,000 deaths

globally in 2020 [1]. Major risk factors for breast cancer include age, obesity, family history, alcohol consumption, genetic mutations, hormonal factors, and lack of physical activity. Despite extensive research, an effective and universally accessible cure for breast cancer remains elusive. Moreover, the drug discovery process is long, costly, complex, and often takes decades to develop. In recent years, targeted therapies have shown promising results by targeting specific cancer cells without damaging healthy tissues [2]. One promising target is Cyclin-Dependent Kinase 2 (CDK2), a key cell cycle regulator. It regulates the G1 to S phase of the cell cycle and ensures DNA replication in the S phase [3]. In the late G1 phase, Cyclin E binds to CDK2 and phosphorylates the Rb protein, which then activates E2 transcription factor (E2F). This activated E2F drives the transcription of S-phase proteins, which facilitates the

* Corresponding author.

E-mail addresses: Anu.Manhas@sot.pdpu.ac.in, anu.manhas15@gmail.com (A. Manhas).

cell's entry into the S phase. Besides its crucial role in cell cycle progression, CDK2 also participates in the DNA damage response (DDR) in cells. It plays a role in the homologous recombination (HR) repair pathway by activating the MRN-CtlP-BRCA1 complex, which is essential for repairing DNA double-strand breaks. CDK2 regulates apoptosis by phosphorylating FOCO1 or MCL-1. Additionally, CDK2 is involved in signal transduction, DNA and RNA metabolism, and translation. It can interact with p53, a key tumor-suppressing transcription factor. Activation of p53 represses CDK2 expression, while ectopic expression of CDK2 can bypass p53-dependent senescence [4]. Overexpression or dysregulation of CDK2 is implicated in various cancers, including breast cancer and ovarian cancer. High nuclear CDK2 protein expression is significantly associated with aggressive breast cancer phenotypes, including high tumor grade, lymphovascular invasion, and poor prognosis [4]. In ovarian cancer, exceptionally high-grade serous ovarian cancer (HGSOC), cyclin E (CCNE1) is frequently amplified. This amplification leads to overexpression of cyclin E, resulting in hyperactivation of CDK2. The increased CDK2 activity drives uncontrolled cell proliferation by bypassing the regular cell cycle checkpoints, contributing to tumor growth and progression [5]. Thus, targeting CDK2 offers a potential therapeutic strategy to inhibit cancer and improve treatment outcomes with fewer side effects than traditional therapies. Recent advancements in CDK2 inhibitors have shown promising potential for cancer therapy; however, to date, there are no FDA-approved drugs available. However, there are many reported crystallized inhibitors of the CDK2 protein. This gives us an opportunity to conduct AI-based calculations along with the computational techniques to explore the large chemical space of CDK2 inhibitors to predict and optimize the novel inhibitors of the CDK2 enzyme.

Over the past decade, Artificial Intelligence (AI) has revolutionized drug discovery by rapidly analyzing large datasets, predicting drug efficacy, and enhancing the drug development process. By reducing the time and resources required for clinical trials, AI significantly improves the efficiency of identifying successful therapeutic candidates [6,7]. Various models, including artificial neural networks (ANN), random forests (RF), and support vector machines (SVM), have been utilized to classify compounds according to their activity. A prevalent method for assessing the similarity of molecules involves transforming their structures into bit vectors. We used this method to conduct virtual screening based on the highest-performing machine learning model. The screened molecules were subjected to molecular docking to explore the binding pattern of the candidates in the active site of the protein [8]. The drug-likeness and toxicity of the selected compounds were assessed. The shortlisted molecules were subjected to theoretical studies by applying density functional theory (DFT) calculations to evaluate the electronic properties of the compounds. Chemical properties like energies of highest occupied molecular orbital (HOMO), lowest unoccupied molecular orbital (LUMO), energy gap within HOMO-LUMO, chemical potential (μ), electronegativity (χ), hardness (η), softness (S), and electrophilicity index (ω) related to the electronic structure were examined. These properties can help to shortlist the compound that has the capacity to be optimized further to enhance the binding affinity, improve the bioavailability, and reduce the toxicity by adjusting the reactivity of the molecules. Also, molecular dynamic simulations and free energy calculations were employed to assess various parameters like root mean square deviation (RMSD), root mean square fluctuation (RMSF), radius of gyration (RoG), hydrogen bond (H-bond) formation, solvent-accessible surface area (SASA), principal component analysis (PCA), free energy landscape (FEL), dynamic correlation matrix (DCCM), and interaction energy, to predict the stability and conformational flexibility of the docked complexes.

2. Materials and methods

2.1. Data collection and preparation

The online open-source tool, the OpenCADD platform for cheminformatics, was used in the current study to generate machine learning models [9]. The reported CDK2 inhibitors were retrieved in SMILES (Simplified Molecular Input Line Entry System) format from the ChEMBL (Chemical European Molecular Biology Laboratory) database [10]. The information on inhibitory activity in terms of IC₅₀ (Inhibitory Concentration-50) values of the respective CDK2 inhibitors was also collected. After collecting the data, only those molecules with unique inhibitory activity were selected (structural duplicates were removed). To maintain the scale of the inhibitory activity data, the IC₅₀ values were converted into pIC₅₀ values (negative logarithms). Thereafter, the prepared data with pIC₅₀ values was subjected to Lipinski's Rule of Five (RO5) to filter out the molecules with acceptable drug-like properties [11]. Assessing the lead-like properties of the compounds is essential to predict their oral bioavailability and pharmacokinetic properties. As per RO5, compounds with molecular weight (MW) \leq 500 Da, hydrogen bond acceptors \leq ten, hydrogen bond donors \leq five, and octanol-water partition coefficient (LogP) value \leq 5 exhibit favorable absorption and permeation properties. This rule provides an initial filter to evaluate the drug-likeness of compounds.

2.2. Model development and assessment

Molecular fingerprints are quantitative representations of molecular structures by transforming structural details into numerical vectors or binary strings. This approach allows quick similarity evaluations crucial in virtual screening and structure–activity relationship (SAR) studies. In this work, we have collected SMILES data and converted it into molecular fingerprints, which were calculated using RDKit [12]. These fingerprints are used as inputs to develop machine learning models. The data was divided into 884 active compounds and 705 inactive ones, using a pIC₅₀ cut-off value of 6.3 nM, which is the standard value we will apply in our study [13,14]. We constructed three RF, SVM, and ANN models using MACCS fingerprints for QSAR classification. In this study, we have used Scikit-learn to develop our classifier models.

To evaluate the classifier models' performance, several metrics were calculated, including sensitivity, specificity, accuracy, and AUC (Area Under the Curve). In this context, the true-positive (TP) value represents the number of inhibitors correctly identified as active, while the true-negative (TN) corresponds to decoys correctly identified as inactive. Likewise, a False-positive (FP) refers to the decoys misclassified as active compounds, and a false-negative (FN) indicates the inhibitors incorrectly classified as inactive. Accuracy reflects the overall correctness of the model, while sensitivity measures its ability to identify true positives correctly. On the other hand, specificity quantifies the model's capability to correctly identify negative cases, ensuring a balanced assessment of its predictive performance. The AUC evaluates the classification models by comparing the true positive rate (TPR) with the false positive rate (FPR).

$$\text{Sensitivity} = \text{TP} / (\text{TP} + \text{FN})$$

$$\text{Specificity} = \text{TN} / (\text{TN} + \text{FP})$$

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{FP} + \text{FN} + \text{TN}). \text{Where TP} = \text{True Positive}, \text{TN} = \text{True Negative}, \text{FP} = \text{False Positive}, \text{FN} = \text{False Negative}.$$

2.3. Virtual screening

We employed the most effective machine learning model for ligand-based virtual screening (VS). We utilized the ensemble-based random forest model to conduct the ligand-based virtual screening (VS). To perform the ligand-based VS, we selected around 4 lakh molecules from

the COCONUT (Collection of Open Natural Products database) database were used as an external library for VS [15]. Before VS, using RDKit [16], the selected database was filtered to remove the compounds with invalid SMILES and Pan Assay Interface Compounds (PAINS), not following Lipinski's RO5 criteria. PAINS can mislead the results by yielding false-positive biological screening assays [17]. Some of the chemical features, such as nitro groups (mutagenic groups), phosphates, and sulfates (which may result in poor pharmacokinetic properties), and reactive groups like 2-halopyridines, were removed, which can negatively impact drug development [18]. Moreover, compounds with substructures like catechols and rhodanines are also known to cause false-positive outcomes [19]. Therefore, filtering for these undesirable features, including PAINS, is essential to assembling a more efficient and reliable screening library. Applying Lipinski's guidelines and eliminating unwanted substances increases the probability of identifying compounds with favorable properties and lower risk profiles for clinical development.

2.4. Molecular docking

Molecular docking studies were conducted on the screened candidates to study the interaction of the molecules with the target protein. As per the literature, CDK enzymes have ATP binding sites, non-competitive binding sites, and allosteric binding sites. Designing the inhibitors that bind to the ATP-binding site is considered the primary strategy in drug development against the CDK2 enzyme [20–22]. This strategy prevents the ATP from binding, thus blocking the functioning of the kinase enzyme [20–22]. Therefore, considering the importance of inhibition of the ATP-binding site, the protein with PDB ID 6Q4G and inhibitor HJK co-crystallized in the ATP-binding site was selected. Molecular docking was performed using the FlexX module [23] of the SeeSAR v13.1.1 software [24] to determine their binding affinity. This module utilizes a modified Böhm scoring function to estimate the binding affinity of the candidates based on their interactions with the protein [25]. The selected molecules were initially in SMILES format, which we converted into SDF (Structure Data File) format using an open-source software, Open Babel [26]. Before docking with the screened molecules, we performed redocking on the reference PDB (6Q4G-HJK). Redocking was performed to validate the software. During the docking calculations, the protein chain was prepared in the Protein Editor Mode of the SeeSAR software [24]. This mode executes the preparation of the protein chain by automatically adding the missing side-chains and hydrogen atoms, assigning bond orders, removing water, and alternate residues, generating hetero protonation states at pH 7.4, and optimizing the hydrogen bonding networks. The prepared protein and reference ligand (HJK) were exported to the Binding Site Mode of SeeSAR software [24] to generate the binding pocket. The binding site was created based on the experimentally reported inhibitor's co-crystallized position and the essential amino acid residues. The top 10 poses were generated for each molecule during the calculations, and the most favorable conformations were selected. The 2D interaction plots of the docked complexes were generated using the Discovery Studio Visualizer. Based on the docking scores, the selected molecules were subsequently subjected to ADME/T studies.

2.5. ADME/T studies

The web-based SwissADME and pkCSM tool was used to evaluate the physicochemical, pharmacokinetics, and Toxicity of the selected molecules [27,28]. The 2D coordinates of the docked compounds were converted into SMILES format to study the ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicity) properties. The SwissADME tool predicts physicochemical properties, including molecular weight, rotatable bonds, hydrogen bond acceptors and donors, total polar surface area (TPSA), lipophilicity (LogP), and water solubility. It also assessed a molecule based on the drug's likeness properties

according to Lipinski's RO5, Ghose, Veber, Egan, and Muegge rules to predict the oral bioavailability of the compounds. The tool also predicts the pharmacokinetic properties, such as gastrointestinal (GI) absorption, P-glycoprotein (P-gp) substrate prediction, partition coefficient (Log K_p), blood-brain barrier (BBB) permeability, and cytochrome P450 interactions. The ADME filtered molecules were further subjected to study the toxicity properties using the pkCSM tool. Finally, the ADME/T filtered candidates were subjected to the DFT studies.

2.6. Quantum mechanical electrostatic potential calculations

The DFT studies over the shortlisted compounds, along with the reference, were performed using the Gaussian 09 program [29]. To conduct the studies, the initial geometries of the selected compounds were subjected to an optimization process by using a hybrid functional B3LYP (Becke three-parameter Lee-Yang-Parr) [30] along with a split-valence atomic basis set 6-31+g(d) [31]. All the calculations were performed using the B3LYP/6-31+g(d) level of theory. The minima of the optimized candidates were assessed via frequency calculations, and the absence of the imaginary frequency indicates the generation of lower energy coordinates at the minima. The frontier molecular orbitals were constructed to analyze the HOMO and LUMO distributions. Through the HOMO-LUMO and energy gap analysis, we can determine the reactivity and charge transfer potential of the molecules. The electrostatic potential (ESP) maps of the compounds were generated to check the electron density and polarity of the molecules. Through these maps, we can interpret the electron-rich and deficient areas, which are important for understanding the bonding of the compounds with the protein. Using HOMO and LUMO energies, other chemical parameters were calculated using the mathematical equations as under [32]:

$$\text{Chemical potential } (\mu) = \frac{E_{\text{HOMO}} + E_{\text{LUMO}}}{2}$$

$$\text{Electronegativity } (\chi) = -\mu$$

$$\text{Hardness } (\eta) = \frac{E_{\text{LUMO}} - E_{\text{HOMO}}}{2}$$

$$\text{Softness } (S) = \frac{1}{\eta}$$

$$\text{Electrophilicity Index } (\omega) = \frac{\mu^2}{2\eta}$$

2.7. Molecular dynamics simulations

The best docked molecules, reference (CDK2_HJK), and apo protein were subjected to classical molecular dynamics simulations. The simulations were conducted for 500 ns using GROMACS 2022.4 software with the AMBER99SB force field [33,34]. The ligand topologies and parameter files were generated using the ACPYPE tool based on the Amber atom-type force field [35]. To study the charge on the ligands, the antechamber AM1-BCC method was implemented [36]. Each system was solvated in a dodecahedron-shaped box with the SPC216 water model, placed with a minimum distance of 1 nm from the centre of the protein. An equivalent quantity of counter ions (Na⁺/Cl⁻) was added in the solvated box to neutralize the prepared systems. Initially, the systems were energy minimized using the steepest descent algorithm to remove the undesired clashes from the docked complexes. The maximum force below 1000 kJ mol⁻¹ nm⁻¹ (50000 steps) was employed to conduct the energy minimization. After energy minimization step, equilibration was carried out using position restraints within NVT (constant number [N], constant volume [V], and constant temperature [T]) and NPT (constant number [N], constant pressure [P], and constant temperature [T]) ensembles for one ns each. The NVT equilibration was conducted at a fixed volume for 100 ps and at a temperature of 300 K,

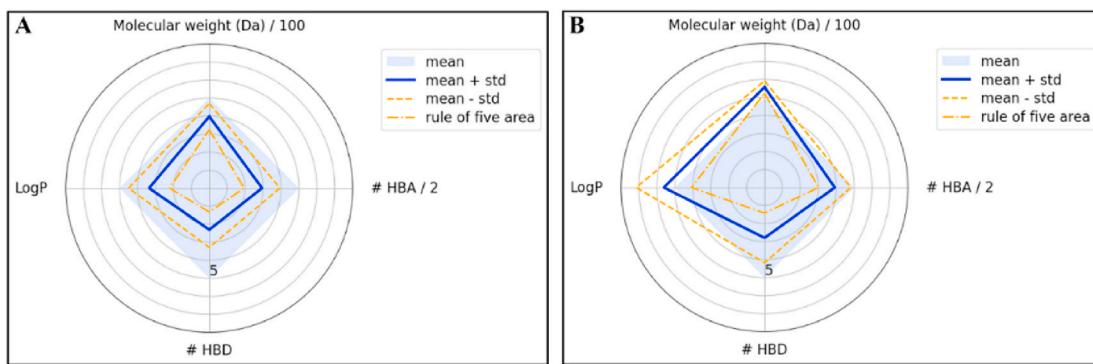


Fig. 1. The radar plot displays the physicochemical properties of the compounds that follow Lipinski's RO5 rule, as shown in 2A, and don't follow the RO5 rule, as shown in 2B.

Table 1

Assessment of the generated Machine learning models based on parameters like sensitivity, specificity, accuracy, and AUC for CDK2 data.

Models	Sensitivity	Specificity	Accuracy	AUC
RF	0.85	0.74	0.80	0.88
SVM	0.85	0.78	0.82	0.88
ANN	0.79	0.72	0.76	0.83

employing the modified Berendsen thermostat algorithm [37]. The Parrinello-Rahman barostat [38] was implemented during NPT equilibrium to maintain the system at a constant pressure of 1 bar. The Particle Mesh Ewald (PME) method [39] was applied using a 1.2 nm cut-off to efficiently compute long-range electrostatic interactions and Coulombic and van der Waals interactions. The Linear Constraints Solver (LINCS) algorithm [40] was employed to impose constraints on bond lengths. Subsequently, a 500 ns production run using the default parameters was performed using the NPT ensemble. To check the stability of the docked complexes, properties such as protein-ligand root mean square deviation (RMSD), root mean square fluctuations (RMSF), hydrogen bond formation, radius of gyration (RoG), solvent accessible surface area (SASA), principal component analysis (PCA), free energy landscape (FEL), dynamic correlation matrix (DCCM), and interaction energies (van der Waals and Electrostatic) were examined. These properties were calculated and plotted by using GRACE software [41]. Thereafter, the docked candidates were subjected to free energy calculations.

2.8. Binding free energy analysis

The docked molecules were analyzed via the Molecular Mechanics/Poisson-Boltzmann Surface Area (MM/PBSA) technique with the gmx_mmpbsa tool, applying default parameters set to idecomp of 2 and dec_verbose of 3 [42]. The analysis also involved setting the print residue option to a default value of 4 Å. This technique is widely used for calculating the binding free energy of protein-inhibitor complexes. This approach combines molecular mechanics energy terms with continuum solvation models to estimate the binding free energy (ΔG_{bind}) of protein-ligand complexes. In general, the binding free energy is calculated as the difference in the free energy of the complex and the combined free energy of the ligand and protein, calculated separately as shown in the following equation.

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

G_{complex} , G_{protein} , and G_{ligand} represent the free energies of the protein-ligand complex, the unbound protein, and the unbound ligand, respectively. Each of these free energies can be estimated using the following equation.

$$G = E_{\text{MM}} + G_{\text{sol}} - TS$$

The molecular mechanics energy (E_{MM}) includes bonded, electrostatic, and van der Waals interactions. The solvation free energy (G_{sol}) consists of polar contributions calculated using the Poisson-Boltzmann equation and non-polar contributions estimated from the solvent-accessible surface area (SASA). Stable trajectories were considered for conducting the free energy calculations. The selection of the time duration was based on the RMSD plots. The frames were selected at a consistent interval of 100 ps to cover various conformational states of the docked complexes successfully.

3. Results and discussion

3.1. Analysis of machine learning classification models

Three machine learning classifiers, including RF, ANN, and SVM, were employed to conduct classification QSAR studies using the Open-CADD pipeline. The models were constructed using the MACCS fingerprints generated from the RDKit toolkit to encode the molecules. We collected a dataset of 2084 CDK2 inhibitors from ChEMBL, and a total of 1657 molecules were used after cleaning. These CDK2 inhibitors were employed for Lipinski's RO5 rule criteria analysis, in which we got 1589 compounds that fulfilled the criteria. Fig. 1 presents the radar (spider) plots illustrating the physicochemical properties of the dataset compounds, categorized based on their compliance with or deviation from Lipinski's Rule of Five (RO5). The blue square represents the region where the physicochemical properties of a compound comply with Lipinski's Rule of Five (RO5). The solid blue line indicates the mean values, while the orange dashed line denotes the standard deviations. Fig. 1A shows the radar plot for the compounds that follow Lipinski's Rule of Five, and Fig. 1B shows the radar plot for the compounds that violate the RO5 rule. As shown in Fig. 1, the mean value and the standard deviations both fall under the RO5 criteria. As illustrated in Fig. 1B, certain deviations from Lipinski's Rule of Five were observed, specifically in the molecular weight (MW) and logP values. These violations indicate that some data exceed the acceptable thresholds for these parameters, potentially impacting their drug-like properties and oral availability. After the filtration of molecules based on the RO5 rule, the remaining 1589 compounds were divided into active and inactive binary labels based on the pIC_{50} cutoff of 6.3. This resulted in 884 active and 705 inactive compounds that were further split into a training/test (80/20) set. The classifiers were developed using the Scikit-learn library in Python, a widely utilized framework for implementing machine learning algorithms. To optimize the performance of each model, hyperparameters were carefully tuned (Supplementary Table 1). For RF, the parameter was set to $n_{\text{estimators}} = 100$, criterion = 'entropy'; for SVM, kernel = 'rbf', gamma = '0.1', probability = True; for ANN,

Table 2

Assessment of the generated Machine learning models based on parameters like sensitivity, specificity, and AUC using cross-validation on CDK2 data.

Models	Sensitivity	Specificity	Accuracy	AUC
RF	0.85 ± 0.01	0.76 ± 0.04	0.81 ± 0.02	0.88 ± 0.02
SVM	0.86 ± 0.01	0.76 ± 0.05	0.81 ± 0.02	0.88 ± 0.01
ANN	0.81 ± 0.01	0.69 ± 0.05	0.76 ± 0.02	0.83 ± 0.02

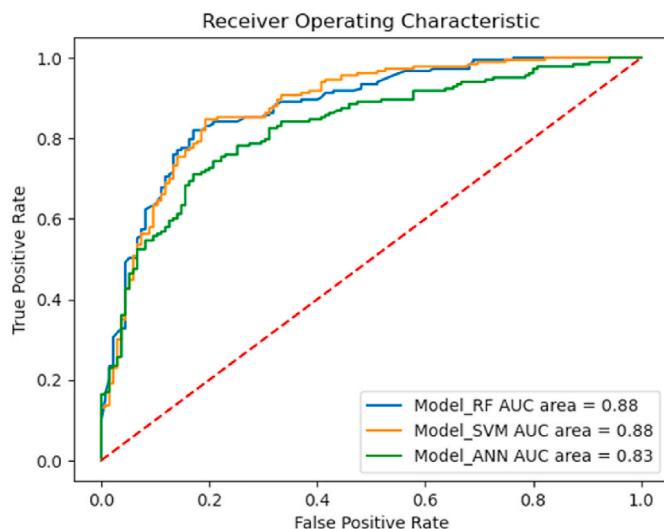


Fig. 2. Receiver Operating Characteristic (ROC) graph depicting the True Positive Rate (TPR) versus False Positive Rate (FPR) for three machine learning models, RF, SVM, and ANN.

Table 3

Assessment of the generated Machine learning models based on parameters like sensitivity, specificity, and AUC using cross-validation on CDK2 data.

Models	Sensitivity	Specificity	Accuracy	AUC
RF	0.86 ± 0.02	0.79 ± 0.04	0.83 ± 0.01	0.91 ± 0.01
SVM	0.96 ± 0.02	0.52 ± 0.04	0.76 ± 0.02	0.89 ± 0.04
ANN	0.85 ± 0.03	0.78 ± 0.02	0.81 ± 0.02	0.89 ± 0.02

hidden_layer_sizes = (5, 3), random_state = 'SEED'.

The performance of these three models was evaluated using the key metrics, including sensitivity, specificity, and the area under the ROC curve (AUC) on the test set, as shown in Table 1.

Further, we employed five-fold cross-validation on all three models, and the results are shown in Table 2. The ROC curve of these models is depicted in Fig. 2. Moreover, we also executed the five-fold validation for molecules encoded using Morgan fingerprints and not MACCS keys as presented in Table 3. We observed that among all three models, the RF and SVM models displayed almost similar sensitivity, specificity, accuracy, and AUC values, as shown in Tables 1, 2, and 3. However, the SVM model demonstrated a specificity of 0.52 ± 0.04 during the cross-validation using Morgan fingerprints, which is below the acceptable threshold for reliable classification (Table 3). Ideally, specificity values should exceed 0.7–0.8 to ensure a model's ability to correctly identify true negatives while minimizing false positives. Therefore, we selected the RF model that has better sensitivity, specificity, accuracy, and AUC for cross-validation.

The virtual screening using the RF model was carried out using the external coconut database of 4,77,975 molecules. After screening, the most suitable 327 molecules with a predicted probability above 85 % were selected for further filtering against the PAINS and unwanted substructures. After this filtration, we retrieved 309 molecules subjected to molecular docking analysis.

3.2. Molecular docking studies

All the screened and filtered 309 molecules were subjected to molecular docking using SeeSAR software. Before docking, redocking was performed to validate the software. The PDB ID 6Q4G with co-crystallized inhibitor HJK was selected to conduct the redocking calculations. From the redocking calculations, it was observed that the redocked pose of HJK was generated with an RMSD of 1.7 Å, docking score of -5.13 kcal/mol, and displayed binding with residues Thr14A, Lys33A, Glu81A, Leu83A, and Asp145A (Fig. 3).

It is observed that, similar to the crystallized pose, our software could show the interaction with the crucial residues, Thr14A, Lys33A, Glu81A, and Leu83A [43–45]. Hence, our redocking calculations validate our software. On conducting docking using 309 filtered molecules, it was observed that only 299 molecules could bind in the active site of the CDK2 enzyme (Supplementary Table 2). After docking calculations, the docked candidates were further shortlisted based on the reference redocking score of -5.13 kcal/mol. Of 299 docked molecules, 245

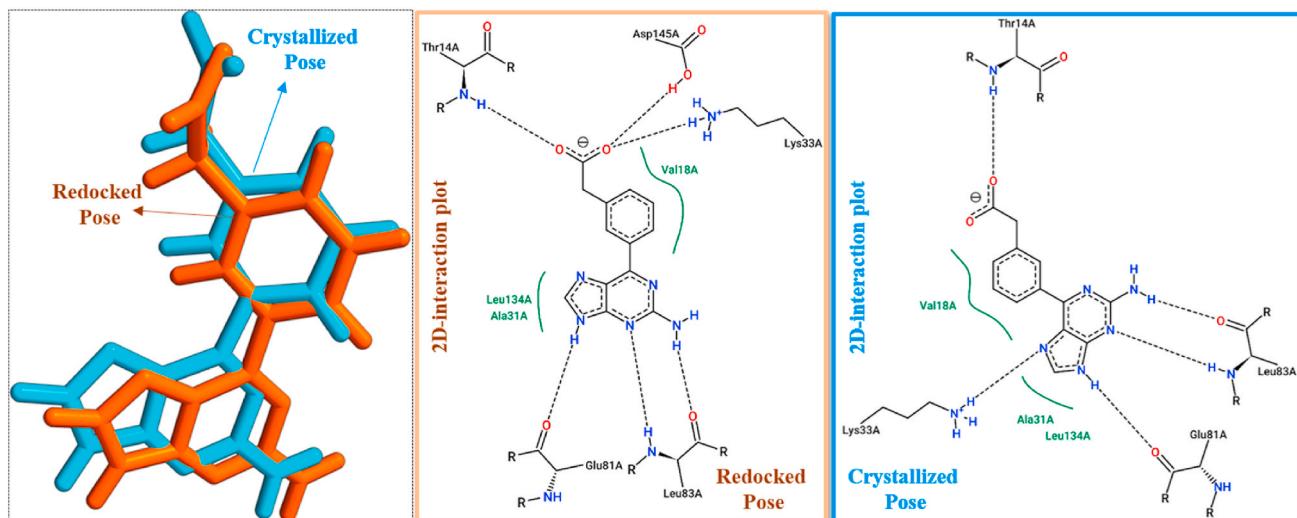


Fig. 3. Pictorial representation of superimposed redocked (orange) and crystallized (blue) pose along with the 2D interaction plot of redocked (orange) and crystallized (blue) pose. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

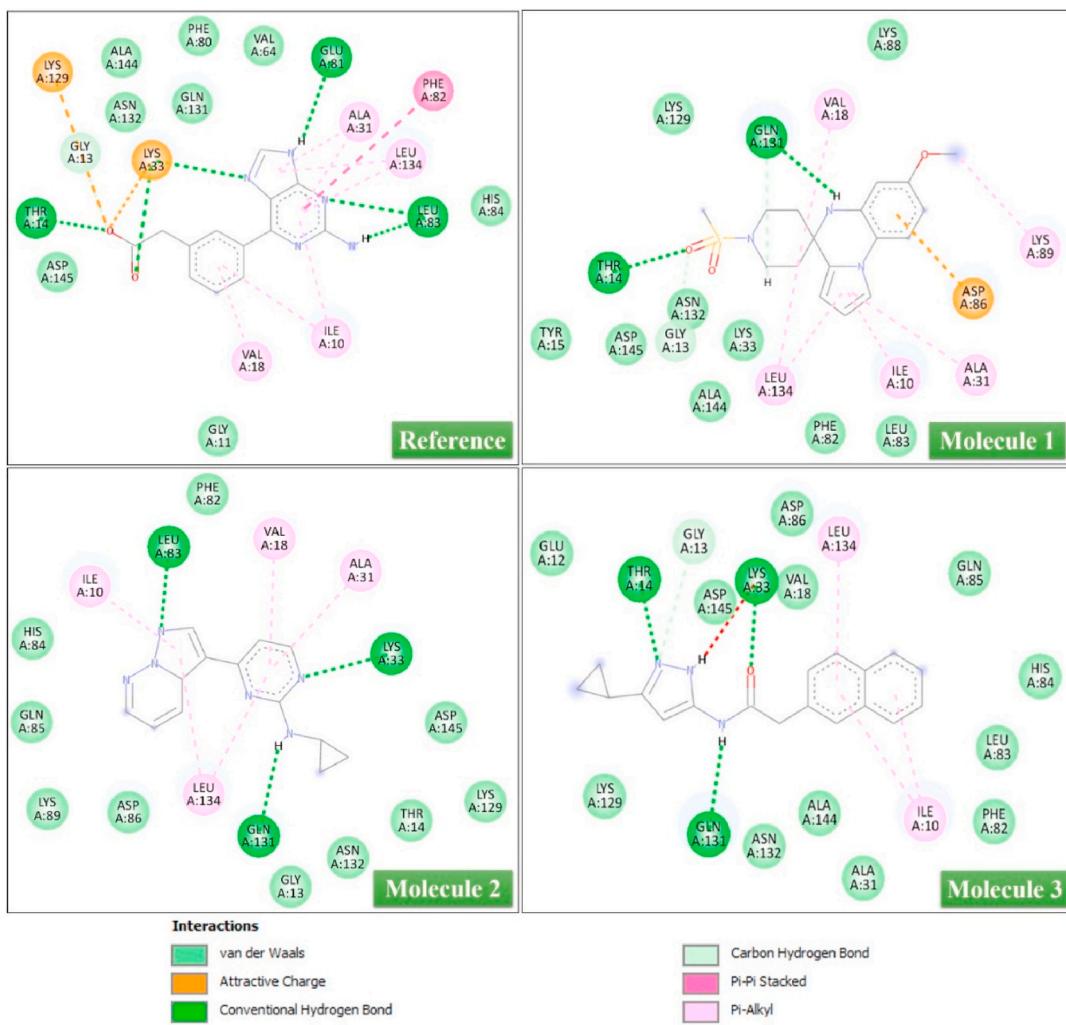


Fig. 4. 2D-interaction plot of the docked candidates Molecule 1, Molecule 2, and Molecule 3 along with the reference compound.

candidates displayed better binding affinity than the reference ([Supplementary Table 2](#)). The top 40 molecules were shortlisted among these candidates for further analysis based on their significantly higher docking score (-15.00 kcal/mol). The molecules were selected based on their superior binding score and favorable interactions. These 40 molecules were further subjected to ADMET analysis to study the pharmacokinetic properties of the molecules ([Supplementary Table 3](#)). Apart from this, we have added interaction table of top 10 compounds retrieved from the docking calculations ([Supplementary Table 4](#)).

3.3. ADME/T analysis

Assessing the ADME properties of the candidates is a crucial aspect of early-stage drug discovery, providing valuable insights into their pharmacokinetics (pk) and safety profiles of the molecules. To perform ADME studies, the SwissADME tool was utilized to predict the pk parameters of the shortlisted 40 docked molecules as mentioned in the [Supplementary Table 5](#). On analyzing the pk profile of the candidates, it was observed that only 3 molecules were able to display the acceptable range of the properties ([Supplementary Table 6](#)). Properties based on the Lipinski rule of five, Ghose, Veber, Egan, and Muegge rules were checked to determine their potential for oral bioavailability and suitability as lead compounds. Among these, only three compounds met all five criteria for drug-likeness. As per Lipinski's rule of 5, a compound must possess a molecular weight $\leq 500 \text{ Da}$, ≤ 5 hydrogen bond donors, ≤ 10 hydrogen bond acceptors, and $\leq 2 \text{ logP}$ value to have a good

bioavailability. Also, as per Ghose's rule, molecular weight of the compound should be in-between 160 to 480 Da, value of logP should vary in-between -0.4 to 5.6 , molar refractivity of the compounds should be within the range of 40–130, and a total of 20–70 atoms for a molecule to possess drug-like property. Veber's rule states that for a molecule to show good absorption and molecular flexibility, the TPSA of a molecule should be $\leq 140 \text{ \AA}^2$, a maximum of 12 hydrogen bond acceptors and donors groups in combination, and a maximum of 10 rotatable bonds. The shortlisted compounds also satisfied the criteria of Egan's rule ($\text{logP} \leq 5.88$, $\text{TPSA} \leq 131.6 \text{ \AA}^2$), suggesting they can be well absorbed in the intestines and may cross the blood-brain barrier. Additionally, compliance with Muegge's rule (molecular weight between 200 and 600 Da, $\text{logP} \leq 5$, ≤ 5 hydrogen bond donors, ≤ 10 hydrogen bond acceptors, $\text{TPSA} \leq 150 \text{ \AA}^2$, 1 to 7 rings, 5 to 35 carbon atoms, and 1 to 15 hetero-atoms) confirmed their suitability as lead compounds. Moreover, these three finalized molecules have 0 alerts as per PAINS and Brenk filters. Additionally, the toxicity analysis was conducted by using pkCSM. The reported parameters include AMES toxicity, hERG I inhibitor, and Skin Sensitization ([Supplementary Table 6](#)). As per the outcome, all three selected molecules with ID CNP0537535.0 (Molecule 1), CNP0590537.0 (Molecule 2), and CNP0522611.0 (Molecule 3) exhibit a favorable toxicity profile, thus making them potentially safer and suitable for further biological evaluations. Consequently, the shortlisted molecules were subjected to DFT calculations to study the stability and reactivity of the filtered molecules.

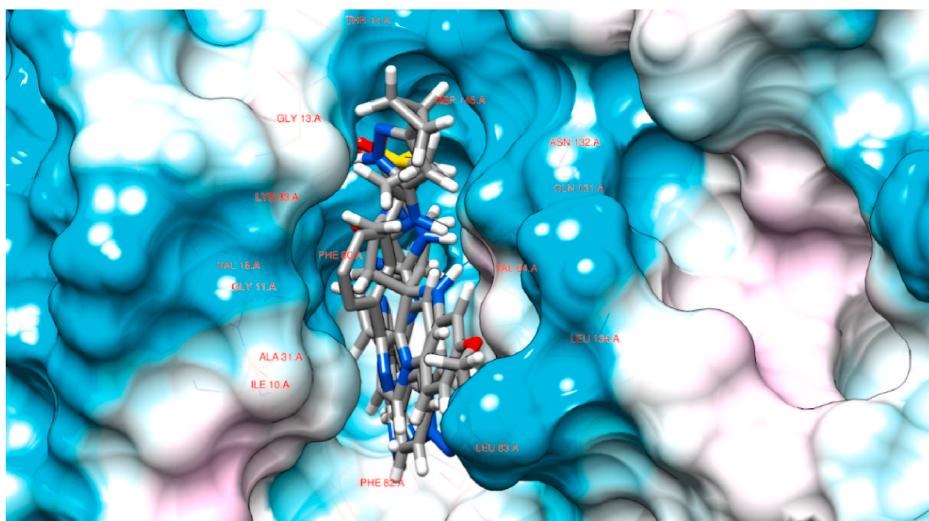


Fig. 5. 3D coordinates of CDK2 protein docked with Molecule 1, Molecule 2, Molecule 3, and reference (HJK). Amino acids of the binding site are represented in red colour. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 4
Table of HOMO-LUMO energy gap (eV) of the selected compounds solved at B3LYP/6-31+g(d) level of theory.

Molecules	E _{HOMO} (eV)	E _{LUMO} (eV)	Energy gap (E _{LUMO} - E _{HOMO}) (eV)
Molecule 1	-5.51	-0.70	4.81
Molecule 2	-6.07	-2.13	3.94
Molecule 3	-6.28	-1.81	4.47
Reference (HJK)	-6.00	-1.83	4.17

3.4. Interaction analysis of three compounds

The 2D interaction plots and docking score of the three shortlisted molecules and the reference ligand are shown in Fig. 4. **Molecule 1** possesses the highest docking score of -15.94 kcal/mol, trailed by **Molecule 2** (-15.46 kcal/mol) and **Molecule 3** (-15.30 kcal/mol). The 2D interaction plot of **Molecule 1** showed that the compound showed hydrogen bonding interactions with amino acids Thr14, Gln131, and Asn132. The residues like Ile10, Ala31, Lys 89, and Leu134 were involved in showing Pi-alkyl interactions. Moreover, the amino acid Asp86 might contribute to the stability of the compound due to the formation of electrostatic interactions with **Molecule 1**. In case of **Molecule 2**, amino acids like Lys33, Leu83, and Gln131 were involved in forming hydrogen bonding, Ile10, Val18, Ala 31, and Leu134 were responsible for showing Pi-alkyl interactions. **Molecule 3** displays hydrogen bonding with Gly13, Thr14, Lys33, Gln131, and Pi-alkyl interactions with Ile10 and Leu134. Apart from all the mentioned conventional hydrogen bonding, pi-alkyl, and charge interactions, some residues like Gly13, Thr14, Val18, Lys33, Phe 82, Asp86, Lys 129, Asn132, and Asp145 were involved in showing van der Waals interactions, thus reflecting the shape complementarity of the docked protein-ligand complexes. Comparatively, Reference shows interaction with Thr14, Glu 81, and Leu83, and pi-alkyl interactions with Ile10, Val18, Ala31, and Leu134. From the 2D (Fig. 4) and 3D interaction plot (Fig. 5), it is observed that all the docked molecules possess the conserved interactions with the residues Lys33 and Asp145, which are considered crucial for the inhibition of the enzyme [20,43]. Also, as per the criteria of the molecule to be efficient CDK2 inhibitors, they must possess flat heterocyclic rings [20]. The shortlisted molecules, **Molecule 1**, **Molecule 2**, and **Molecule 3**, contain aromatic heterocyclic rings [20]. Furthermore, among them, **Molecule 2** possesses an extended fused heterocyclic system that can potentially improve the inhibition. Thus, all three molecules possess the capability of inhibiting the CDK2

enzyme. These docked compounds were further subjected to molecular dynamics simulations.

3.5. Frontier molecular orbital

All three shortlisted compounds and the reference molecule (Supplementary Figure 1) were subjected to DFT using the Gaussian 09 program [29]. Table 4 and Supplementary Table 7 represent the electronic properties of the shortlisted compounds. Table 4 depicts the energy of HOMO and LUMO and energy gap between them. To understand the electronic properties affecting the molecular binding, we analyze the FMOs (HOMO-LUMO) of the shortlisted molecules (Molecule 1, 2, 3, and Reference). The HOMO energy depicts the donating capacity of the molecule, whereas, LUMO energy depicts the accepting capability of the molecule during binding interaction with the macromolecule or any target. The energy gap determines the stability and molecular reactivity. A molecule with a smaller energy gap represents higher chemical reactivity, thus it can interact well within the target binding site. Among the shortlisted molecules, **Molecule 2** possess the smallest energy gap of 3.94 eV, thus, suggesting greater reactivity and strong interaction in the binding site of the enzyme. However, **Molecule 1** with the energy gap of 4.81 eV indicate more stable, but is expected to be less reactive. With the help of HOMO and LUMO energy, we can calculate the other parameters related to the reactivity, stability and electron affinity, i.e., chemical potential, electronegativity, hardness, softness and electrophilicity index, as shown in the Supplementary Table 5. It is evident from Table 4 that HOMO energy of **Molecule 1** (-5.51 eV) is highest among all, thus suggesting that the compound have better electron donating capability. **Molecule 1** is trailed by **Molecule 2** (-6.07 eV) and **Molecule 3** (-6.28 eV). However, if we observe, the reference molecule showed the HOMO energy of -6.00 eV. The energy of **Molecule 2** is slightly closer to the energy of the reference compound. Similarly, the LUMO energy of **Molecule 1** is higher (-0.70 eV), suggesting its weaker electron-accepting tendency. **Molecule 1** is followed by **Molecule 3** (-1.81 eV) and **Molecule 2** (-2.13 eV). The LUMO energy of the reference compound is -1.83 eV. It is observed that the energy of **Molecule 2** and the reference compounds is close. On comparison, **Molecule 1** can be considered as the most stable and least reactive compound, **Molecule 2** is observed to be the most reactive compound, thus, can be considered as highly involved in electron transfer during drug-target interactions. The remaining **Molecule 3** possesses a balanced reactivity and stability, making it a versatile compound that shows moderate stability compared to **Molecule 2** but better reactivity

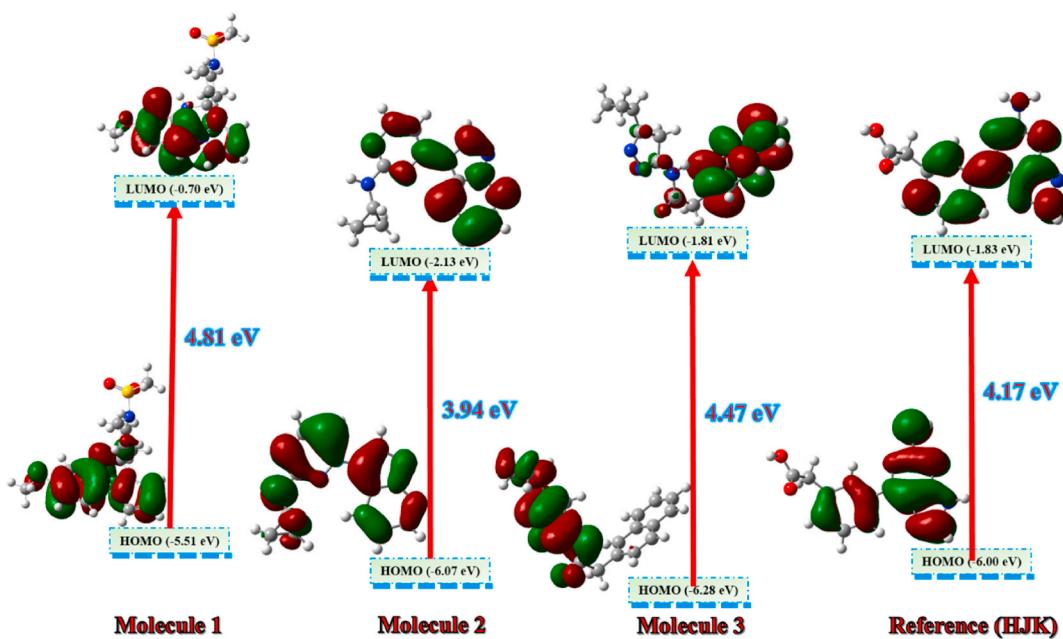


Fig. 6. Frontier molecular orbitals (HOMO and LUMO) for Molecule 1, Molecule 2, Molecule 3 and reference generated at B3LYP/6–31+g(d) level of theory.

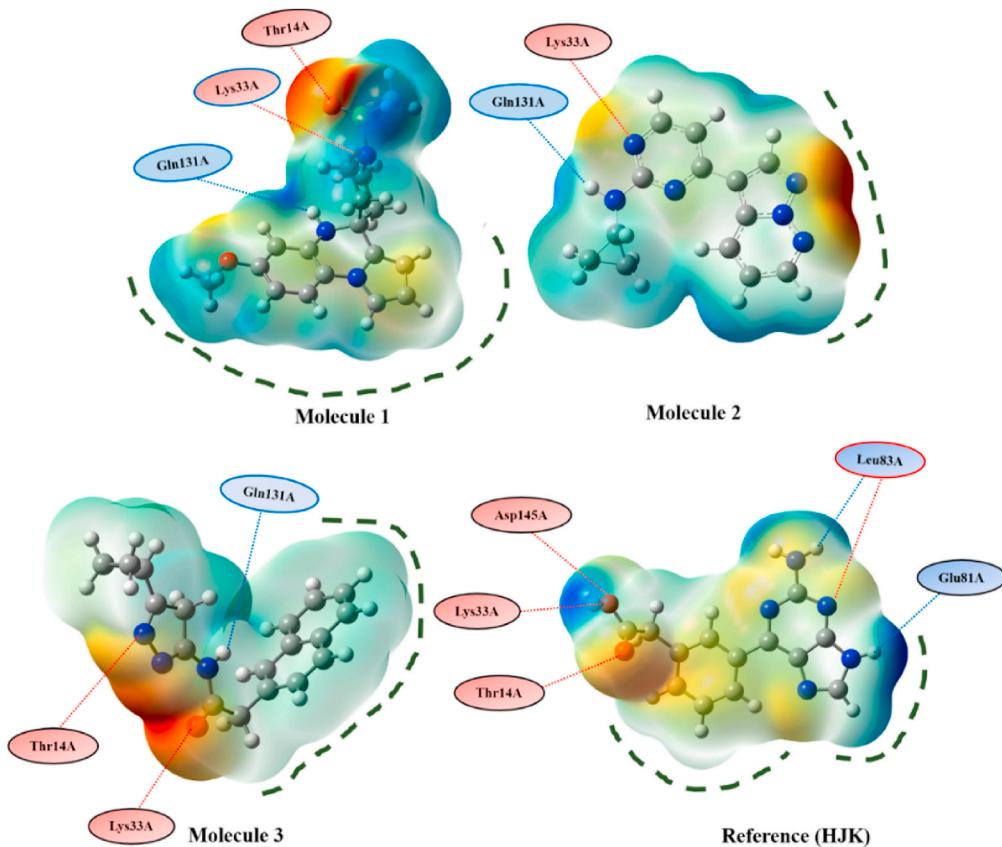


Fig. 7. The electrostatic potential (ESP) maps for Molecule 1, Molecule 2, Molecule 3, and reference generated at B3LYP/6–31+g(d) level of theory.

than **Molecule 1**.

Along with the HOMO and LUMO energies, the pictorial form of the frontier molecular orbitals are shown in Fig. 6. The red region in the FMO represents higher electron density and, thus, has good electron donation capability (nucleophilicity). The green area indicates lower electron density, making them suitable electron acceptors

(electrophilicity). It is depicted from Fig. 6 that the electron donor region in the HOMO of **Molecule 1** is more in comparison to **Molecule 2**, followed by **Molecule 3**. Also, **Molecule 2** possesses the best electron acceptor region in the LUMO. Overall, **Molecule 2** is considered the best candidate owing to its high reactivity, **Molecule 1** is ideal for its stability, and **Molecule 3** acts as a versatile molecule.

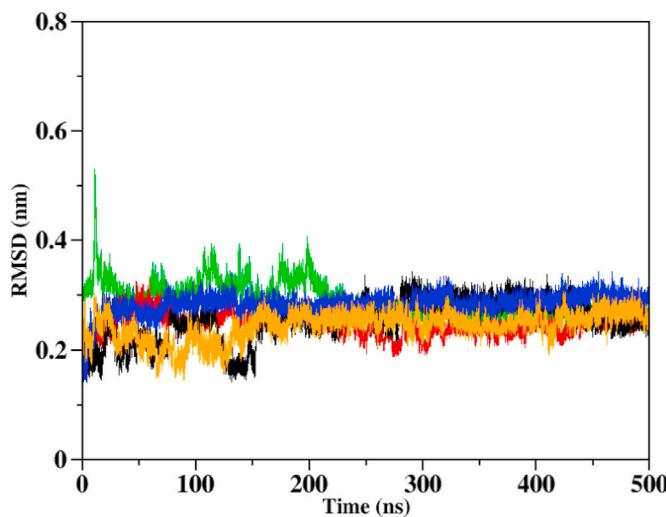


Fig. 8. The RMSD plot illustrating the conformational variations of protein backbone over a 500 ns simulation run for apo-protein (black), reference molecule (red), Molecule 1 (green), Molecule 2 (blue), and Molecule 3 (orange). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

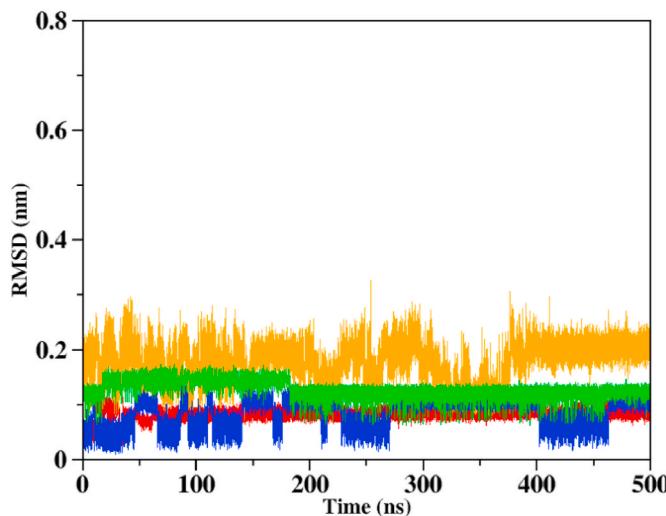


Fig. 9. The RMSD plot illustrating the variations of the ligand throughout a 500 ns simulation run for apo-protein (black), reference molecule (red), Molecule 1 (green), Molecule 2 (blue), and Molecule 3 (orange). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

3.6. Key chemical properties

From the HOMO and LUMO energies, various other electronic properties were calculated to study the electronic properties of the compounds. It is well documented in the literature that molecules with appropriate electronic properties can be optimized further to enhance the binding affinity, improve the bioavailability, and minimize the toxicity by controlling the reactivity. All the calculated parameters, such as chemical potential, electronegativity, hardness, softness, and electrophilicity index, are shown in [Supplementary Table 7](#). Chemical potential represents the molecule's tendency to give electrons, electronegativity describes the ability to attract electrons, hardness indicates the resistance to electron transfer, softness explains the chemical reactivity, and electrophilicity index describes the electron-accepting ability of the molecule. On observing [Supplementary Tables 7](#) and it is

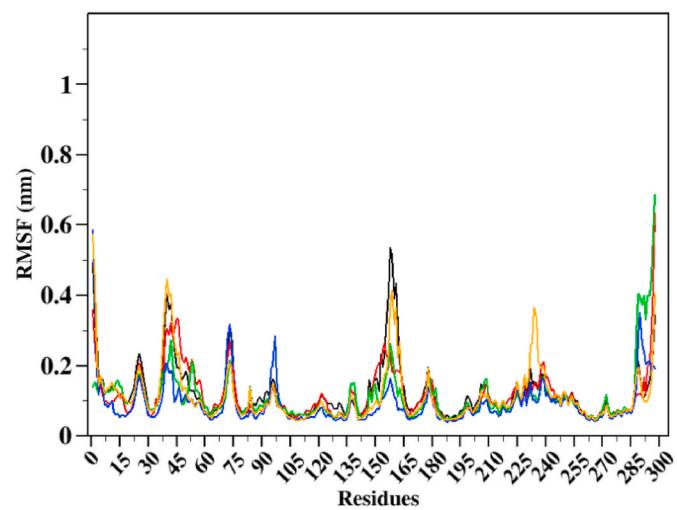


Fig. 10. The RMSF plot illustrating the residue-wise fluctuations of protein backbone over a 500 ns simulation run for apo-protein (black), reference molecule (red), Molecule 1 (green), Molecule 2 (blue), and Molecule 3 (orange). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

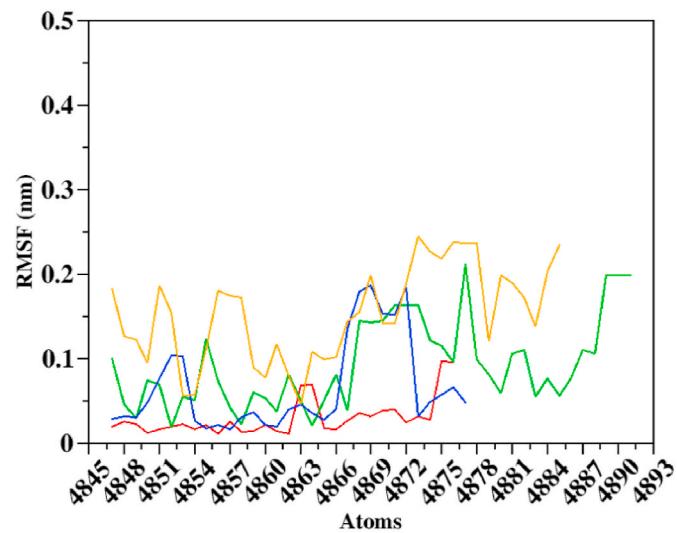


Fig. 11. The RMSF plot illustrating the atom-wise fluctuations of ligands over a 500 ns simulation run for apo-protein (black), reference molecule (red), Molecule 1 (green), Molecule 2 (blue), and Molecule 3 (orange). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

evident that **Molecule 2** possess the highest tendency to accept the electrons (electrophilic) owing to its chemical potential (-4.10 eV), electronegativity (4.10 eV) and electrophilicity index (4.27 eV), most reactive due to lowest hardness (1.97 eV) and high softness (2.41 eV). Thereafter, **Molecule 1** possesses a lesser electron accepting tendency due to a lesser chemical potential (-3.11 eV), low electronegativity (3.11 eV), most stable due to high hardness (2.41 eV) and low softness (0.42 eV). The remaining Molecule 3 is less reactive than **Molecule 2** but more reactive than **Molecule 1**. The electronegativity value of 4.05 eV represents a high capability of accepting electrons.

3.7. Electrostatic potential map analysis

Electrostatic potential (ESP) maps help to understand the molecular interactions, charge distribution, and binding affinity within protein-

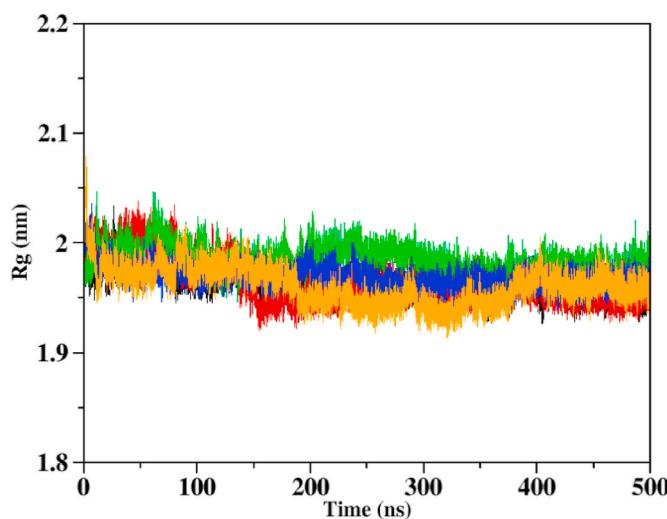


Fig. 12. Radius of gyration plot illustrating the compactness and structural stability of protein Co atoms over a 500 ns simulation run for apo-protein (black), reference molecule (red), Molecule 1 (green), Molecule 2 (blue), and Molecule 3 (orange). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

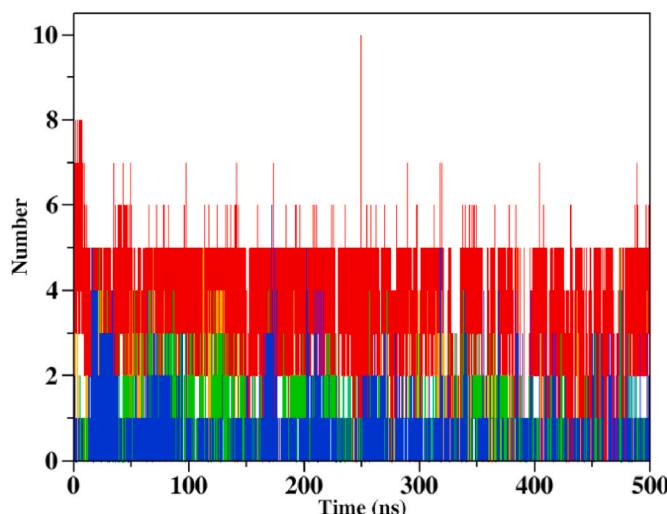


Fig. 13. Hydrogen bond plot of the protein-ligand complexes illustrating the number of H-bonds formed over a 500 ns simulation run for reference molecule (red), Molecule 1 (green), Molecule 2 (blue), and Molecule 3 (orange). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

ligand interactions. The ESP maps contain red, blue, and green/yellow regions, as shown in Fig. 7. Red represents the electron-rich area, blue represents the electron-deficient area, and green/yellow represents the neutral charge density. The electron-rich area tends to interact with the positively charged amino acids, the blue region interacts with negatively charged residues, and the green/yellow region interacts through van der Waals forces (hydrophobic interactions). As observed in Fig. 7, Molecule 1 interacts with the amino acids Thr14A and Lys33A through electron-rich regions, forming hydrogen bonds. Also, the greenish-blue region interacts with Gln131A, suggesting a weaker hydrogen bonding. In Molecule 2, the amino acid Lys33A interacts with the red/yellow region of the compound, suggesting an electrostatic interaction. The interaction of Gln131A with the weak electronegative region suggests the formation of hydrogen bonding. The electron-rich region of Molecule 3 interacts with Lys33A and Thr14A, the blue region interacts with Asp145A,

indicating strong bonding, and Gln131A interacts with the blue region of the molecule, indicating interaction. The ESP map of reference shows the interaction of Lys33A, Thr14A, and Asp145A with the negative region of the compound; amino acids Glu81A and Leu83A form an interaction with the blue region of the molecule. Thus, the presence of positively and negatively charged regions over the molecule's surface ensures the formation of strong and various interactions.

3.8. Molecular dynamics simulation

Molecular dynamics of apo protein, selected docked molecules, i.e., Molecule 1 (M1), Molecule 2 (M2), Molecule 3 (M3), and the reference molecule HJK (6Q4G), were carried out for a time duration of 500 ns. The simulations aimed to investigate structural stability, binding interactions, and conformational changes observed in the docked complexes over time. In this examination, the key indicators of structural flexibility and stability, i.e., RMSD, RMSF, Rg, and H-bonds, were evaluated to estimate the stability of the complex. We have also evaluated SASA to analyze the surface area of protein, PCA, to study the conformational alteration in protein, FEL, to provide insight into the energy distribution in the protein, and DCCM to study the influence of ligand binding on the internal motion of the protein. Additionally, the interaction energy between the protein and ligand was also calculated. The mean values of all the properties are presented in Supplementary Table 8.

3.9. Protein-ligand RMSD analysis

RMSD quantifies how the structure of a protein-ligand complex differs over time from its initial conformation. A low RMSD value indicates the stability of the complex, thus displaying minimal structural changes, implying a consistently strong binding interaction throughout the simulation [46,47]. Fig. 8 depicts the RMSD plot for the CDK2_docked complexes generated from the 500 ns simulation run. The mean average RMSD backbone value for CDK2-APO, CDK2-HJK, CDK2-M1, CDK2-M2, and CDK2-M3 were reported to be 0.254 ± 0.038 nm, 0.246 ± 0.020 nm, 0.288 ± 0.033 nm, 0.283 ± 0.020 nm, and 0.244 ± 0.026 nm, respectively. As shown in Fig. 8, CDK2-HJK, CDK2-M1, CDK2-M2, and CDK2-M3 show almost similar and stable behaviour. All the complexes attain stability after a time duration of 300 ns. CDK2-HJK and CDK2-M3 showed lower RMSD values than CDK2-APO, indicating that the protein stabilizes after binding with the inhibitor. Among all the complexes, CDK2-M3 exhibits an RMSD lower than the reference CDK2-HJK complex with a reduction of 0.002 nm. CDK2-M1 and CDK2-M2 exhibit slightly higher RMSD values. However, their RMSD plot reveals consistent variations after 200 ns. This suggests that compound CDK2-M1, CDK2-M2, and CDK2-M3 complexes have attained stable converged behaviour during the simulations.

Ligand RMSD analysis for the docked complexes was carried out to check the ligand's binding stability (Fig. 9). As shown in Supplementary Table 8, the average RMSD values for ligands HJK, M1, M2, and M3 are 0.083 ± 0.009 nm, 0.129 ± 0.015 nm, 0.085 ± 0.026 nm, and 0.178 ± 0.035 nm, respectively. The CDK2-M1 complex displays a similar trend to the reference and relatively stable ligand RMSD behaviour throughout the simulation. Docked complexes of CDK2 with M2 and M3 exhibit a minor fluctuation in the ligand RMSD plot. The average RMSD value of M2 is nearly identical to that of the reference HJK ligand. However, they eventually converged and attained stability, which is the most critical aspect of simulation.

3.10. Protein-ligand RMSF analysis

RMSF quantifies the dynamic mobility of specific atoms or amino acid residues across time intervals. Lower RMSF values signify stability in particular regions of the protein-ligand complex, indicating minimal movement and suggesting strong and stable interactions in their

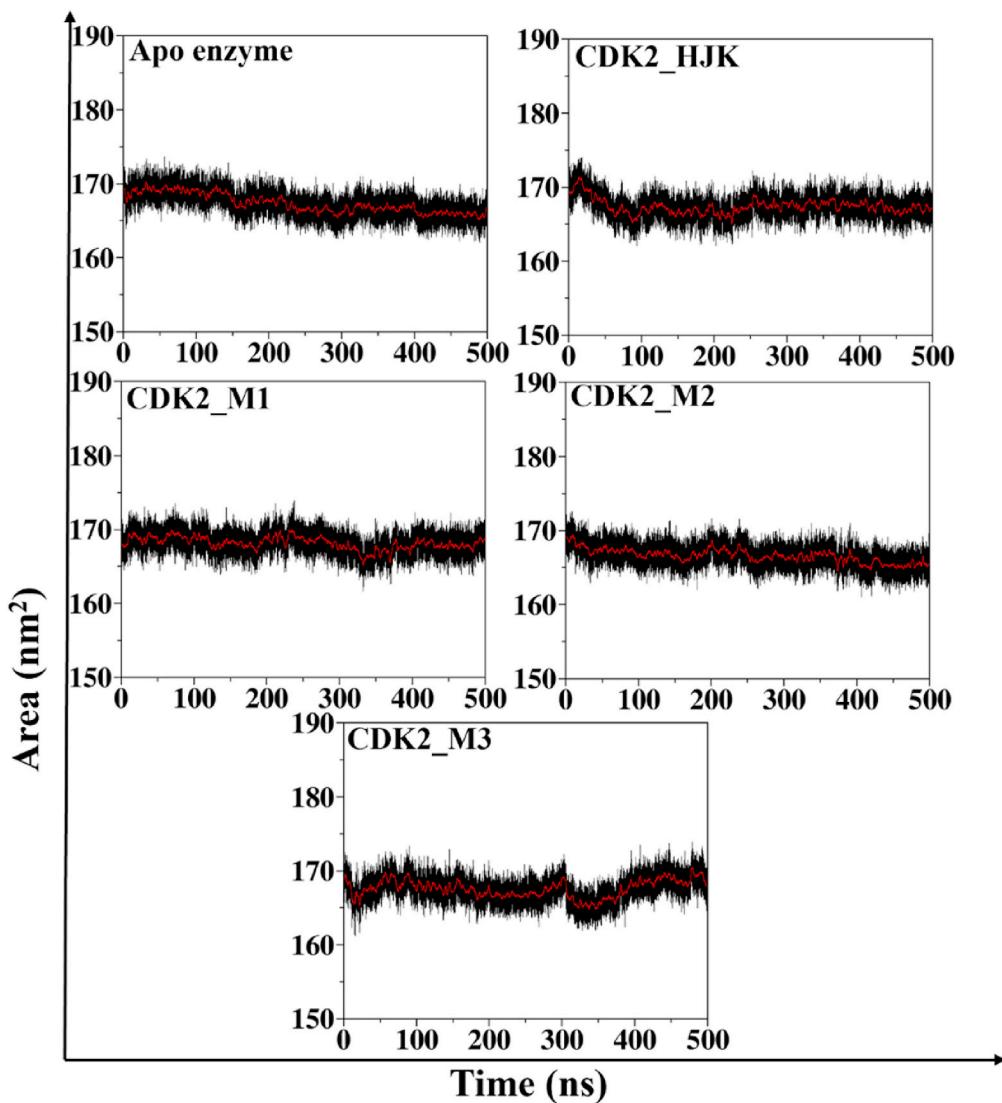


Fig. 14. The solvent accessible surface area (SASA) plot of CDK2-APO, CDK2-HJK, CDK2-M1, CDK2-M2, and CDK2-M3 generated over a 500 ns simulation run.

structural framework [48]. Fig. 10 illustrates the protein RMSF plot of the docked complexes over a time duration of 500 ns. The average protein backbone RMSF values for CDK2-APO, CDK2-HJK, CDK2-M1, CDK2-M2, and CDK2-M3 are 0.122 ± 0.079 nm, 0.116 ± 0.070 nm, 0.111 ± 0.076 nm, 0.091 ± 0.060 nm, and 0.113 ± 0.080 nm, respectively. RMSF values for all the complexes CDK2-HJK, CDK2-M1, CDK2-M2, and CDK2-M3 are lower than CDK2-APO enzyme. This shows that binding of a ligand to the protein induces stabilization, minimizes conformational fluctuations, and enhances the overall stability of the protein structure. The RMSF value of all three complexes of M1, M2, and M3 displayed lower RMSF values in comparison to the reference CDK2-HJK. The CDK2-M2 complex displayed the least RMSF value, showing the most stable protein-ligand complex with reduced flexibility.

The mobility of the ligands within the protein binding site was assessed by examining the ligand RMSF plots. The average ligand RMSF values for HJK, M1, M2, and M3 are 0.030 ± 0.022 nm, 0.092 ± 0.052 nm, 0.066 ± 0.054 nm, and 0.153 ± 0.057 nm, respectively. The RMSF plot for the ligands is presented in Fig. 11. On observing the figure, the CDK2-M2 exhibits the lowest RMSF value among the docked complexes, indicating reduced fluctuations and stable binding. Overall, the analysis of the Protein and Ligand RMSF indicates that molecule M2 forms a stable complex with CDK2, exhibiting lower RMSF values.

3.11. Radius of gyration (R_g)

The R_g examines how the protein mass is distributed relative to its center of mass at each trajectory frame. It compares the distribution with the hydrodynamic radius obtained from the experiments. The R_g analysis evaluates the stability and compactness of protein structures throughout the molecular dynamics simulations. As depicted in Fig. 12, all the docked complexes display consistently stable R_g values during the simulation. This stability indicates the preservation of structural integrity, with no significant conformational expansion or contraction throughout the simulation. The mean R_g values for CDK2-APO, CDK2-M1, CDK2-M2, and CDK2-M3 are 1.964 ± 0.013 nm, 1.965 ± 0.020 nm, 1.985 ± 0.012 nm, 1.970 ± 0.010 nm, and 1.961 ± 0.018 nm, respectively. CDK2-HJK exhibits nearly equivalent R_g values as CDK2-APO, suggesting comparable compactness and structural integrity throughout the simulation. CDK2-M3 exhibited a lower R_g value than CDK2-APO and the reference CDK2-HJK by 0.003 nm and 0.004 nm, respectively. The lowered R_g values signify a more compact structure, implying that the interaction with M3 results in protein stabilization and decreased protein structure flexibility.

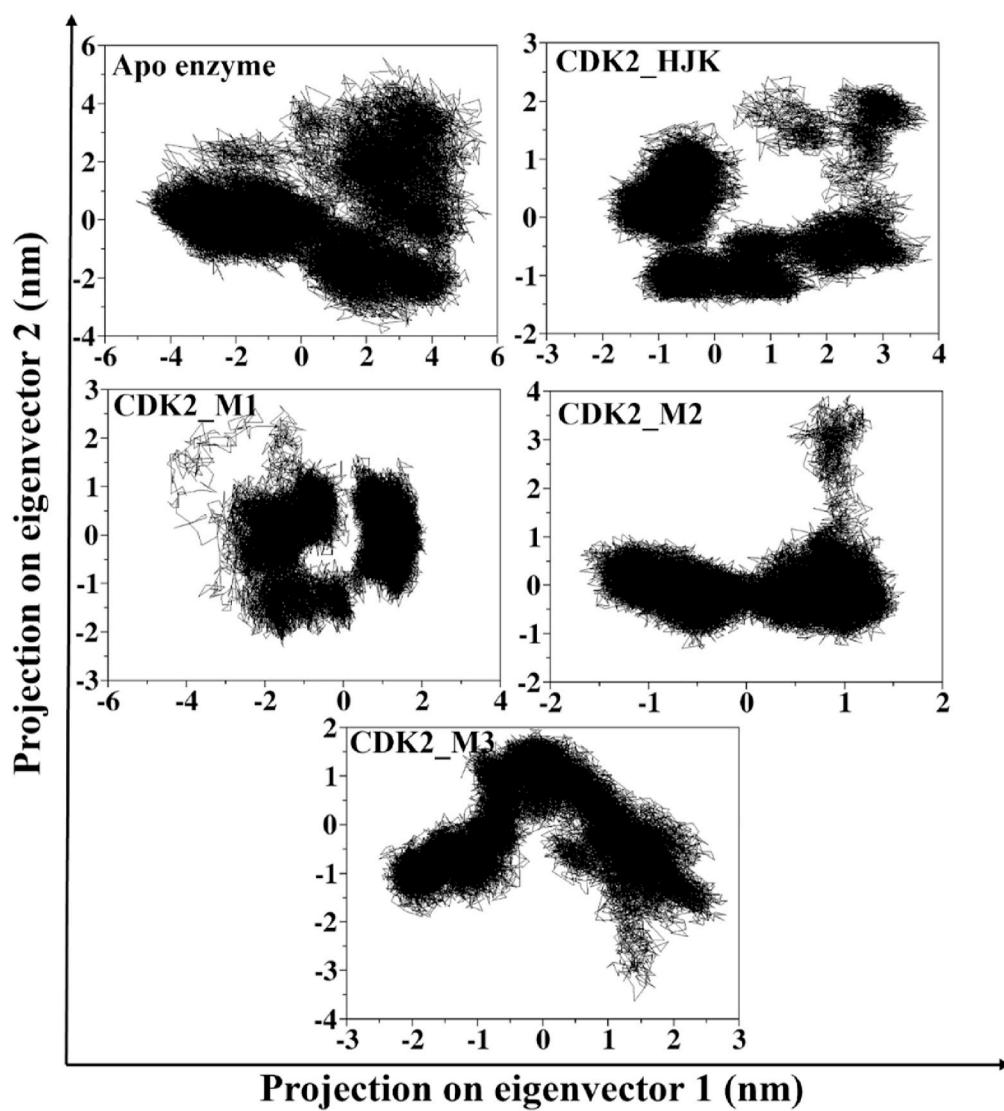


Fig. 15. The principal component analysis (PCA) plot showing projection of motion of C α chain of the protein along PC1 and PC2 principal components for CDK2-APO, CDK2-HJK, CDK2-M1, CDK2-M2, and CDK2-M3 over the molecular dynamic simulation of a 500 ns.

3.12. Hydrogen bonding analysis

Hydrogen bond formation between a protein and a ligand is a crucial molecular interaction that plays a pivotal role in modulating binding affinity and stability of protein-ligand complexes. Hydrogen bonding interactions are important for stabilizing protein-ligand complexes and impact ligand biological activity by reinforcing binding interactions. An increased number of hydrogen bonds typically signifies stronger and more stable interactions between the protein and the ligand. The plot illustrating the formation of hydrogen bonds for both the reference complex and the docked complexes is presented in Fig. 13. The mean H-bond values for CDK2-HJK, CDK2-M1, CDK2-M2, and CDK2-M3 are 2.980 ± 1.665 , 0.460 ± 0.640 , 0.194 ± 0.515 , and 0.211 ± 0.554 , respectively. Among all the docked complexes, CDK2-M1 shows a relatively higher mean value of H-bond, suggesting its ability to interact with the CDK2 enzyme. However, these docked complexes exhibited a moderate level of hydrogen bonding compared to the CDK2-HJK complex. Nonetheless, these docked complexes demonstrated stability based on RMSD, Rg analysis. These analyses indicate that while some of these complexes may exhibit fewer H-bonds, their stability and binding affinity can still be considerable because of beneficial structural dynamics and protein-ligand interactions.

3.13. SASA analysis

SASA analysis was carried out to understand the structural compactness of the hydrophobic core of the complexes. This measurement is essential when analyzing the interaction between protein and ligands and evaluating conformation and structural stability changes. The SASA plot for the reference complex and the docked complexes is illustrated in Fig. 14. The mean SASA values for CDK2-APO, CDK2-HJK, CDK2-M1, CDK2-M2, and CDK2-M3 are $167.382 \pm 1.666 \text{ nm}^2$, $167.428 \pm 1.465 \text{ nm}^2$, $168.202 \pm 1.367 \text{ nm}^2$, $166.640 \pm 1.411 \text{ nm}^2$, and $167.624 \pm 1.604 \text{ nm}^2$, respectively. Among all the docked complexes, CDK2-M2 exhibited the lowest SASA, lower than the reference complex and the apo enzyme. This suggests a more compact structure of M2 with reduced solvent accessibility, possibly reflecting a tighter binding conformation. In contrast, CDK2-M1 and CDK2-M3 complexes exhibited slightly higher SASA values compared to the apo enzyme, indicating relatively more solvent-exposed surfaces. These observations imply that M2 may enhance structural stability by promoting hydrophobic core burial, whereas M1 and M3 may lead to minor conformation expansion.

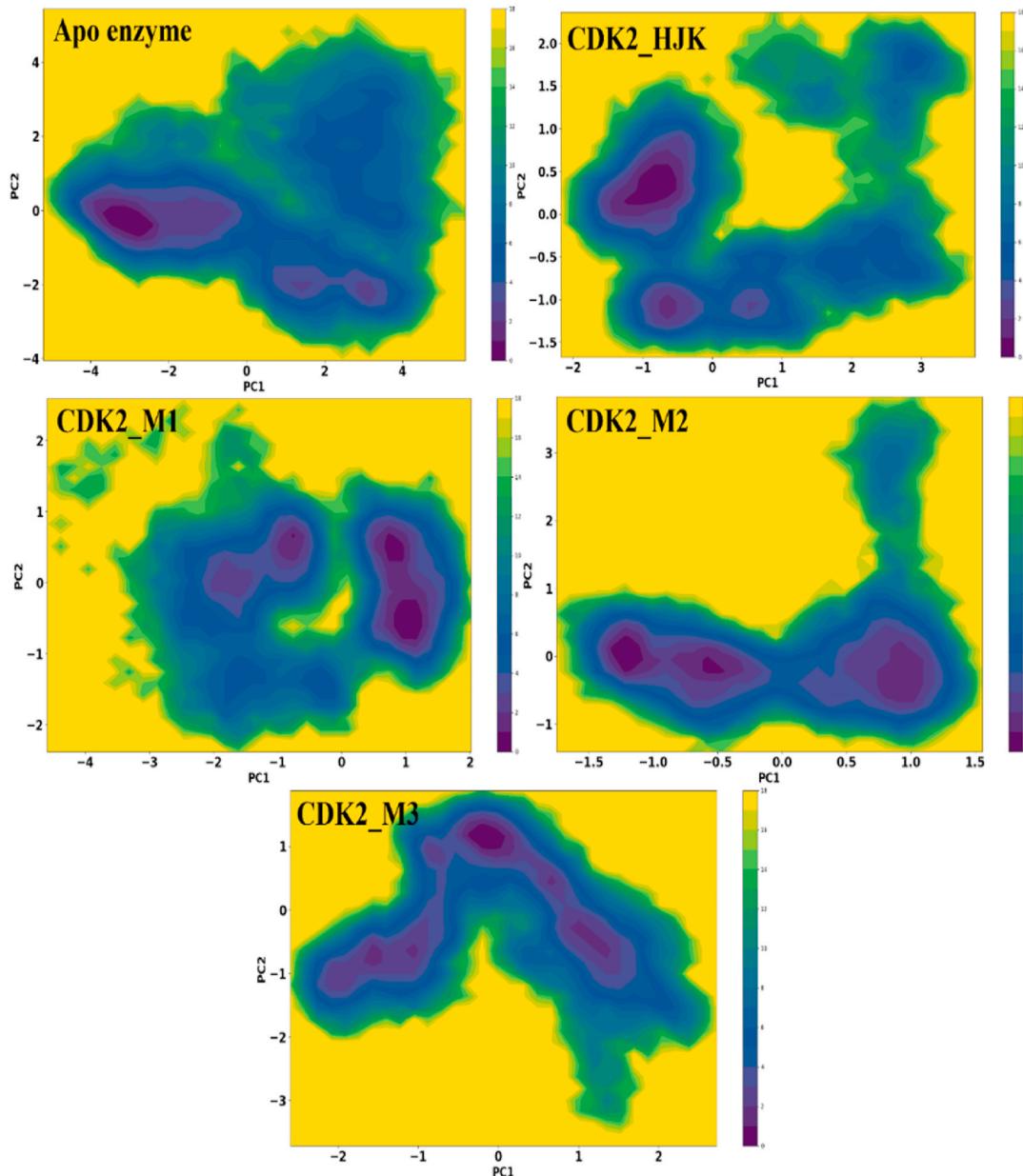


Fig. 16. The principal component analysis (PCA) based 2D free energy landscape (FEL) plot of apo-protein, reference-protein, and selected protein-ligand complexes with respect to the C α chain.

3.14. PCA analysis

PCA is commonly used in simulations to reveal significant collective movements in biomolecular systems. It identifies principal components that capture the largest variations in atomic positions, enabling the distinction of significant conformational changes from random fluctuations. PCA is particularly valuable for examining binding mechanisms, evaluating structural stability, and understanding functionally significant dynamics of protein-ligand complexes. The PCA analysis reveals distinct clustering patterns demonstrating variations in the stability and flexibility of complexes. Compounds within a smaller phase space display minimal conformational changes and more stability. On the other hand, more dispersed clusters suggest higher flexibility and less stability. This analysis projected the CDK2-M1, CDK2-M2, and CDK2-M3 trajectories onto the first two principal components, PC1 and PC2. These docked complexes were compared with the reference CDK2-HJK and CDK2-APO as depicted in Fig. 15. We observed that among all the

docked complexes and the reference complex, M2 occupied the least phase space. This indicates that the CDK2-M2 complex exhibits the highest stability among the complexes examined. The CDK2-M3 shows a similar phase space to reference CDK2-HJK and a similar scattering pattern. CDK2-M1 occupied a large phase space, indicating a relatively higher degree of flexibility.

3.15. FEL analysis

FEL analysis is used to identify the most stable conformational states by mapping the energy minima in the reduced dimensional space defined by the principal components PC1 and PC2. These energy minima indicate the energetically favorable state. This approach helps to find the global minima of CDK2-APO, CDK2-HJK, CDK2-M1, CDK2-M2, and CDK2-M3 complexes. Figs. 16 and 17 display the 2D and 3D FEL plots, highlighting the distinct energy minima and transition states that provide a deeper understanding of the conformational behavior and binding

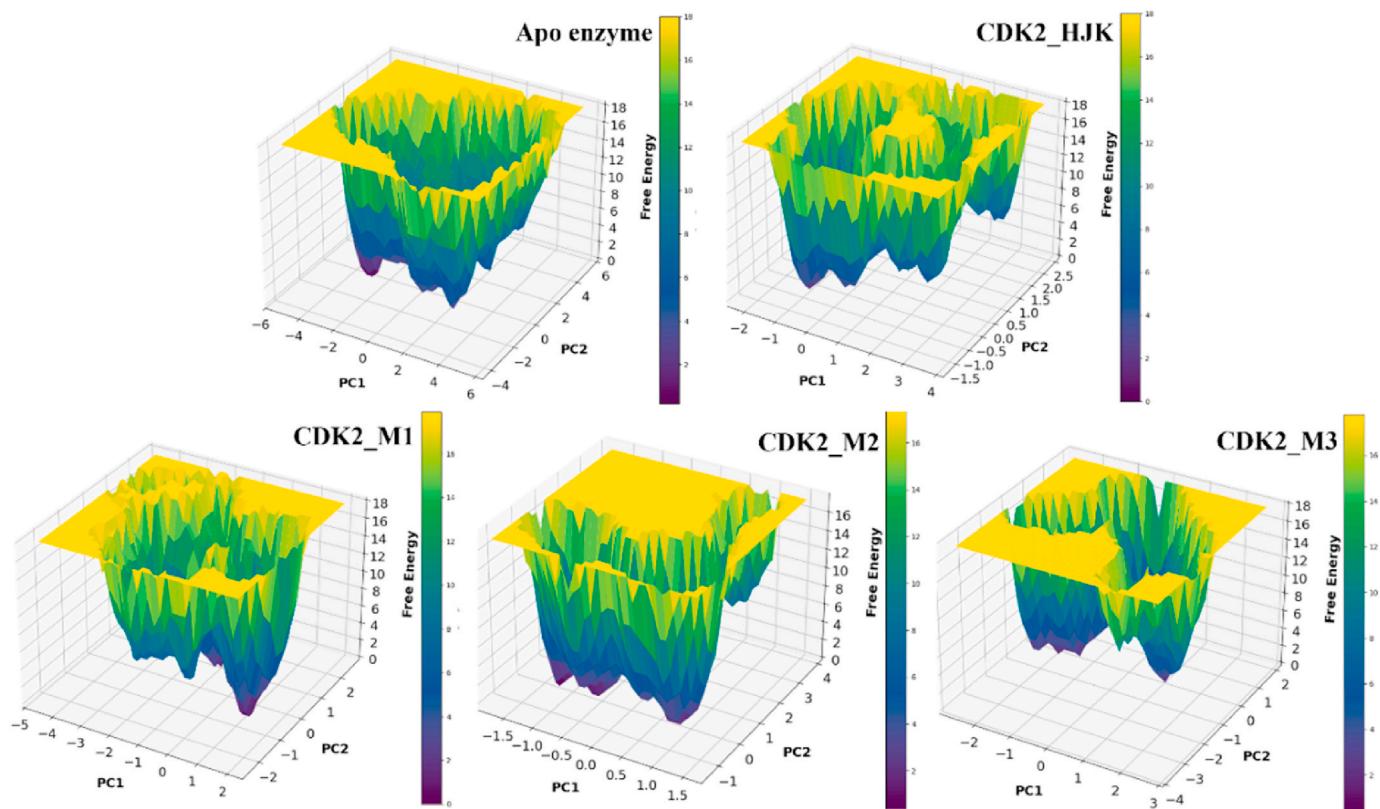


Fig. 17. The 3D-free energy landscape (FEL) plot based on principal component analysis of C α chain for apo-protein, reference-protein, and selected protein-ligand complexes.

energetics. In a 2D FEL pot, the dark blue color signifies low-energy regions or energy minima corresponding to stable, highly populated states. In contrast, lighter shades like yellow represent higher energy regions associated with less stable states.

In the 3D FEL plot, low-energy states are represented by blue color and deep curves, indicating the conformer's stability at certain PCs. As shown in Fig. 16, CDK2-HJK and CDK2-M3 form a similar phase space. Among all the systems, CDK2-M2 occupies the smallest phase space, which signifies the stable conformer compared to CDK2-APO and reference CDK2-HJK. These results were also justified by the 3D FEL plot as depicted in Fig. 17. In the 3D FEL plot, we observed that CDK2-M2 showed more compact phase space and lower energy minima. Similarly, the references CDK2-HJK and CDK2-M3 exhibit comparable phase space. These results demonstrate that among all the docked complexes, M2 is more stable than the reference HJK, while M3 exhibits stability equivalent to that of the reference HJK.

3.16. DCCM analysis

To understand the collective motions and communication between residues in the protein-ligand complexes, we conducted a DCCM analysis using the C α atom trajectories from simulations. DCCM evaluates the extent to which atomic displacements of residues are correlated over the simulation time. In the matrix, each element C_{ij} represents the degree of correlated motion between residues i and j , with values ranging from +1 (fully correlated motion) to -1 (fully anti-correlated motion). DCCM plots of the reference complex and other docked complexes are represented in Fig. 18. In this plot, red regions represent the positively correlated motions (i.e., residues move in the same direction). In contrast, the blue regions indicate negatively correlated or anti-correlated motions (i.e., residues move in opposite directions). Meanwhile, the white areas represent uncorrected or independent motions, and the diagonal line in each matrix signifies perfect self-correlation (C_{ii}

= 1). Among all the docked complexes, CDK2-M1 displays a correlation pattern most similar to the reference complex. M1 preserves key correlated motions and maintains the dynamic integrity. CDK2-M2 and CDK2-M3 exhibit a slightly more dispersed correlation pattern than CDK2-M1 and CDK2-HJK. Nevertheless, these complexes maintain crucial correlated regions that are essential for functional dynamics, suggesting that significant long-range residue interactions are predominantly conserved upon ligand binding.

3.17. Protein-ligand interaction energy

The analysis of interaction energy involves calculating non-bonded interaction energies, such as van der Waals and electrostatic interactions, between a protein and ligand throughout the simulation. The van der Waals interaction (ΔE_{vdw}) is calculated based on the Lennard-Jones potential, representing hydrophobic interaction and non-covalent attractions between protein and ligand. Electrostatic interactions follow Coulomb's rule, which describes ionic interactions, hydrogen bonding, and polar interactions. We have calculated Coulombic energy and Lennard-Jones energy for the reference CDK2-HJK, CDK2-M1, CDK2-M2, and CDK2-M3 as shown in Table 5. Among all the docked complexes, CDK2-M2 exhibits the highest total interaction energy of -146.43 ± 05.40 kJ/mol and the highest Lennard-Jones energy of -121.17 ± 02.50 kJ/mol. The Lennard-Jones energy of CDK2-M2 is nearly identical to that of the reference CDK2-HJK, which has an energy of -134.83 ± 01.40 kJ/mol. However, the Coulombic energy of the reference is higher owing to the presence of charge. The Lennard-Jones energy of CDK2-M3 is relatively good (-106.94 ± 10.00 kJ/mol), but its Coulombic energy is comparatively low (-19.91 ± 02.50 kJ/mol). CDK2-M1 exhibits the highest coulombic energy (-36.25 ± 00.00 kJ/mol) among all docked complexes; however, its Lennard-Jones energy is relatively low (-100.94 ± 00.00 kJ/mol). Overall, the findings indicate that M2 exhibits more favorable interactions with the highest total

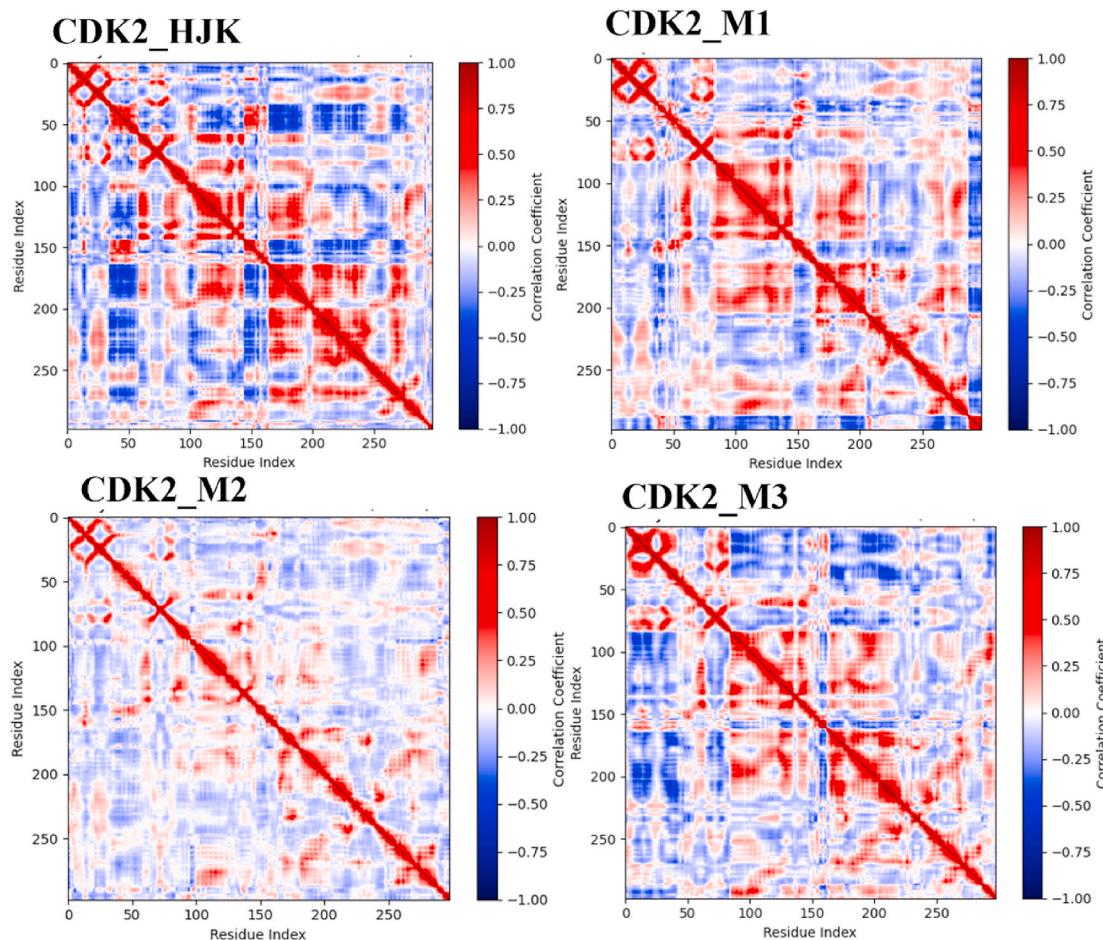


Fig. 18. DCCM analysis plots for the CDK2 protein in complex with the reference and selected docked complexes.

Table 5

Protein-ligand interaction energies (average short-range Coulomb and short-range Lennard-Jones energies) from molecular dynamics simulations of 500 ns for CDK2-HJK, CDK2-M1, CDK2-M2, and CDK2-M3.

Complexes	Coulombic Energy (kJ/mol)	Lennard-Jones Energy (kJ/mol)	Total Interaction Energy (kJ/mol)
CDK2-HJK	-206.94 ± 02.10	-134.93 ± 01.40	-341.87 ± 03.50
CDK2-M1	-36.25 ± 00.00	-100.94 ± 0.00	-137.19 ± 00.00
CDK2-M2	-25.26 ± 02.90	-121.17 ± 02.50	-146.43 ± 05.40
CDK2-M3	-19.91 ± 02.50	-106.94 ± 10.00	-126.85 ± 02.50

interaction energy than the other shortlisted molecules.

3.18. Binding free energy calculations

MM/PBSA analysis of the docked complexes was carried out on the stable trajectory extracted from 300 to 500 ns using the gmx_MMPBSA tool. *Supplementary Table 9* presents the total binding energies and the energy values corresponding to each energy component contribution for the docked complexes. The total binding energies for CDK2-M1, CDK2-M2, and CDK2-M3 are -7.65 kcal/mol, -9.39 kcal/mol, and -13.88 kcal/mol, respectively. CDK2-M3 exhibited the highest binding energy of all the docked complexes, with CDK2-M2 following closely, demonstrating relatively strong binding energy. The highest value of M3 is mainly because of the strong van der Waals interaction (-21.05 kcal/mol), which suggests effective hydrophobic interactions with the binding site. However, the electrostatic interactions (-2.91 kcal/mol) are relatively weak, and the lower polar solvation energy (12.20 kcal/mol)

contributes to the favorable binding. CDK2-M2 shows moderate binding energy, primarily influenced by van der Waals interactions (-20.44 kcal/mol) and significant contributions from electrostatic interactions (-6.06 kcal/mol). However, the high polar solvation penalty (18.95 kcal/mol) diminishes the overall stability of the complex. M1 exhibits the least favorable binding free energy value due to lower van der Waals interaction (-11.43 kcal/mol) and electrostatic energy (-2.99 kcal/mol), alongside a high polar solvation energy (7.95 kcal/mol). The higher binding energy of M3 and M2 complexes suggests strong and stable interactions between the protein and the ligand molecules.

3.19. pCDK2i_v1.0: an advanced AI-powered cheminformatics tool to predict/screen CDK2 inhibitors

In the current work, with an aim to support the work of experimentalists, a machine learning model was developed for predicting the potential of a molecule to be CDK2 active. An innovative online cheminformatics tool, named pCDK2i_v1.0, accessible via Google Colab, is designed to help researchers efficiently screen potential CDK2 inhibitors. This is openly available through https://github.com/Amincheminform/pCDK2i_v1. This tool leverages a pre-trained, rigorously validated AI/ML model to distinguish between active and inactive compounds with high reliability.

Key features of pCDK2i_v1.0:

- (1) Fingerprint-Based Molecular Analysis: The tool utilizes molecular fingerprints computed directly from the input SMILES (Simplified Molecular Input Line Entry System) string, a widely used text-based representation of chemical structures.

- (2) Extensive model validation: The model has undergone comprehensive validation using high-quality, curated datasets, ensuring robust performance in identifying active (Class 1: likely to exhibit CDK2 inhibitory activity) and inactive (Class 0: unlikely to inhibit CDK2) CDK2 inhibitors.
- (3) Clear visualization: This tool generates a visual representation of the 2D structure of the query molecule, allowing researchers to inspect and analyze the compound directly within the Colab notebook.

This AI-powered cheminformatics tool significantly lowers the barrier for researchers in drug discovery and CDK2-related research, offering a freely accessible and user-friendly tool to accelerate the identification of promising anti-cancer candidates. By integrating cheminformatics with AI-driven predictions, pCDK2i_v1.0 enhances the efficiency of virtual screening, enabling scientists to explore new avenues in drug development and precision medicine. To validate our tool, we have performed the screening of the reported 265 CDK2 inhibitors retrieved from the RCSB protein databank. We observed that out of 265 inhibitors, our model was able to predict 174 molecules as active (Supplementary Table 10).

4. Conclusions

CDK2 is a promising target for cancer treatment, and extensive research is underway to discover its inhibitors. Nevertheless, no drug has received FDA approval for CDK2 to date. All three molecules shortlisted in the current work possess the capability of inhibiting the CDK2 enzyme. With respect to this, we utilized machine learning techniques to perform ligand-based virtual screening to identify inhibitors for the CDK2 enzyme. In this work, we selected an RF model to perform virtual screening of 4,77,975 molecules from the COCONUT database. The screened molecules were subsequently utilized for molecular docking studies and interaction analysis. The top 40 molecules with good docking scores were selected for ADMET analysis, shortlisting three molecules that adhered to all ADME/T criteria. The 2D interaction analysis reveals that all three docked molecules demonstrate conserved interactions with the residues Lys33 and Asp145, which are crucial for enzyme inhibition. Therefore, all three molecules possess the capability of inhibiting the CDK2 enzyme. We also observed that Molecule 2 features an extended fused heterocyclic system, which may enhance its inhibitory potential. Subsequently, we conducted a DFT analysis, which indicated that Molecule 2 was the most reactive among the selected molecules. The stability of these three chosen complexes was assessed via simulation studies with a time duration of 500 ns. The MD analysis results, including RMSD, RMSF, H-bond, Rg, SASA, PCA, FEL, DCCM, and protein-ligand interactions, indicate that all three molecules exhibit stable behavior with the CDK2 enzyme. The protein-RMSD analysis shows that CDK2-M3 has greater stability, reflected by its lower RMSD value when compared to the apo CDK2 enzyme. Meanwhile, ligand RMSD indicates M1's stability, reflecting a stable trajectory trend comparable to the reference. Protein RMSF and Ligand RMSF analysis reveal that M2 exhibits the lowest fluctuation and maintains stability. Furthermore, CDK2-M3 showed a lower Rg value than apo and reference, indicating a more compact structure. CDK2-M1 shows a comparatively higher average value of H-bond than other docked molecules. Furthermore, M2 demonstrated stable behaviour in SASA, PCA, FEL analyses, and protein-ligand interaction evaluations. Subsequently, MM/PBSA analysis revealed that M3 exhibited the lowest binding energy, with M2 following closely behind. In summary, the findings from the different computational methods employed in this study indicate that all three molecules, M1, M2, and M3, may serve as inhibitors of the CDK2 enzyme. Moreover, in the current contribution, an online tool "pCDK2i_v1.0" was designed to correlate the activity of a molecule as CDK2 active or inactive. This tool plays an important role in advancing the research on breast cancer involving the CDK2 enzyme. It will help

experimentalists to use the machine learning model to predict the activity of the novel molecules.

CRediT authorship contribution statement

Priyanka Solanki: Writing – original draft, Visualization, Validation, Formal analysis, Data curation. **Sk. Abdul Amin:** Writing – review & editing, Software. **Anu Manhas:** Writing – review & editing, Supervision, Resources, Methodology, Conceptualization.

Ethics statement

We confirm that this study does not involve human subjects, animals, or other components necessitating ethical approval under current institutional or legal guidelines.

Declaration of competing interest

The authors declare no conflict of interest.

Acknowledgments

PS acknowledges the Pandit Deendayal Energy University (PDEU) for the financial assistance. PS and AM acknowledge PDEU for the infrastructure and facilities.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.combiomed.2025.110712>.

References

- [1] C.P. Wild, E. Weiderpass, B.W. Stewart (Eds.), World Cancer Report: Cancer Research for Cancer Prevention, International Agency for Research on Cancer, Lyon, France, 2020. Available from: <http://publications.iarc.fr/586>. Licence: CC BY-NC-ND 3.0 IGO.
- [2] M. Athar, A. Manhas, N. Rana, I. Ahmad, Computational and bioinformatics tools for understanding disease mechanisms, *Biozell* 48 (2024) 935.
- [3] P. Solanki, S. Sarwadiya, M. Athar, P.C. Jha, A. Manhas, Targeting the cyclin-dependent kinase family in anticancer drug discovery: from computational to experimental studies, *Chem. Phys. Impact* (2024) 100768.
- [4] A. Lashen, S. Alqahtani, A. Shoaqifi, M. Algethami, J.N. Jeyapalan, N.P. Mongan, E. A. Rakha, S. Madhusudan, Clinicopathological significance of cyclin-dependent kinase 2 (CDK2) in ductal carcinoma *in situ* and early-stage invasive breast cancers, *Int. J. Mol. Sci.* 25 (2024) 5053.
- [5] V.E. Brown, S.L. Moore, M. Chen, N. House, P. Ramsden, H.J. Wu, S. Ribich, A.R. Grassian, Y.J. Choi, CDK2 regulates collapsed replication fork repair in CCNE1-amplified ovarian cancer cells via homologous recombination, *NAR Cancer* 5 (2023) zcad039.
- [6] A. Lavecchia, Deep learning in drug discovery: opportunities, challenges and future prospects, *Drug Discov. Today* 24 (2019) 2017–2032.
- [7] F. Zhong, J. Xing, X. Li, X. Liu, Z. Fu, Z. Xiong, D. Lu, X. Wu, J. Zhao, X. Tan, Artificial intelligence in drug design, *Sci. China Life Sci.* 61 (2018) 1191–1204.
- [8] P. Solanki, N. Rana, P. Jha, A. Manhas, A comprehensive analysis of the role of molecular docking in the development of anticancer agents against the cell cycle CDK enzyme, *Biozell* 47 (2023) 707.
- [9] D. Sydow, A. Morger, M. Driller, A. Volkamer, TeachOpenCADD: a teaching platform for computer-aided drug design using open source packages and data, *J. Cheminf.* 11 (2019) 1–7.
- [10] D. Mendez, A. Gaulton, A.P. Bento, J. Chambers, M. De Veij, E. Félix, M. P. Magarinos, J.F. Mosquera, P. Mutowo, M. Nowotka, ChEMBL: towards direct deposition of bioassay data, *Nucleic Acids Res.* 47 (2019) D930–D940.
- [11] C.A. Lipinski, F. Lombardo, B.W. Dominy, P.J. Feeney, Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings, *Adv. Drug Deliv. Rev.* 23 (1997) 3–25.
- [12] G.A. Landrum, Rdkit: Open-Source Cheminformatics, 2014. Release 2014.03.1.
- [13] B. Merget, S. Turk, S. Eid, F. Rippmann, S. Fulle, Profiling prediction of kinase inhibitors: toward the virtual assay, *J. Med. Chem.* 60 (2017) 474–485.
- [14] C. Shen, J. Luo, W. Ouyang, P. Ding, X. Chen, IDDKin network-based influence deep diffusion model for enhancing prediction of kinase inhibitors, *Bioinformatics* 36 (2020) 5481–5491.
- [15] M. Sorokina, P. Merselberger, K. Rajan, M.A. Yirik, C. Steinbeck, COCONUT online: collection of open natural products database, *J. Cheminf.* 13 (2021) 2.

- [16] A.P. Bento, A. Hersey, E. Félix, G. Landrum, A. Gaulton, F. Atkinson, L.J. Bellis, M. De Veij, A.R. Leach, An open source chemical structure curation pipeline using RDKit, *J. Cheminf.* 12 (2020) 1–16.
- [17] J.B. Baell, G.A. Holloway, New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays, *J. Med. Chem.* 53 (2010) 2719–2740.
- [18] R. Brenk, A. Schipani, D. James, A. Krasowski, I.H. Gilbert, J. Frearson, P.G. Wyatt, Lessons learnt from assembling screening libraries for drug discovery for neglected diseases, *ChemMedChem: Chemistry Enabling Drug Discovery* 3 (2008) 435–444.
- [19] E. Gilberg, D. Stumpfe, J. Bajorath, Activity profiles of analog series containing pan assay interference compounds, *RSC Adv.* 7 (2017) 35638–35647.
- [20] Y. Li, J. Zhang, W. Gao, L. Zhang, Y. Pan, S. Zhang, Y. Wang, Insights on structural characteristics and ligand binding mechanisms of CDK2, *Int. J. Mol. Sci.* 16 (2015) 9314–9340.
- [21] T.G. Davies, D.J. Pratt, J.A. Endicott, L.N. Johnson, M.E. Noble, Structure-based design of cyclin-dependent kinase inhibitors, *Pharmacol. Ther.* 93 (2002) 125–133.
- [22] I. Hope, J.A. Endicott, J.E. Watt, Emerging approaches to cdk inhibitor development, a structural perspective, *RSC Chem. Biol.* 4 (2023) 146–164.
- [23] M. Rarey, B. Kramer, T. Lengauer, G. Klebe, A fast flexible docking method using an incremental construction algorithm, *J. Mol. Biol.* 261 (1996) 470–489.
- [24] BioSolveIT, SesarR v13.1.1 Software, Sankt Augustin, Germany, SeeSAR, accessed on 25 July, 2023.
- [25] H.-J. Böhm, The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure, *J. Comput. Aided Mol. Des.* 8 (1994) 243–256.
- [26] N.M. O'Boyle, M. Banck, C.A. James, C. Morley, T. Vandermeersch, G. R. Hutchison, Open babel: an open chemical toolbox, *J. Cheminf.* 3 (2011) 1–14.
- [27] A. Daina, O. Michelin, V. Zoete, SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules, *Sci. Rep.* 7 (2017) 42717.
- [28] D.E. Pires, T.L. Blundell, D.B. Ascher, pkCSM: predicting small-molecule pharmacokinetic and toxicity properties using graph-based signatures, *J. Med. Chem.* 58 (2015) 4066–4072.
- [29] A. Frisch, Gaussian 09W Reference, vol. 25p, 2009, p. 470. Wallingford, USA.
- [30] I.Y. Zhang, J. Wu, X. Xu, Extending the reliability and applicability of B3LYP, *Chem. Commun.* 46 (2010) 3057–3070.
- [31] C.E. Check, T.O. Faust, J.M. Bailey, B.J. Wright, T.M. Gilbert, L.S. Sunderlin, Addition of polarization and diffuse functions to the LANL2DZ basis set for p-block elements, *J. Phys. Chem.* 105 (2001) 8111–8116.
- [32] R. Vijayaraj, V. Subramanian, P. Chatteraj, Comparison of global reactivity descriptors calculated using various density functionals: a QSAR perspective, *J. Chem. Theor. Comput.* 5 (2009) 2744–2753.
- [33] M.J. Abraham, T. Murtola, R. Schulz, S. Páll, J.C. Smith, B. Hess, E. Lindahl, GROMACS: high performance molecular simulations through multi-level parallelism from laptops to supercomputers, *SoftwareX* 1 (2015) 19–25.
- [34] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J.L. Klepeis, R.O. Dror, D. E. Shaw, Improved side-chain torsion potentials for the amber ff99SB protein force field, *Proteins: Struct., Funct., Bioinf.* 78 (2010) 1950–1958.
- [35] A.W. Sousa da Silva, W.F. Vranken, ACPYPE-antechamber python parser interface, *BMC Res. Notes* 5 (2012) 1–8.
- [36] A. Jakalian, B.L. Bush, D.B. Jack, C.I. Bayly, Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method, *J. Comput. Chem.* 21 (2000) 132–146.
- [37] H.J. Berendsen, J.v. Postma, W.F. Van Gunsteren, A. DiNola, J.R. Haak, Molecular dynamics with coupling to an external bath, *J. Chem. Phys.* 81 (1984) 3684–3690.
- [38] M. Parrinello, A. Rahman, Crystal structure and pair potentials: a molecular-dynamics study, *Phys. Rev. Lett.* 45 (1980) 1196.
- [39] T. Darden, D. York, L. Pedersen, Particle mesh ewald: an N·log (N) method for ewald sums in large systems, *J. Chem. Phys.* 98 (1993) 10089–10092.
- [40] B. Hess, H. Bekker, H.J. Berendsen, J.G. Fraaije, LINCS: a linear constraint solver for molecular simulations, *J. Comput. Chem.* 18 (1997) 1463–1472.
- [41] B. Biçak, S.K. Gündüz, Y. Kökcü, A.E. Öznel, S. Akyüz, Molecular docking and molecular dynamics studies of L-glycyl-L-glutamic acid dipeptide, *Bilge Int. J. Sci. Technol. Res.* 3 (2019) 1–9.
- [42] M.S. Valdés-Tresanco, M.E. Valdés-Tresanco, P.A. Valiente, E. Moreno, gmx_MMPBSA: a new tool to perform end-state free energy calculations with GROMACS, *J. Chem. Theor. Comput.* 17 (2021) 6281–6291.
- [43] D.J. Wood, J.D. Lopez-Fernandez, L.E. Knight, I. Al-Khawaldeh, C. Gai, S. Lin, M. P. Martin, D.C. Miller, C. Cano, J.A. Endicott, FragLites—minimal, halogenated fragments displaying pharmacophore doublets. An efficient approach to druggability assessment and hit generation, *J. Med. Chem.* 62 (2019) 3741–3752.
- [44] J. Chen, L. Pang, W. Wang, L. Wang, J.Z. Zhang, T. Zhu, Decoding molecular mechanism of inhibitor bindings to CDK2 using molecular dynamics simulations and binding free energy calculations, *J. Biomol. Struct. Dyn.* 38 (2020) 985–996.
- [45] J. Chen, X. Wang, J.Z. Zhang, T. Zhu, Effect of substituents in different positions of aminothiazole hinge-binding scaffolds on inhibitor–CDK2 association probed by interaction entropy method, *ACS Omega* 3 (2018) 18052–18064.
- [46] P. Solanki, A. Manhas, D. Patel, P. Jha, Structure-based computational approach in search of the potent molecules targeting phosphoglycerate dehydrogenase (PHGDH) enzyme for cancer treatment, *ChemistrySelect* 9 (2024) e202403205.
- [47] N. Rana, P. Solanki, R. Mehra, A. Manhas, Identification of natural compound inhibitors for substrate-binding site of MTHFD2 enzyme: insights from structure-based drug design and biomolecular simulations, *Chem. Phys. Impact* 10 (2025) 100809.
- [48] A. Manhas, S.P. Kumar, P.C. Jha, Molecular modeling of Plasmodium falciparum peptide deformylase and structure-based pharmacophore screening for inhibitors, *RSC Adv.* 6 (2016) 29466–29485.