Computers in Biology and Medicine 195 (2025) 110672

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

De novo design and bioactivity prediction of mitotic kinesin Eg5 inhibitors using MPNN and LSTM-based transfer learning

Damilola Samuel Bodun [a,b,c,*], Isiaka O. Ibrahim [d], Mujeebat Bashiru [c,e], Rachael Oluwakamiye Abolade [c,f,g], Wilberforce K. Ndarawit [c,h,i], Adedoyin John-Joy Owolade [c,j], Adegbenro Temitope [c,k], Ibeh John Somtochukwu [c,k], Ezekiel Abiola Olugbogi [l], Katukoliya Gamage Anuththara Samadhi [a], Gbolahan Oduselu [b], Suliman Sharif [a], Damilola Alex Omoboyowa [m], Olayinka O. Ajani [b]

[a] Standard Seed Corporation, Smyrna, USA
[b] Covenant University Bioinformatics Research (CUBRe), Covenant University, Ota, Nigeria
[c] ChemoInformatics Academy, Lagos, Nigeria
[d] Department of Engineering Technology, Arkansas State University, Jonesboro, USA
[e] Department of Chemistry, University of Arkansas at Little Rock, Little Rock, USA
[f] Department of Information Science, University of Arkansas at Little Rock, Little Rock, USA
[g] Department of Pharmaceutical Sciences, University of Arkansas for Medical Sciences, Little Rock, USA
[h] Department of Physical Sciences, University of Embu, Embu, Kenya
[i] Natural Product Chemistry and Computational Drug Discovery Laboratory, Embu, Kenya
[j] Department of Pharmaceutical Chemistry, Faculty of Pharmacy, Obafemi Awolowo University, Ile Ife, Nigeria
[k] Faculty of Pharmacy, University of Lagos, Lagos, Nigeria
[l] Department of Biochemistry, School of Basic Medical Sciences, Babcock University, Ilishan-Remo, Nigeria
[m] Phytomedicine and Computational Biology Lab. Department of Biochemistry, Adekunle Ajasin University, Akungba-Akoko, Nigeria

ARTICLE INFO

Keywords:
Long short term memory (LSTM)
Mitotic kinesin Eg5
Molecular dynamics simulation
Message passing neural network
Transfer learning

ABSTRACT

Breast cancer, the most commonly diagnosed disease worldwide, has been linked to the overexpression of the kinesin Eg5 protein, a spindle motor protein crucial for the assembly and maintenance of the bipolar spindle during mitosis. This makes Eg5 an attractive therapeutic target for tumor treatment. To address the urgent need for effective treatments for this life-threatening illness, we utilized generative AI to design novel and potential inhibitors of this protein.

In this study, a generative LSTM model was pretrained on SMILES data from ChEMBL and subsequently fine-tuned using SMILES of compounds with reported activity against the Eg5 protein. The fine-tuned model generated valid compounds, which were screened using a machine learning model, drug-likeness filters, molecular docking, and molecular dynamics (MD) simulations conducted over 200 ns.

Five novel compounds with better binding affinities to Eg5 compared to the co-crystallized ligand were identified. The top compound, Compound 103 (a bioisostere of the co-crystallized ligand), demonstrated a significantly improved binding free energy (−82.68 kcal/mol) compared to the co-crystallized ligand (−76.98 kcal/mol), as determined by MM-GBSA calculations. ADMET predictions and MD simulations further confirmed that the top compounds interacted effectively with the target protein and exhibited drug-like properties.

This study shows the potential of generative AI to explore our vast chemical space and find promising drug candidates. However, further *in vitro* and *in vivo* studies are needed to confirm the predicted biological effects of the top compounds.

* Corresponding author
E-mail address: dbodun56@gmail.com (D.S. Bodun).

https://doi.org/10.1016/j.compbiomed.2025.110672
Received 31 March 2025; Received in revised form 17 May 2025; Accepted 21 June 2025
Available online 27 June 2025
0010-4825/© 2025 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

1. Introduction

Breast cancer still ranks among the most common and lethal cancer type globally with over 2.3 million cases reported annually and nearly 685,000 mortalities [1]. Despite significant advancements in both detecting the disease in its early stages and therapeutic interventions, the prognosis for patients diagnosed with aggressive and therapeutic-resistant form of breast cancer remains a problem of medical practice globally, hence necessitating the need to develop a novel and effective therapeutic agent. The kinesin spindle protein Eg5, a molecular motor essential for the proper segregation of chromosomes during mitosis, has recently received much attention as a potential anti-cancer drug candidate [2]. Inhibiting Eg5 disrupts mitotic spindle formation, resulting to mitotic arrest and subsequent apoptosis of cancer cells, especially in the rapidly proliferating tumors such as breast cancer [2]. Targeting Eg5 with bioactive molecules holds promise as viable proposition and an effective therapeutic strategy against breast cancer. However, discovering potent, selective and bioactive inhibitor for Eg5 remains an intricate and resource intensive endeavor.

Traditional drug discovery often involves high-throughput screening of chemical libraries to identify potential molecules that can act as drug candidates. However, this process is often costly, time-consuming, and limited by the chemical diversity of the molecule's libraries available [3,4]. Machine learning, particularly deep learning, offers an innovative alternative approach by generating novel chemical structures, exploring a wider chemical space and predicting their bioactivity in-silico, thereby reducing the dependency on extensive wet lab experiments. The emergence of long short-term memory (LSTM), recurrent neural networks (RNNs) and transfer learning techniques has enabled rapid generation and evaluation of drug-like molecules, potentially expediting the discovery of valuable therapeutic drug candidates. LSTM recurrent neural networks have the capacity to process and generate sequential data, which makes them well-suited for handling molecular representation like simplified molecular input line entry system (SMILES) strings [5]. SMILES notations encode molecules as sequences of characters representing atoms, bonds and molecular structures, hence enabling RNNs with a structured language to learn and manipulate [6]. Therefore, by training the model on large datasets of existing drug-like molecules, generative AI models can learn the "grammar" and rules governing the chemical composition, enabling them to generate novel SMILES strings that could possibly represent plausible molecular structure.

In this study, an LSTM-based transfer learning approach was employed to leverage knowledge from ChEMBL dataset (https://www.ebi.ac.uk/chembl/), a comprehensive bioactivity database curated and maintained by the European Bioinformatic Institute (EBI) [7]. The database is designed to keep bioactivity data on small molecules with drug-like features, focusing on the relationship between chemical compounds, molecular targets, and the biological system they interact with [7]. By initially training the LSTM model with ChEMBL database, the model was able to learn the correct structure of SMILES sequences then further training the model using compounds with activity toward Mitotic Kinesin Eg5, made it possible for the model to generate compounds specific for our target protein, a process known as transfer learning or fine tuning. These generated molecules were further screened based on their physicochemical properties, and their binding affinities, employing advanced techniques and molecular dynamic simulations.

2. Materials and method

2.1. Data collection and preprocessing

1,576,904 compounds were collected from the ChEMBL database, a manually curated database of bioactive molecules with drug-like properties. This data was imported and preprocessed using the Pandas and RDKit libraries in Python. First, salt components were removed from the compounds using the 'SaltRemover' function of RDKit. Then, the molecules were standardized using the RDKit MolStandardize module. Finally, very large and very small molecules were filtered out by selecting only those with SMILES lengths between 34 and 74 characters.

2.2. Long short term memory (LSTM)

A Recurrent Neural Network (RNN) is a type of feedforward neural network designed to process sequential data, such as text and SMILES [8]. It contains feedback loops in its recurrent layers, allowing it to maintain information in "memory" over time. However, RNNs face a significant issue known as the vanishing gradient problem, where the gradient of the loss function decays exponentially with time. This causes the loss to become so minimal that it has little to no effect on the network's learning. Long Short-Term Memory (LSTM) networks, an improved type of RNN, introduce a memory cell along with input and forget gates, which help to address the gradient problems and enable the network to learn longer-term dependencies [9] (Fig. 1). We trained an LSTM model using the preprocessed SMILES data from ChEMBL.

2.3. Training the LSTM model

In this study, we trained a character-level LSTM. To begin, we created a dictionary to map characters to integers, which was then used to encode the preprocessed SMILES strings into numbers, as neural networks only understand numerical input. Our neural network consisted of an embedding layer, followed by a single-layer LSTM network, and a fully connected output layer. The embedding layer mapped input sequences of size 31 to an embedding space with 256 dimensions. The LSTM layer had 512 hidden units, and the network had approximately 2 million trainable parameters. Input sequences were processed in batches of 64. After the LSTM layer, a fully connected (dense) layer with 31 output features was used, and the output was transformed into probabilities using a Softmax layer during sampling. The network was trained using backpropagation through time, with the ADAM optimizer using a learning rate of 0.001 and cross-entropy loss function. The PyTorch library was used to implement the model. During training, the model's accuracy was tested on the validation dataset for 5000 epochs. This model "the prior model" was then sampled to generate 1,000,000 characters.

2.4. FineTuning the model

After the initial training, we confirmed that the "prior model" could generate valid SMILES (i.e., compounds). However, to generate molecules specific to a particular target, it was necessary to update the "prior model" weights through retraining.

To achieve this, the FASTA sequence of the target protein, Mitotic Kinesin Eg5 (PDB ID - 3KEN), was blasted against the ChEMBL database to obtain a list of compounds known to be active against Eg5. The data

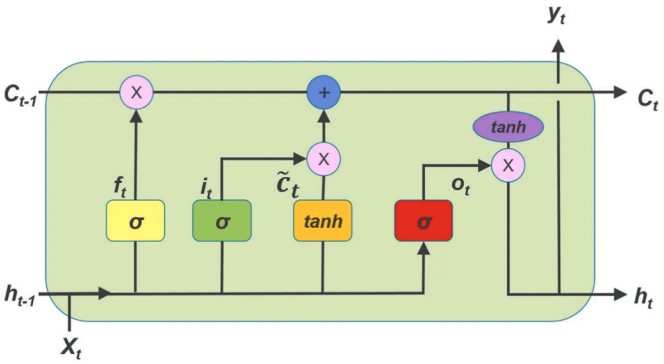


Fig. 1. An LSTM cell.

was then cleaned, duplicates were removed, and it was preprocessed as before. This processed data was used to update the weights of the "prior model" by retraining it for 2000 epochs using a reduced learning rate of 0.0005 to help prevent overfitting. This smaller step size helps prevent drastic updates to the pretrained weights and facilitates smoother adaptation to the Eg5 data. Afterward, the model was sampled to generate 1,000,000 characters. Both the smiles generated by the "prior model" and the "fine-tuned model" were evaluated for validity, novelty, and uniqueness after sampling [10].

Validity evaluates whether the generated SMILES strings are correct, as determined by RDKit's MolFromSmiles function.

Novelty assesses whether the generated molecules are new by checking if the SMILES strings are absent from the training dataset.

Uniqueness ensures that the novel molecules are distinct by checking for duplicates among them.

The unique SMILES (n = 3906) were then saved in a.CSV file for further study.

To evaluate the performance of the generative models, the Hamiltonian diversity value was measured [11].

2.5. Chemical space analysis

We generated Morgan fingerprints for the compounds produced by the prior model and the fine-tuned model, using a radius of 2 and 2048-bit fingerprint vectors. The RDKit molecular representations were obtained using RDKit in Python. Subsequently, we performed t-Distributed Stochastic Neighbor Embedding (t-SNE) clustering analysis using scikit-learn in Python, reducing the data points from 2048 dimensions to 2 dimensions [12]. Default parameters provided by scikit-learn were used for the t-SNE analysis [13]. The chemical space was visualized through histograms and scatter plots created using the Plotly library.

2.6. Virtual screening by machine learning model

Using ChemProp, a directed message-passing neural network model that learns to predict molecular properties directly from the graph structure of molecules—where atoms are represented as nodes and bonds as edges—we created a classification model based on the list of Eg5 inhibitors from ChEMBL [14]. Compounds with pIC50 values less than 6.5 were classified as 0, while those with values higher than 6.5 were classified as 1. This dataset was divided into training, test, and validation sets in an 8:1:1 ratio. The model was then used to predict the activity of the unique SMILES generated from the fine-tuned model. The molecules (n = 1328) predicted to be active (value = 1) were converted to a.CSV file. This model was also evaluated with an external validation list of SMILES and classification metrics, accuracy, precision, recall, F1 score and Specificity.

$$\text{Accuracy} = \frac{\textit{Correct classifications}}{\textit{Total classifications}} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Recall} = \frac{\textit{Correctly classified actual positives}}{\textit{all actual positives}} = \frac{TP}{TP + FN}$$

$$\text{Precision} = \frac{\textit{Correctly classified actual positives}}{\textit{everything classified as positives}} = \frac{TP}{TP + FP}$$

$$\text{Specificity} = \frac{\textit{Correctly classified actual negatives}}{\textit{everything classified as negatives}} = \frac{TN}{TN + FP}$$

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

2.7. Virtual screening by PAINs

PAINs, or Pan Assay Interference Compounds, are small molecules that can lead to false positives in drug discovery tests [15]. They interfere with various biological assays and often produce non-specific or misleading activity that appears to be genuine pharmacological effects. In this study, we filtered the compounds predicted to be active by using an external list of PAINs to reduce the list of compounds to 963 [16].

2.8. Structure based screening

The target protein, Human Eg5 in complex with S-trityl-L-cysteine (STLC) (PDB ID: 3KEN), was downloaded in PDB format from the Protein Data Bank (PDB) and imported into Maestro software for protein preparation. During this process, water molecules were removed, and the protein was minimized and optimized. The active site of the protein was then mapped based on the position of the co-crystallized ligand (STLC) using the grid generation module [17]. The list of active compounds (n = 963) identified from the virtual screening using PAINS was converted into.SDF format using DataWarrior software and subsequently imported into Maestro, where it was prepared using the LigPrep module. Molecular docking was carried out in a stepwise approach. First, Standard Precision (SP) docking was used to quickly screen the large number of compounds (n = 963) and identify those with favorable binding scores. Compounds that scored better than −8.0 kcal/mol in SP docking were then redocked using Extra Precision (XP) mode, which uses more advanced scoring functions and thorough sampling to improve accuracy. This docking methodology was validated by calculating the RMSD (Root Mean Square Deviation) of the co-crystallized ligand (STLC) between its initial position and the re-docked position. Additionally, the correlation between docking scores and the experimental values (pIC50) of the compounds was assessed. The co-crystallized ligand (STLC) was used as the reference ligand for this study. Later in this study, this docking protocol was later validated by superposing the co-crystallized redocked to its native position and also by correlating the experimental values of sampled molecules reported to have an activity towards Eg5 and the XP docking score (see Supplementary Fig. 2 and Supplementary Fig. 3).

2.9. Estimation of binding energy

The XP docking results showed a correlation between good poses and good scores for the docked complexes. However, to assess the potential biological response of these complexes, the free binding energy was calculated using molecular mechanics generalized born surface area (MM-GBSA) with the Prime module of Maestro. The top five ranked compounds including the standard drug were selected for this analysis. The results of this analysis would give information on the stability of the complex and help to determine if the association between the ligands and the target protein is strong enough to elicit a potential biological response [18]. The free binding energy was calculated with:

$$\Delta G^{bind} = G^{complex} X - \left(G^{protein} + G^{Ligand}\right)$$

2.10. ADMET prediction

The top five compounds as well as the co-crystallized ligand (the standard), were analyzed for their ADMET properties using the ADMETlab (https://admetlab3.scbdd.com/). and SwissADME (http://www.swissadme.ch/) web servers.

2.11. Molecular dynamics simulation

The top compounds based on their docking scores with the co-crystalized ligand were subjected to a 200 ns molecular dynamics (MD) simulation using the Desmond tool of the Schrodinger. The simulation was carried out with an explicit solvent system, and to maintain neutrality, Na+, and Cl-were added at 300^{O}K and 1.01325 bar. An orthorhombic box with a 10 × 10 × 10 Å buffer and 0.15 M salt was

used to prepare the model for the physiological environment, with the transferable intermolecular potential-4 point (TIP4P) force field. After the MD simulation, the trajectories were analyzed for RMSF, RMSD, and protein-ligand contacts using the interaction diagram module of the Desmond-Schrodinger suite.

3. Results and discussion

Breast cancer is the most prevalent malignancy among women and a leading cause of cancer-related mortality worldwide. To address this challenge, various studies have utilized various drug discovery approaches, including virtual screening of large chemical databases and advanced drug design techniques. The human kinesin Eg5, a spindle motor protein, plays a crucial role in the assembly and maintenance of the bipolar spindle during mitosis [19]. Its essential function in cell cycle progression has made it an attractive therapeutic target for halting tumor growth by disrupting mitosis.

In recent years, advancements in deep learning, particularly long short-term memory (LSTM) recurrent neural networks (RNN), transformers, and state-space models (SSM), have shaped various domains of artificial intelligence, including natural language processing, time-series prediction, and drug discovery [20]. These architectures have demonstrated remarkable success in modeling sequential data (e.g SMILES), capturing temporal dependencies, and learning complex patterns that were previously unattainable with traditional machine learning approaches.

In this study, we leverage one of these techniques to design novel inhibitors for kinesin Eg5, A character-level long short-term memory (LSTM) model was initially trained on the SMILES representations of over 1.8 million compounds sourced from the ChEMBL database. This model was subsequently fine-tuned using compounds known to exhibit activity against mitotic kinesin Eg5. The fine-tuned model generated novel molecules, which were subjected to a rigorous screening pipeline. First, a classification model, built using message passing neural networks (MPNN) implemented via ChemProp, was employed to predict the activity of the generated compounds. Next, Pan Assay Interference Compounds (PAINS) filters were applied to eliminate potential off-target molecules. Finally, structure-based screening methods were utilized to identify promising candidates with strong potential as novel inhibitors of kinesin Eg5.

3.1. Long short term memory (LSTM) model

A prior model was trained with the objective of enabling the model to understand the correct sequence of SMILES strings required to form valid chemical compounds. This model underwent training for 5000 epochs, during which a notable reduction in error was observed, as shown in Fig. 2. To evaluate the model's generative capabilities, it was used to generate 1 million characters which was equivalent to 18,557 SMILES strings. Of these, 9404 (50.68 %) were valid SMILES strings.

Further analysis revealed 15 duplicate entries, resulting in a high uniqueness of 96.92 %. This left 9009 (48.55 %) SMILES strings classified as novel molecules. Following this, the model was fine-tuned on a dataset of molecules with known inhibitory activity against Eg5. The same evaluation metrics were applied to assess the fine-tuned model's performance. Upon generating another 1 million characters, this translated to 21,142 SMILES strings, among which 13,915 (65.82 %) were valid. However, 1011 duplicate entries were identified, significantly reducing the uniqueness metric to 35.34 % (Table 1). This resulted in 3906 novel molecules.

The fine-tuning process led to a marked improvement in the number of generated and valid SMILES strings, which can be attributed to the additional training on Eg5-specific molecules. This improvement is also reflected in the more pronounced descent observed in the loss curve across epochs (Fig. 3). However, the high number of duplicates

Table 1
Metrics of the generative models.

Metric	Prior Model	Fine Tuned Model
Total Samples Generated	1,000,000	1,000,000
Total Generated SMILES	18557	21142
Valid SMILES	9404	13,915
Unique Generated SMILES	9114	4917
Validity (%)	50.68	65.82
Uniqueness (%)	96.92	35.34
Novelty (%)	48.5	18.5
Duplicates Found	15	1011
Novel Valid SMILES	9009	3906

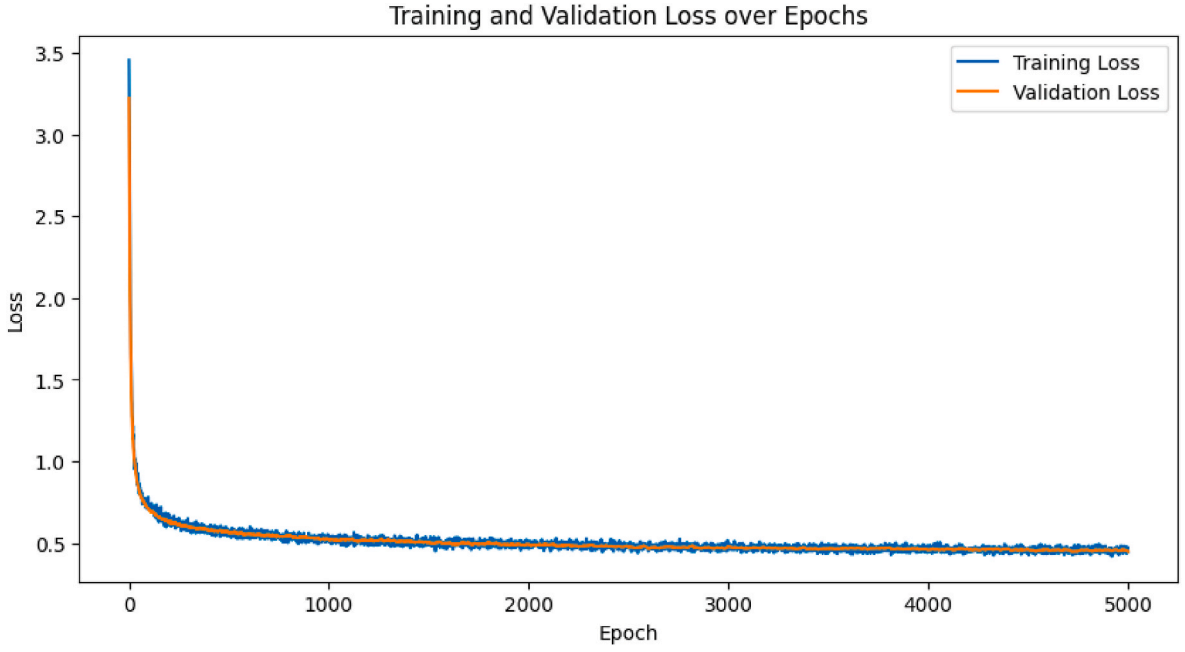


Fig. 2. Training and Validation Loss over 5000 epochs for the Prior model.

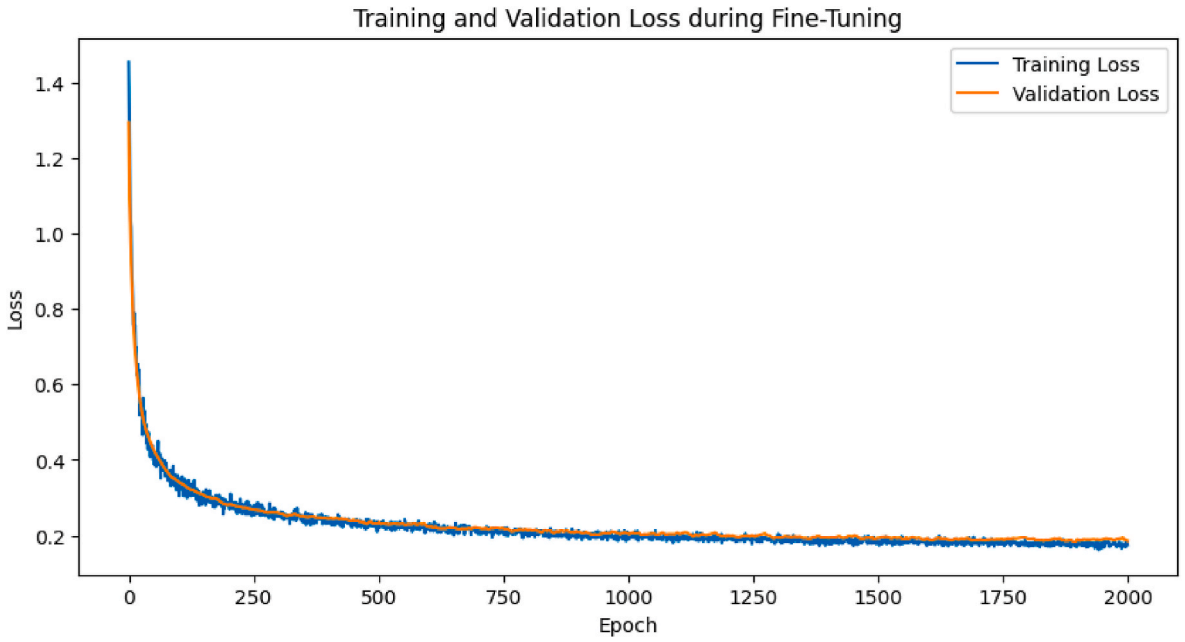


Fig. 3. Training and Validation Loss over 2000 epochs for the fine tuned model.

generated by the fine-tuned model highlights a potential overfitting issue. Additionally, this outcome may be influenced by the choice of the scale factor (1.0) or temperature parameter, which controls the randomness of the generated SMILES characters. A lower scale factor corresponds to a higher temperature, leading to greater randomness in the outputs, whereas a higher scale factor reduces randomness, making the generation more deterministic [21].

3.2. Chemical space visualization

The novel molecules generated by both the prior and fine-tuned models were visualized using a t-SNE plot. The plot revealed a clear separation between the random molecules generated by the prior model and the potential inhibitors produced by the fine-tuned model (Fig. 4). When the novel molecules from the fine-tuned model were compared to

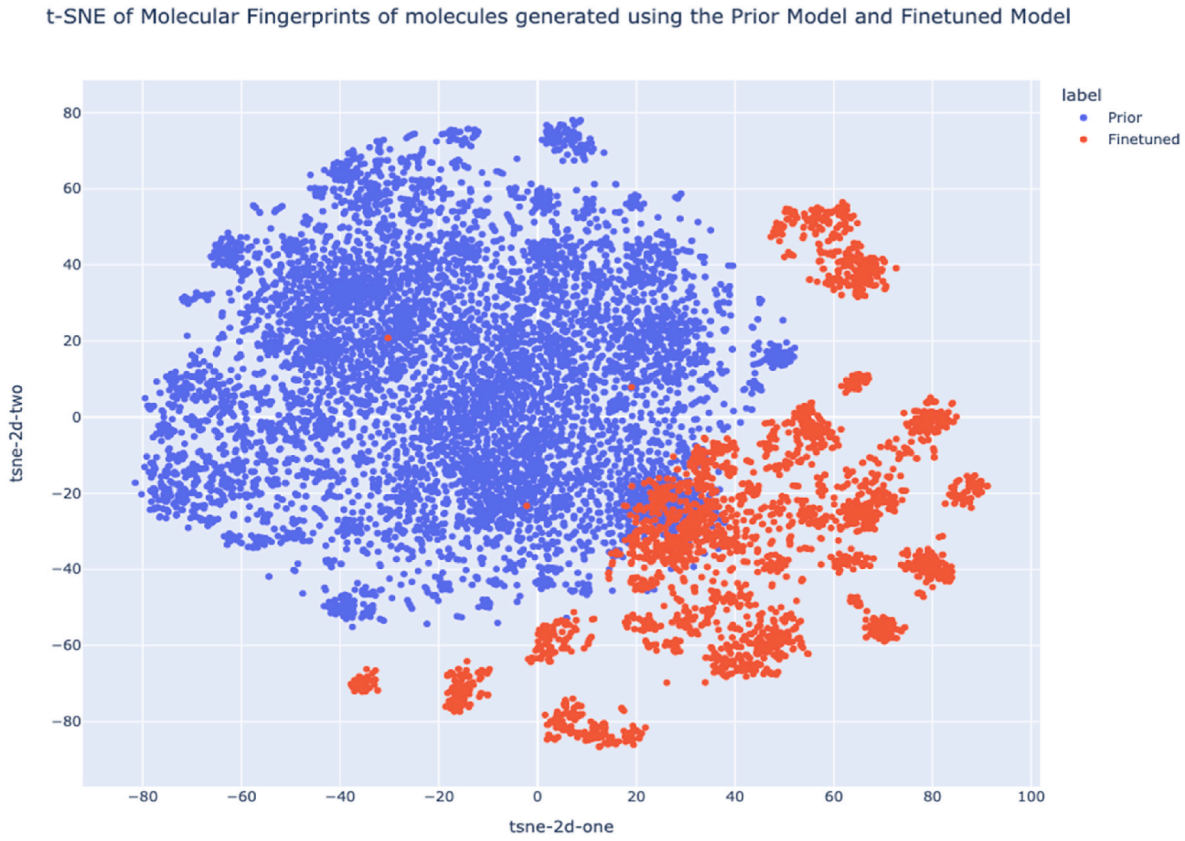


Fig. 4. t-SNE plot of the molecules from the prior and finetuned model.

known Eg5 inhibitors, there was significant overlap, indicating structural similarity (Fig. 5). This overlap shows the potential of these newly generated compounds as prospective Eg5 inhibitors.

Further analysis was performed using grouped bar charts to compare selected 1D-descriptors of the molecules generated by the two models. The fine-tuned model produced molecules with key characteristics such as having within 3–4 hydrogen bond acceptors, a higher number of rings, and more aromatic rings compared to the molecules generated by the prior model (Fig. 6). Scatter plots comparing various 1D-descriptors and medicinal chemistry parameters further highlighted the advantages of the fine-tuned model. The molecules it generated exhibited a better distribution of QED (Quantitative Estimate of Drug-likeness) values, suggesting an improved drug-likeness profile. Additionally, a negative correlation was observed between QED and molecular weight (MW), which aligns with the general observation that smaller molecules often have higher QED scores. The analysis also revealed a reduced TPSA (Topological Polar Surface Area) for molecules from the fine-tuned model, a parameter linked to better membrane permeability. Furthermore, the synthetic accessibility (SAS) scores of these molecules were concentrated between 3 and 4, indicating moderate ease of synthesis (Fig. 7). The fine-tuned model generated novel compounds with improved properties, making them promising Eg5 inhibitor candidates.

3.3. Scaffold diversity and novelty

We analyzed the scaffolds of the generated novel molecules, comparing them to existing Eg5 inhibitors and identifying newly generated scaffolds. Four of the top ten scaffolds were shared between the generated molecules and Eg5 inhibitors (Fig. 8). Additionally, new scaffolds were identified, though some appeared overly complex and potentially challenging to synthesize, this has also been noticed in another study too [10].

Also using the Hamiltonian diversity, we measured the diversity of the generative models (Prior and Fine Tuned). Hamiltonian diversity (HamDiv) is a novel metric that measures molecular diversity by calculating the length of the shortest Hamiltonian circuit in a complete graph [11]'. In this graph, each node represents a molecule, and the weight of each edge is the distance between a pair of molecules. Hamiltonian diversity is a more comprehensive and principled approach to measuring molecular diversity compared to existing metrics. It addresses the limitations of metrics like Richness and IntDiv by considering both the quantity and dissimilarity of molecules and adhering to the principles of monotonicity and dissimilarity. In this study, we employed the greedy approach to calculate HamDiv due to the number of molecules (>1000) generated by both models. With a HamDiv value of 5081.98 for the prior model (n = 9009 molecules) indicates the molecular set that is more diverse than that of the fine-tuned model, which has a HamDiv value of 1632.49 (n = 3906 molecules). This means that the molecules generated by the prior model are either more dissimilar to one another or more numerous, resulting in a broader exploration of the chemical space compared to the fine-tuned model. This outcome is expected, as the prior model was trained on general ChEMBL molecules, while the fine-tuned model was further trained specifically on Eg5 inhibitors.

3.4. Machine learning model

Using ChemProp, a message-passing neural network that represents molecules as graphs with atoms as nodes and bonds as edges, we developed a classification model to predict the bioactivity of compounds against Eg5 [14]. The model was evaluated using standard classification metrics such as accuracy, precision, recall, F1 score, specificity, and ROC-AUC, along with confusion matrices (Fig. 9 and Table 2) to assess its prediction performance. As shown in Table 2, the model achieved consistent accuracy on the training (0.85), test (0.83), and validation (0.80) sets. This indicates that the model generalizes well to new data. Precision, which tells us how many of the predicted active compounds were actually correct, remained steady on the training and test sets (0.83) but dropped to 0.72 on the validation set. This suggests a slight increase in false positives on unseen data. In contrast, recall, which measures how well the model identifies actual active compounds, remained high across datasets. It was 87 percent for training, 78 percent for test, and 85 percent for validation. This is important in virtual screening because missing potential actives (false negatives) can be more risky than identifying some inactive compounds as active. The F1 scores, which combine both precision and recall, ranged from 0.78 to 0.85, showing balanced performance. Specificity, which reflects the model's ability to correctly classify inactive compounds, ranged from

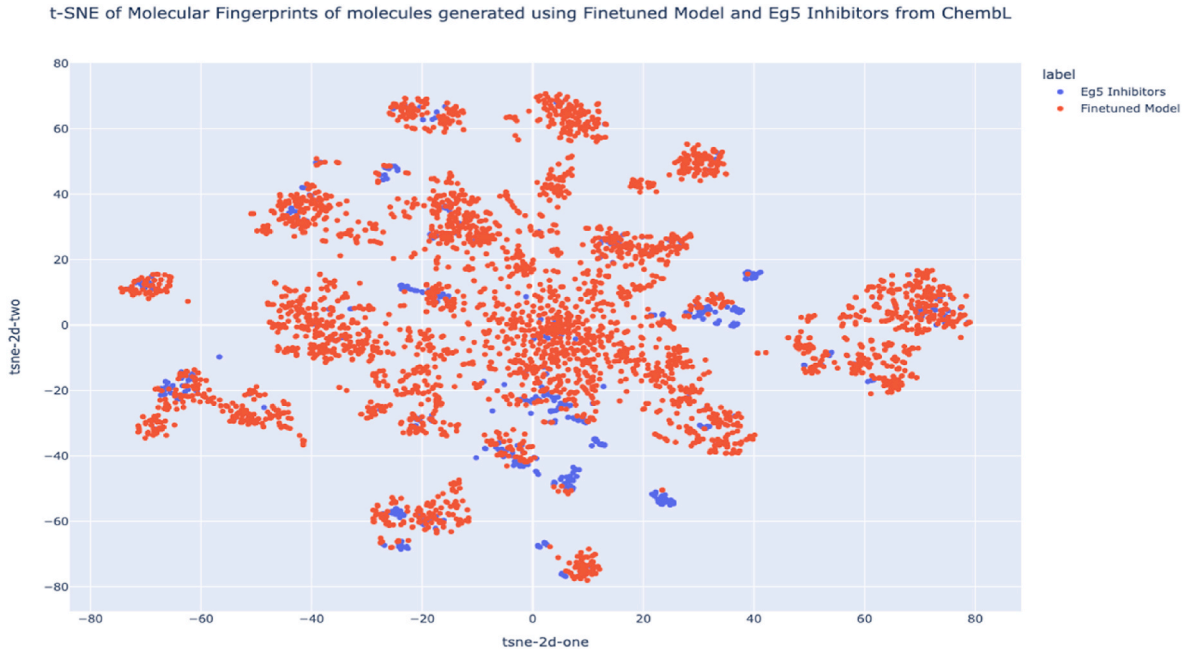


Fig. 5. t-SNE plot of the molecules from the finetuned model and current Eg5 inhibitors.

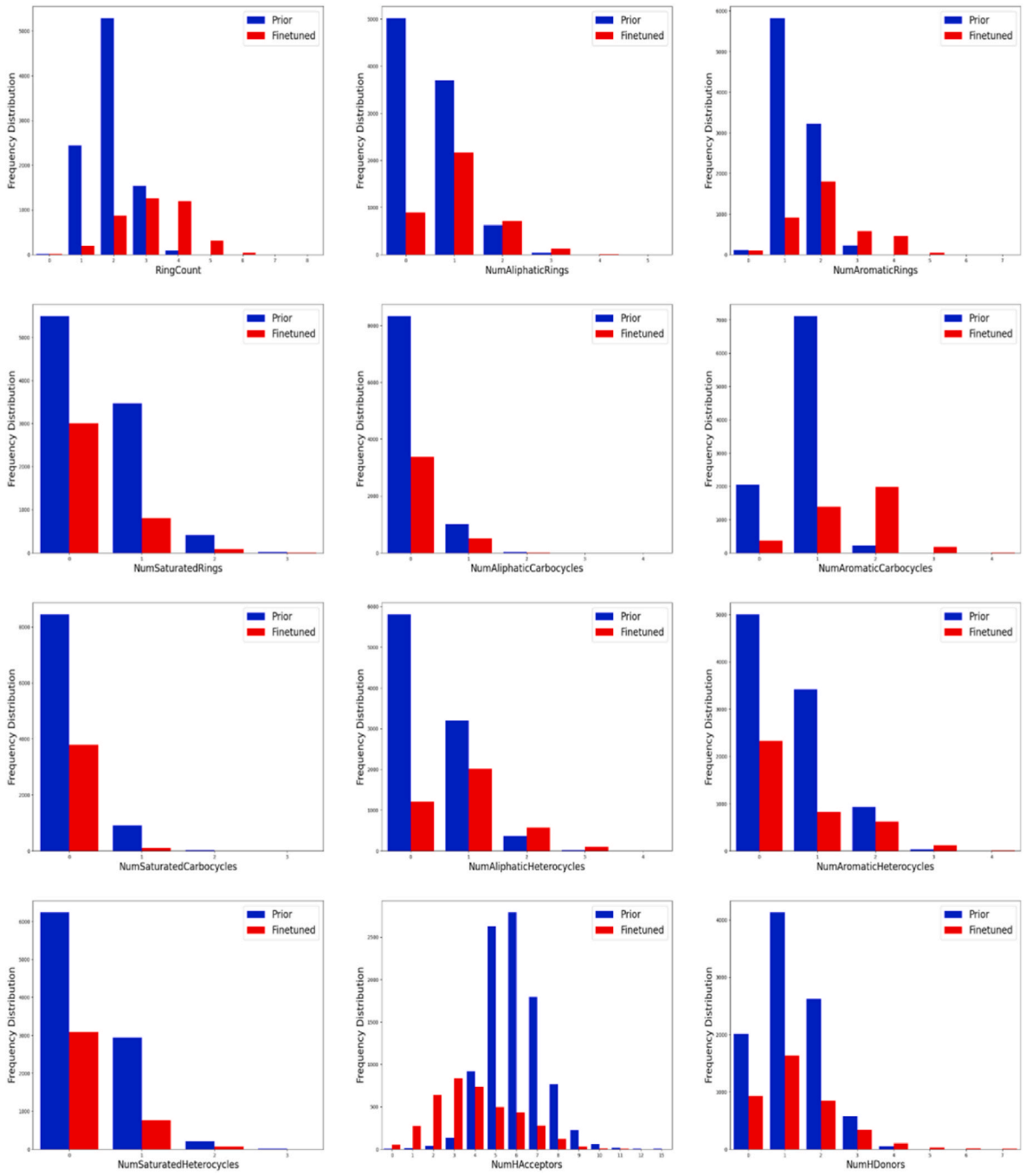


Fig. 6. Comparison of the molecules generated via the prior and finetuned model by selected 1D-descriptors.

0.77 to 0.86. This helps reduce the number of false positives. In addition, the ROC curves showed strong overall performance, with area under the curve (AUC) values exceeding 0.89 for all datasets. A high AUC means the model is effective at distinguishing between active and inactive compounds, confirming its robustness and reliability for virtual screening.

The model was then applied to filter 3906 novel molecules generated from the fine-tuned model, classifying them into 2578 inactives and 1328 actives. To validate its performance, the model was tested on an external dataset of known inhibitors and decoys. It accurately identified most of the inhibitors, although it struggled with certain natural products reported to inhibit Eg5, such as Terpendole E, Gossypol, and Adociasulfate-2. The model also performed well with the decoys, with the exception of Metformin, a known antidiabetic agent (see Supplementary Fig. 1).

3.5. Virtual screening to filter pan-assay interference compounds (PAINs)

The goal of drug discovery is to identify a ligand that binds specifically to a target protein (on-target) while avoiding interactions with unintended protein targets (off-target). Achieving selectivity is crucial to minimize side effects and improve therapeutic efficacy. However, certain compounds or substructures, known as PAINS (Pan-Assay Interference Compounds), have been reported to interfere with assay results, leading to false-positive or misleading biological activity across unrelated targets and testing methods [22]. PAINS compounds can exhibit promiscuous behavior due to various mechanisms, including chemical aggregation, chelation, singlet oxygen production,

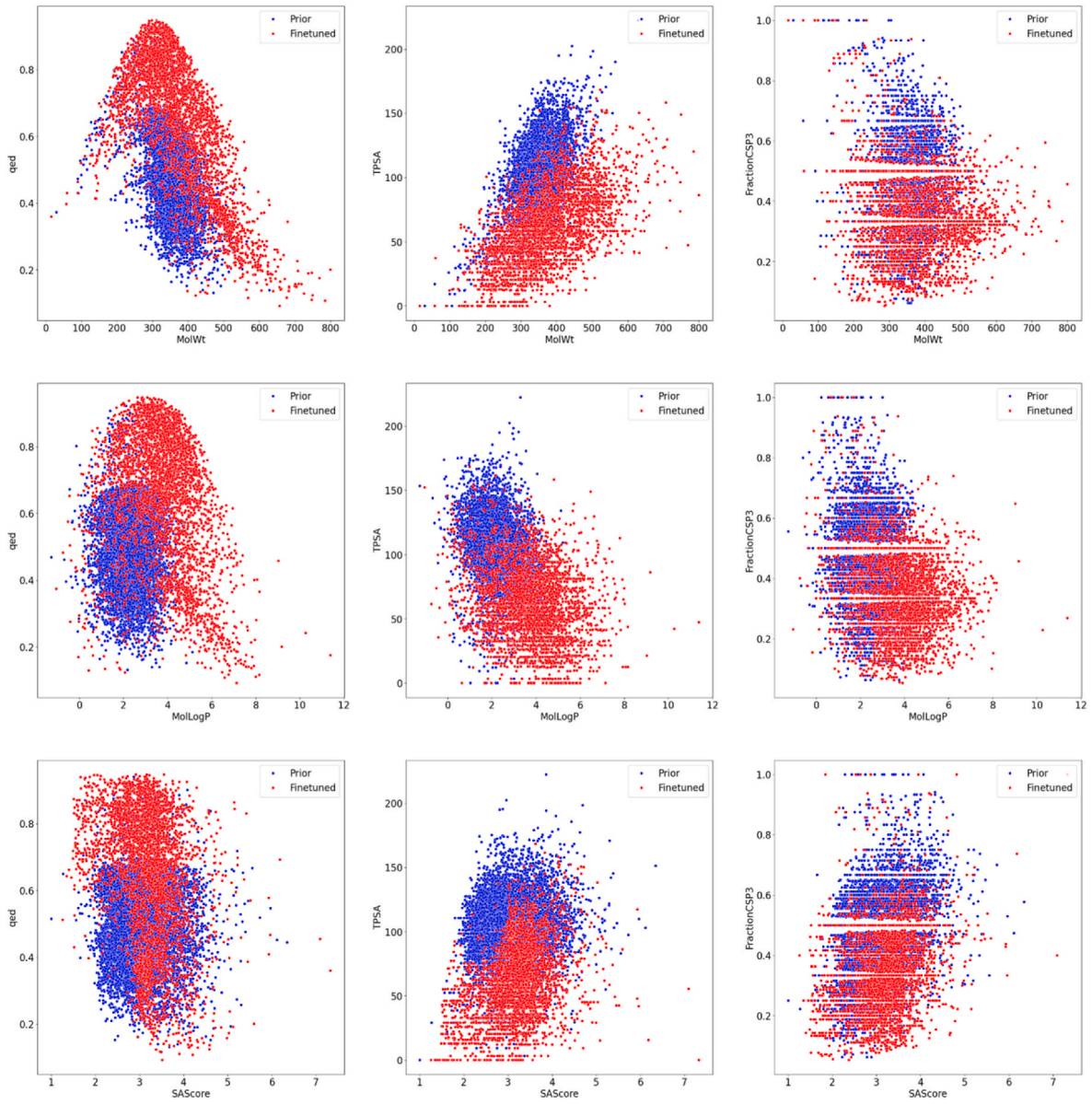


Fig. 7. Correlation of selected descriptors of the molecules generated via the prior and fine tuned model.

fluorescence effects, redox activity, and nonselective reactivity with proteins [23]. To address these issues, we filtered the predicted active compounds in this study to exclude those containing PAINS. This reduced the number of compounds to 963. By doing so, we increased the likelihood that the potential hits have true on-target activity, minimizing the chances of assay interference or off-target effects. We also highlighted the structures and names of the PAINs in some of the compounds we filtered (Fig. 10).

3.6. Molecular docking and MMGBSA

Molecular docking is a technique that is used to screen for compounds that can bind to a target protein and can therefore be used as a medication. The binding affinity of compounds to a target protein evaluates the stability of that compound when it is bound to the protein, the higher the binding affinity the more chances of success for the compound as a drug. The 963 already PAINS-filtered compounds were docked to the target protein in a precision based flow, initially docked using the SP scoring function and finally XP scoring function which we reported with. The SP (Standard Precision) scoring function in Glide is designed for fast and accurate screening of large compound libraries, making it suitable for docking tens to hundreds of thousands of ligands efficiently [24]. It uses a softer scoring approach to identify promising ligands quickly, even when the predicted binding poses have minor imperfections. While the XP (Extra Precision) scoring function is intended for smaller ligand libraries and offers higher accuracy. It uses more thorough sampling and an advanced scoring method that considers detailed molecular interactions such as water displacement, hydrogen bonding, desolvation, entropic effects, and metal–ligand interactions [25]. This helps to better separate active compounds from inactive ones and reduce false positives.

The top compounds had XP docking scores which range from −8.332 kcal/mol to −8.839 kcal/mol, all higher than the XP docking score of the co-crystallized ligand, −8.323 kcal/mol (Fig. 11; Table 3). Compound 103 had a docking score of −8.839 kcal/mol, by forming hydrogen bonds with GLU 116 and GLY 117. It forms hydrophobic interactions with ALA 133, ALA 218, ALA 219, GLY 217, ILE 136, LEU 160, LEU 214, PRO 137, TRP 127 and TYR 211. Compound 69 had a docking score of −8.827 kcal/mol, by forming hydrogen bonds with GLU 116 and GLY 117. It made hydrophobic contacts with ALA 133,

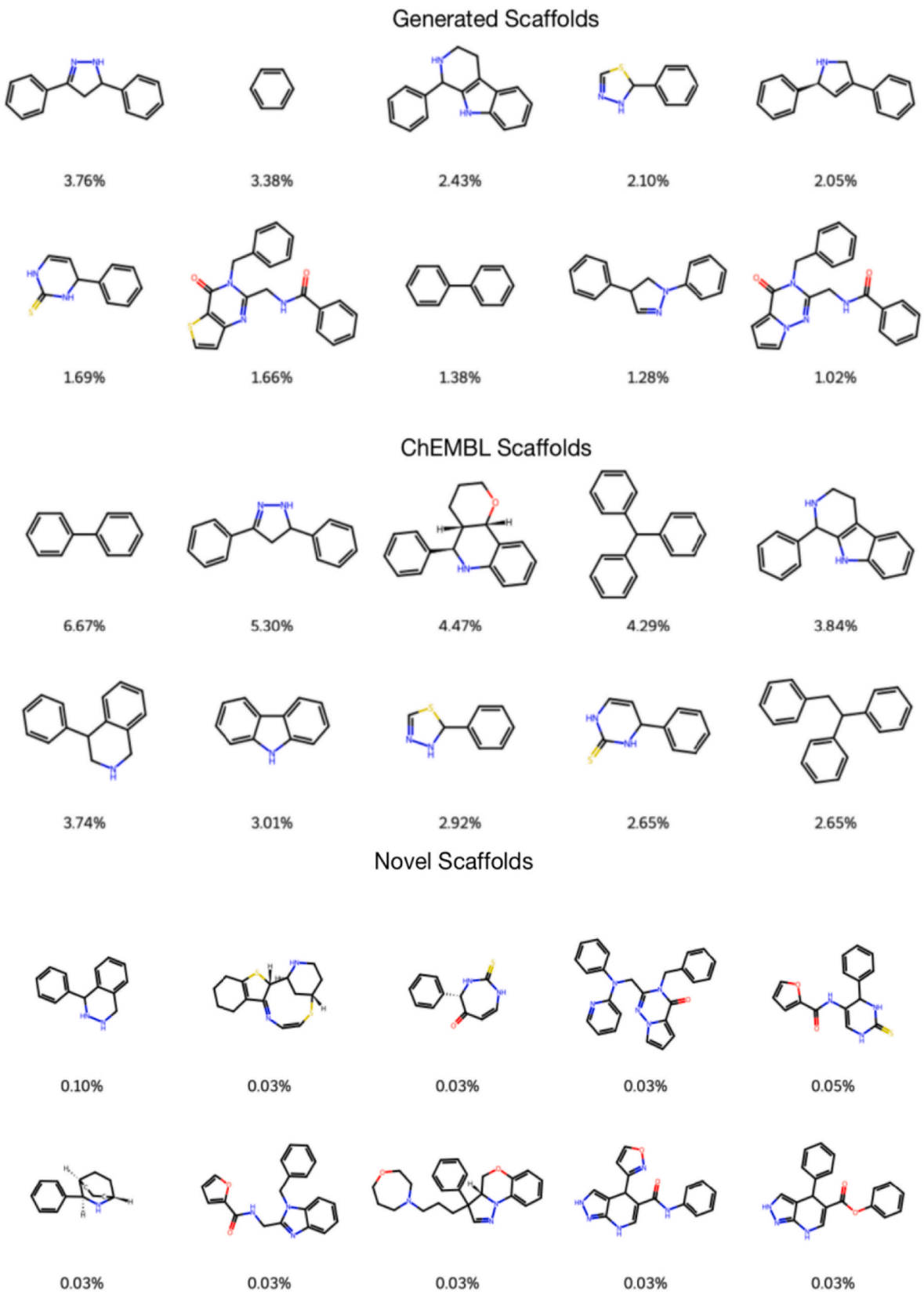


Fig. 8. Top 10 most common scaffolds from the fine-tuned model's generated molecules, current Eg5 inhibitors, and newly identified novel scaffolds. The prevalence of each scaffold is expressed as a percentage.

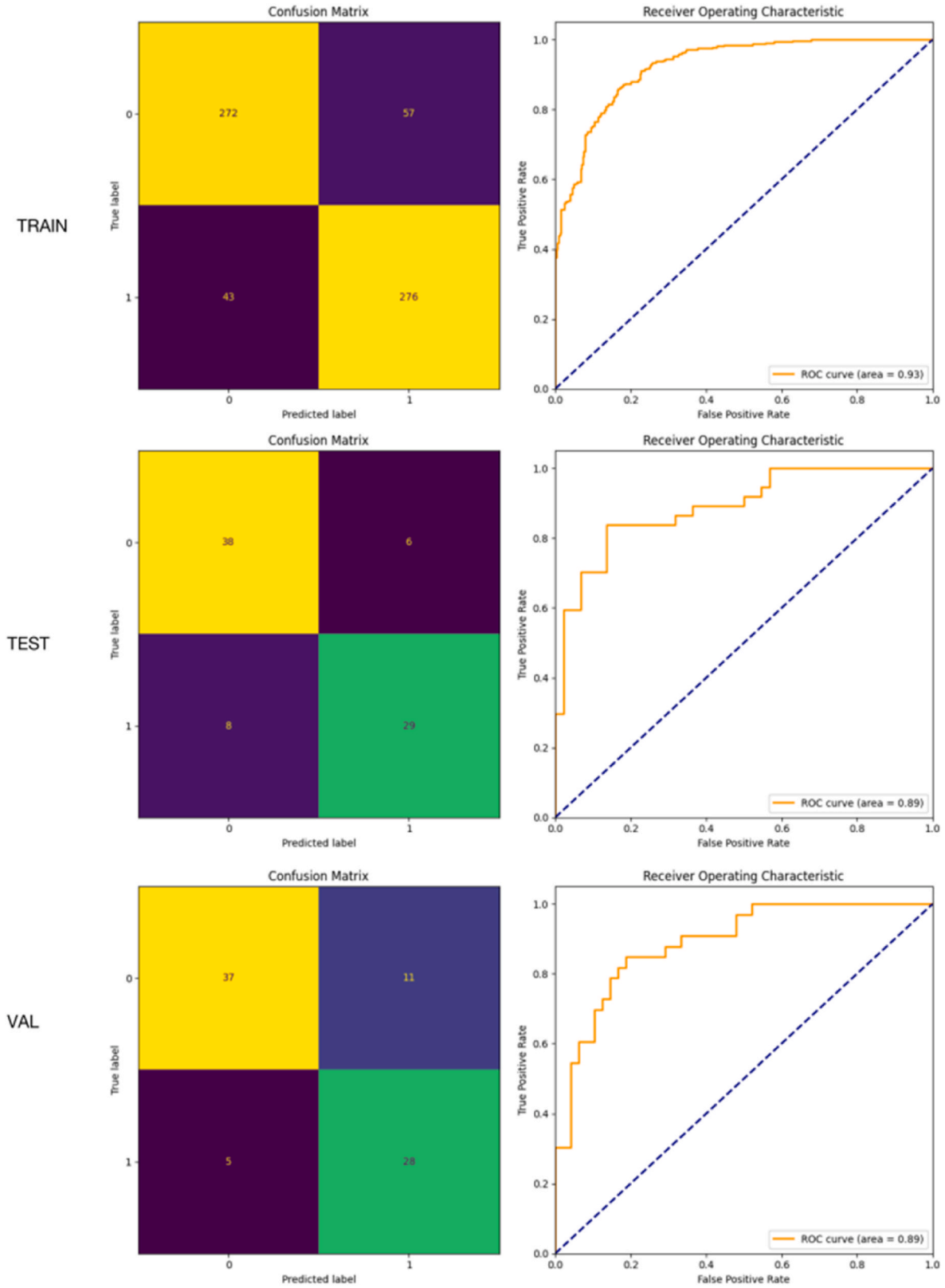


Fig. 9. Confusion matrices and ROC curve plots for the train, test, and validation sets.

ALA 218, GLY 217, ILE 136, LEU 160, LEU 171, LEU 172, LEU 214, PHE 239, PRO 137, TRP 127, and TYR 211. Compound 888 had a docking score of −8.44 kcal/mol, by forming hydrogen bonds with GLY 117. It forms hydrophobic bonds with ALA 133, ALA 218, GLY 217, ILE 136, LEU 160, LEU 171, LEU 214, PHE 239, PRO 137, TRP 127, and TYR 211. Compound 177 had a docking score of −8.374 kcal/mol, by forming hydrogen bonds with GLU 116 and GLY 117. It forms hydrophobic bonds ALA 133, ALA 218, GLY 217, ILE 136, LEU 160, LEU 171, LEU 214, PHE 239, PRO 137, TRP 127, and TYR 211. Compound 797 had a docking score of −8.368 kcal/mol, by forming hydrogen bonds with

Table 2
Classification metrics of the machine learning model.

Dataset	Accuracy	Precision	Recall	F1 Score	Specificity
Train	0.85	0.83	0.87	0.85	0.83
Test	0.83	0.83	0.78	0.81	0.86
Val	0.8	0.72	0.85	0.78	0.77

GLU 116 and GLY 117. It forms hydrophobic bonds ALA 133, ALA 218, GLY 217, ILE 136, LEU 160, LEU 214, PHE 239, PRO 137, TRP 127, and TYR 211 (Fig. 12). As a reference, the co-crystallized ligand had a docking score of −8.323 kcal/mol,It formed hydrogen bonds with GLU 116 and GLY 117 and hydrophobic interactions with many of the same residues as the generated compounds, including ALA 133, GLY 217, and LEU 214. The generated compounds demonstrated stronger binding affinity due to little differences in chemical structure, which likely enhanced the interaction made.

The Molecular Mechanics Generalized Born Surface Area (MM-GBSA) method is a widely used and reliable way to calculate binding free energy [26]. It is more accurate than most traditional scoring methods [27]. In this study, the top ligands showed strong binding, with free energy values lower than −56.8, which indicates good binding strength. Among them, Compound 103 stood out as the best, with a binding free energy of −82.68. This is better than the co-crystallized ligand (STLC), which had a binding free energy of −76.98. The stronger binding energy of Compound 103 suggests it has a better fit and stronger interactions with the receptor. Compound 103 stood out with the strongest binding affinity (XP docking score: 8.839 kcal/mol) and the most favorable predicted binding free energy (MM-GBSA: 82.68 kcal/mol), outperforming the reference ligand, STLC (−8.323 kcal/mol and −76.98 kcal/mol, respectively). While several top compounds exhibited better docking scores than the reference, only Compound 103 also surpassed it in MM-GBSA binding energy. This superior binding profile, combined with its hydrogen bonding interactions provides a clear and quantitative basis for its selection as the lead compound in this study.

3.7. Drug-likeness and ADMET prediction

The key physicochemical properties and drug-likeness of the top compounds were evaluated, with the results summarized in the table below. The analysis revealed that the molecular weights of the top compounds, including the co-crystallized ligand, range from 354.14 to 431.19 g/mol—well within the standard acceptable limit of less than 500 g/mol. Lipophilicity, which measures a compound's ability to dissolve in lipids or non-polar solvents, plays a crucial role in drug absorption. Drugs with optimal lipophilicity can effectively cross the lipid-rich cell membranes of the gastrointestinal tract (GIT), enhancing oral absorption [28,29]. According to widely accepted criteria such as Lipinski's Rule of Five, a lipophilicity range (log P) of 0–5 is considered optimal for drug-likeness [30–32]. The co-crystallized ligand (STLC) exhibits a log P value of 2.91, while the consensus log P values of the top compounds fall within a range of 2.80–4.05. These optimal lipophilicity values suggest that the compounds are likely to be efficiently absorbed from the GIT into the systemic circulation, enabling them to exert their therapeutic effects effectively (see Supplementary Table 1).

The solubility patterns of the compounds were analyzed using various classification methods. According to the Log S (ESOL) classification, all compounds were soluble, except for Compound 69, which was moderately soluble. A similar trend was observed with the Ali solubility classification, where Compound 69 and Compound 103 were categorized as moderately soluble. The predicted lipophilicity of the compounds supports their potential for good oral absorption, as previously noted. All the compounds, including the co-crystallized ligand, demonstrated high gastrointestinal tract (GIT) absorption. Regarding blood-brain barrier (BBB) permeability, only Compounds 888 and 177 showed the ability to cross the BBB, indicating a lower likelihood of central nervous system (CNS) toxicity and side effects for the other compounds. Furthermore, most compounds, including Compounds 69, 103, 177, 797, and the reference ligand, were identified as substrates for permeability glycoprotein (P-gp). P-gp, a key efflux transporter, moves its substrates from intracellular to extracellular spaces, such as the GIT lumen, thereby reducing their potential toxicity. These findings provide valuable insights into the pharmacokinetic behavior of the drug candidates, guiding further pharmacological evaluations and potential therapeutic applications [28,33,34].

Understanding how compounds interact with the cytochrome P450 (CYP) enzyme system is vital for predicting the pharmacokinetics of potential drugs, as these enzymes play a central role in the metabolism and elimination of medications [35]. Inhibition of CYP isoforms can lead to poor drug clearance and drug-induced toxicity, making it essential for potential drug candidates to exhibit minimal inhibitory activity against these enzymes [32]. In this study, compounds 69 and 797 inhibited only one of the five tested CYP isoforms, while the other compounds inhibited no more than three of the five isoforms. This suggests that these compounds are likely to be efficiently metabolized by liver enzymes, enabling their effective removal from the body. Additionally, all compounds adhered to Lipinski's Rule of Five, demonstrating favorable drug-like properties, with consistent bioavailability values of 0.55. To further assess the suitability of these compounds, PAINS (Pan-Assay Interference Compounds) and BRENK alert analyses were conducted, even though initial filtering had already removed PAINS-positive compounds. PAINS alerts are used to identify molecular substructures that may produce false-positive results in biological assays, while BRENK alerts detect structural features linked to potential toxicity or metabolic instability [35]. Notably, none of the compounds exhibited PAINS or BRENK alerts, indicating a reduced likelihood of false positives, toxicity, or metabolic instability. These findings show the potential of the compounds in this study to serve as potential safe and metabolically stable drug candidates. Toxicity prediction of each compound showed that none is predicted to be hematotoxic or carcinogenic. The co-ligand and Compound 103 are predicted to be toxic on the liver. In contrast, moderate hepatotoxicity was predicted for Compounds 69 and 797, while Compounds 888 and 177 were predicted to cause no drug-induced liver injury (DILI).

The QED (Quantitative Estimate of Drug-likeness) value evaluates the drug-likeness of a compound based on eight drug-related properties. For our top compounds, all except Compound 103 have QED values greater than 0.5. Compound 177 and Compound 888 have QED values of 0.829 and 0.917, respectively, approaching the ideal threshold of 1. In addition, none of the compounds have a synthetic accessibility score exceeding 3.597, indicating that they are relatively easy to synthesize [36]. To further demonstrate the feasibility of synthesizing these compounds, we predicted the synthetic route for Compound 103, the top-ranked compound, using the ASKCOS webserver [37]. Supplementary Fig. 6 shows a three-step synthetic route for Compound 103. The analysis indicates that all starting materials are commercially available, and each has a history of use as both reactants and products in previous syntheses, thereby supporting the synthetic tractability of this compound.

3.8. Molecular dynamics (MD) simulations

In this study, we ran MD simulations on the top compounds, including the co-crystallized ligand, throughout 200 ns (ns). This was carried out to examine the essential functions, underlying principles, and molecular mechanisms of action of the target protein and protein-ligand complexes by examining their conformational dynamics [38]. In addition to evaluating the stability and flexibility of the receptor-ligand complexes, this will allow us to determine how well the best-performing ligands bind to the Eg5 protein binding site. We used the Root Mean Square Deviations (RMSD), Root Mean Square

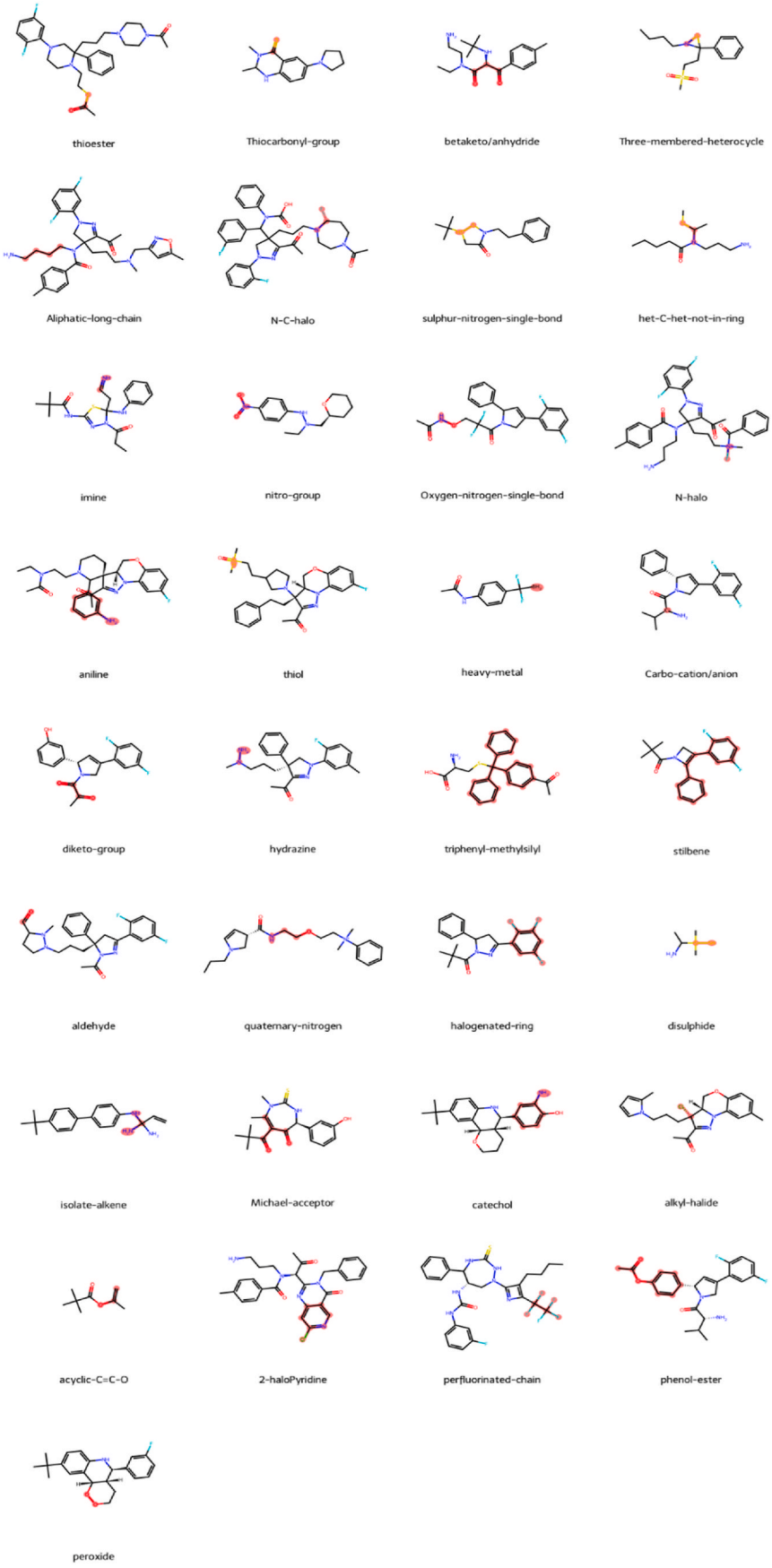


Fig. 10. PAINs highlighted from compounds filtered out.

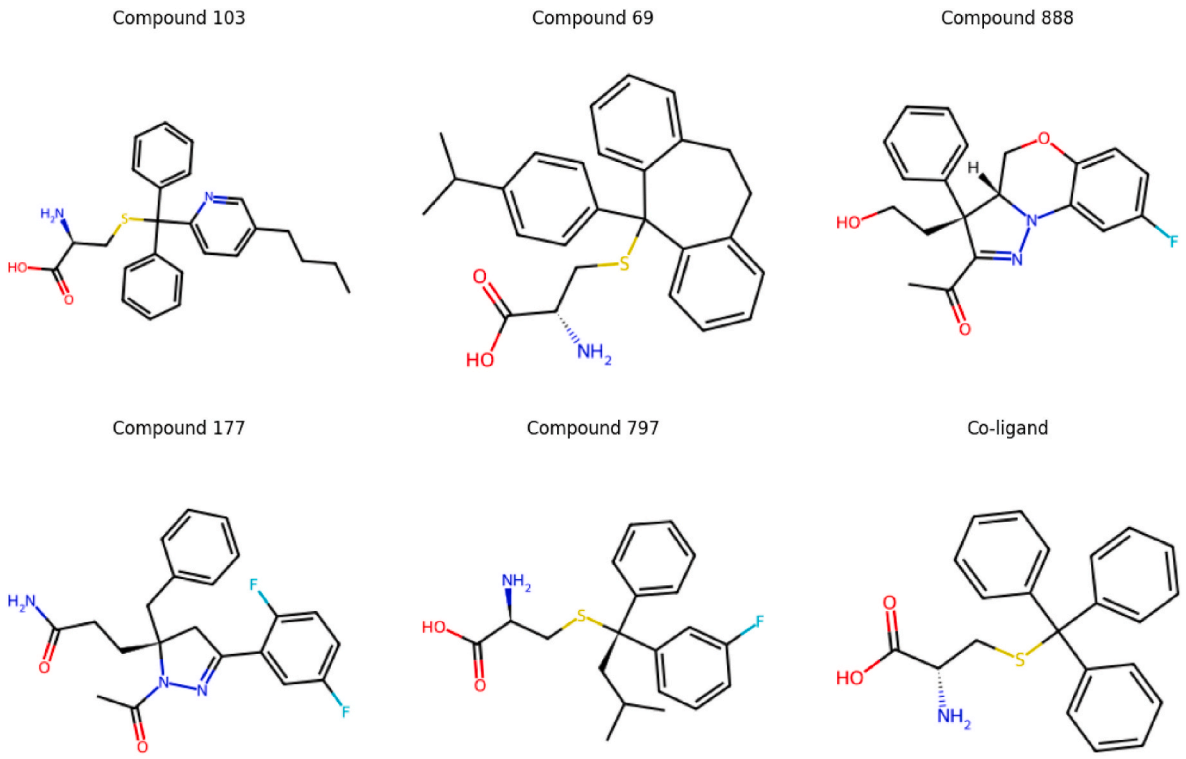


Fig. 11. Top five compounds and co-crystallized ligand (STLC) from this study.

Table 3
Molecular docking, MMGBSA scores and interacting amino acid residues of the top five compounds and co-crystallized ligand.

Title	Docking score (kcal/mol)	MMGBSA dG Bind	Hydrogen bond	Hydrophobic Interactions	Other Interactions
103	−8.839	−82.68	GLU 116, GLU 117	ALA 133, ALA 218, ALA 219, GLY 217, ILE 136, LEU 160, LEU 214, PRO 137, TRP 127, TYR 211	ARG 119, ARG 221, ASP 130, GLU 118, GLU 215, THR 112, THR 222
69	−8.827	−67.35	ARG 221, GLY 117	ALA 133, ALA 218, GLY 217, ILE 136, LEU 160, LEU 171, LEU 172, LEU 214, PHE 239, PRO 137, TRP 127, TYR 211	ARG 119, ASP 130, GLU 116, GLU 118, GLU 215, THR 112
888	−8.440	−65.29	GLY 117	ALA 133, ALA 218, GLY 217, ILE 136, LEU 160, LEU 171, LEU 214, PHE 239, PRO 137, TRP 127, TYR 211	ARG 119, ARG 221, ASP 130, GLU 116, GLU 118, GLU 215, GLY 134, THR 112
177	−8.374	−56.80	GLU 116, GLY 117	ALA 133, ALA 218, GLY 217, ILE 136, LEU 160, LEU 171, LEU 214, PHE 239, PRO 137, TRP 127, TYR 211	ARG 119, ARG 221, ASP 130, GLU 118, GLU 215, THR 112
797	−8.368	−75.74	GLU 116, GLY 117	ALA 133, ALA 218, GLY 217, ILE 136, LEU 160, LEU 214, PHE 239, PRO 137, TRP 127, TYR 211	ARG 119, ARG 221, ASP 130, GLU 118, GLU 215, THR 112
Co-Crystallized Ligand (S-trityl-L-cysteine)	−8.323	−76.98	ARG 221, GLY 117	ALA 133, ALA 218, GLY 217, ILE 136, LEU 160, LEU 214, PHE 239, PRO 137, TRP 127, TYR 211	ARG 119, ASP 130, GLU 116, GLU 118, GLU 215, THR 112

Fluctuation (RMSF), and protein-ligand interactions to evaluate the trajectory of the MD simulations.

3.8.1. Root Mean Square Deviation (RMSD)

RMSD is the mean variation in the displacement of a group of atoms with respect to a reference frame. It is used to measure the scalar distance for the protein and ligand throughout the trajectory of the simulation [39]. Following the alignment of each protein's frame on the reference frame backbone, the RMSD is computed using the side chains or C alpha [40]. Monitoring the protein's RMSD throughout the simulation can reveal information about the structural conformations. It can show if the simulation is equilibrated and whether its fluctuations are within a certain range of thermal averages. Changes of 1–3 Å are quite acceptable for very small, globular proteins. However, larger alterations than 3 Å indicate that the protein undergoes a substantial conformational change throughout the simulation [36].

The RSMD of the Eg5 protein is shown on the left y-axes of Fig. 13, which displays the RSMD plots. First, the backbone of the reference frame was used to align all protein frames, and the RMSD was calculated based on the atom selection [41]. The ligand's relative stability in relation to the protein and its binding is shown by the ligand's RSMD on the right y-axis. The "Lig fit Prot" displays the ligand's RMSD after the protein-ligand complex was first aligned at the reference's protein backbone, and the ligand-heavy atoms' RSMD was then measured. The x-axis shows the overall duration for the simulation in nanoseconds, the stability period, and the fluctuation period.

As illustrated, the RMSD of the protein for compound 103 increased from 1.2 Å to approximately 2.2 Å within the first 20 ns; this suggests that the protein remains close to the initial conformation (Supplementary Fig. 4). From 20 ns to 50 ns, protein RMSD fluctuates and increases until a noticeable peak is seen at 2.8 Å. After 50 ns, the RMSD stabilizes around 2.8 Å; this stabilization continued till the period of 180 ns, after which the RMSD increased to 3.2 Å and stabilized throughout the remaining period of the simulation. The RMSD values fall within the acceptable range, i.e., from 1.2 Å to 3.2 Å throughout the period of 200ns, with more stabilization phase observed at 2.8 Å

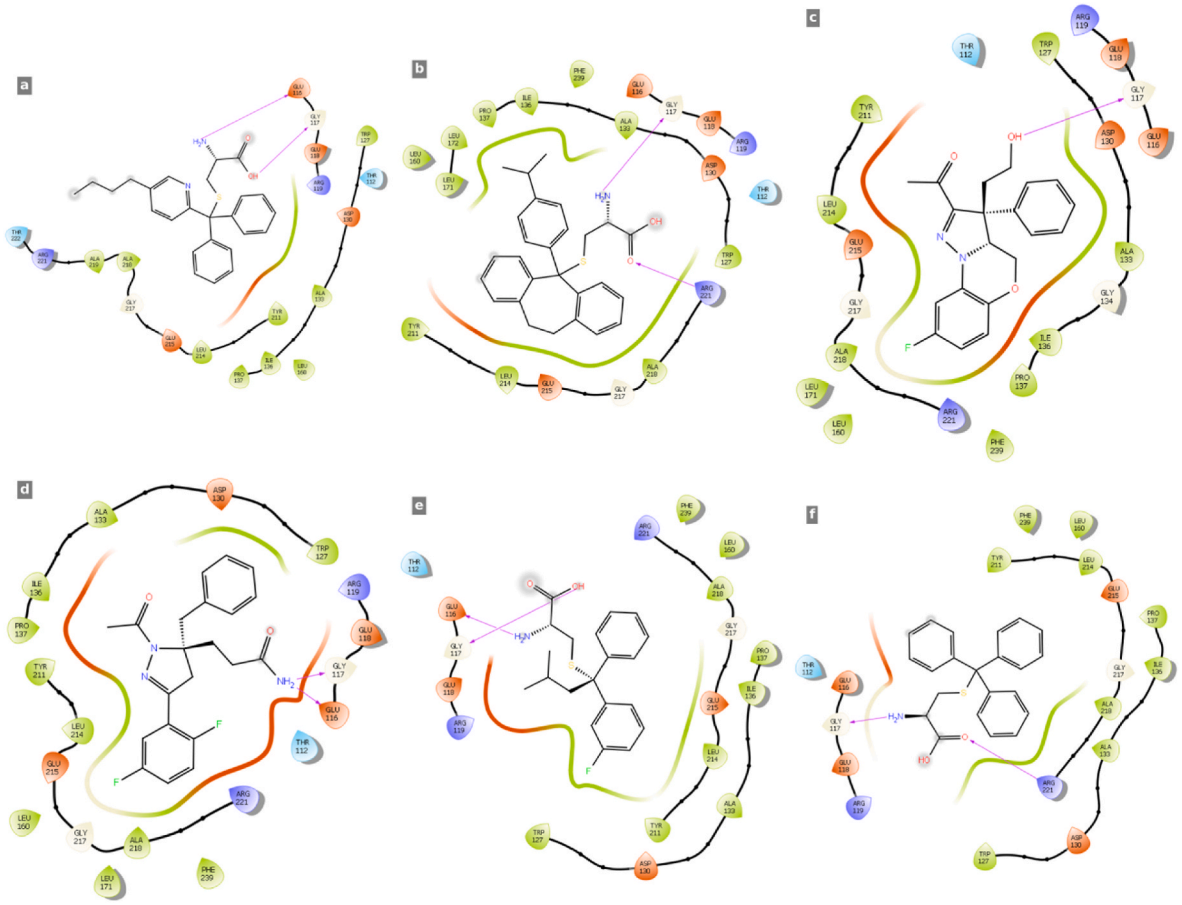


Fig. 12. 2D interacting plots of docking between *(A) Compound 103; (B). Compound 69 (C). Compound 888 (D). Compound 177 (E). Compound 797 (F). Co-crystallized ligand* (S-trityl-L-cysteine) with residues at the binding pocket of EG5.

between 50ns and 180ns. This suggests that the protein reaches a relatively stable conformation with occasional fluctuations. The ligand RMSD (Compound 103) ranges from 1.0 Å to 3.0 Å with sharp fluctuations, suggesting unstable binding.

As shown in Supplementary Fig. 4, the RMSD of the protein for compound 69 here was an initial significant deviation from the protein's backbone or Cα atoms for the first 75ns. After 75ns, a stable equilibration was observed from 3.5 Å to 4.0 Å throughout the remaining period of the simulation. This suggests an initial increasing conformational change of the protein, which eventually remains stabilized over time. However, the ligand RMSD (compound 69) showed a high fluctuation starting from 0.5 Å and continuously oscillated, reaching a peak at 3.0 Å around 125ns. After this, the RMSD decreases and stabilizes around 1.5 Å from 150 ns onwards. The ligand RMSD shows a more pronounced fluctuation, which suggests the ligand does not stay bound to the protein.

Fig. 13 shows a minimal fluctuation between 1.2 Å and 2.1 Å for the RMSD of the protein for compound 888. However, a sharp peak was observed at 2.7 Å around 175ns. After the initial fluctuation, which halted around 20ns, the oscillation stabilized until a sharp peak was observed. Since the conformational changes observed fall within the range of 1–3 Å, which is acceptable, this suggests the protein RSD is undergoing a minimal conformational change throughout the simulation [42]. The ligand RSMD showed large fluctuations; however, it had minimal deviation, which indicates that the ligand stays unfitted into the protein binding site.

As illustrated in Supplementary Fig. 4, the RMSD of the protein for compound 177 increased initially from 1.6A to 3.2A for the first 50ns. However, a sharp peak was seen at 3.6A, stabilizing at 3.2A at around 100ns until the simulation was completed. This suggests a minimal fluctuation and deviation from the throughout the protein's backbone. This minimal deviation also indicates minimal conformational change throughout the simulation. The ligand RSMD showed increased fluctuation at first, around 75ns, but remained stable at equilibrium throughout the simulation. Also, its minimal fluctuation shows a stable binding ability with the binding site of the protein.

The RMSD of the protein for compound 797, which increased slightly from 1.2A to 2.4A around 20ns; however remained stable for an extended period, the protein RSMD remained stable around 2.0 A to 2.4A from 50ns to 140ns until a sharp peak was noticeable after 140ns around 3.2A; the RSMD deviated minimally and remained very stable till the end of the simulation (Supplementary Fig. 4). The results of protein RSMD showed minimal fluctuation, which indicates good protein stability and reduced conformational change. The ligand RSMD suggested strong binding; this is due to its relatively constant values between 0.8 Å and 1.8 Å (Fig. 13). Compound 797 demonstrated highly stable binding, showing potential for effective interaction with the protein binding site. Regarding the co-crystalized ligand, the protein RMSD initially increased from around 1.6A to approximately 2.4A within the first 25 ns; another further increase was observed at 2.8A, indicating significant conformational changes. This is then followed by a stable oscillation from 30ns throughout the entire simulation, suggesting a low fluctuation and high stability. The ligand's RMSD increases from around 1.2 Å to approximately 2.4 Å, suggesting early binding. However, the ligand remains stable with minimal fluctuations, indicating a consistent binding within the protein's binding site.

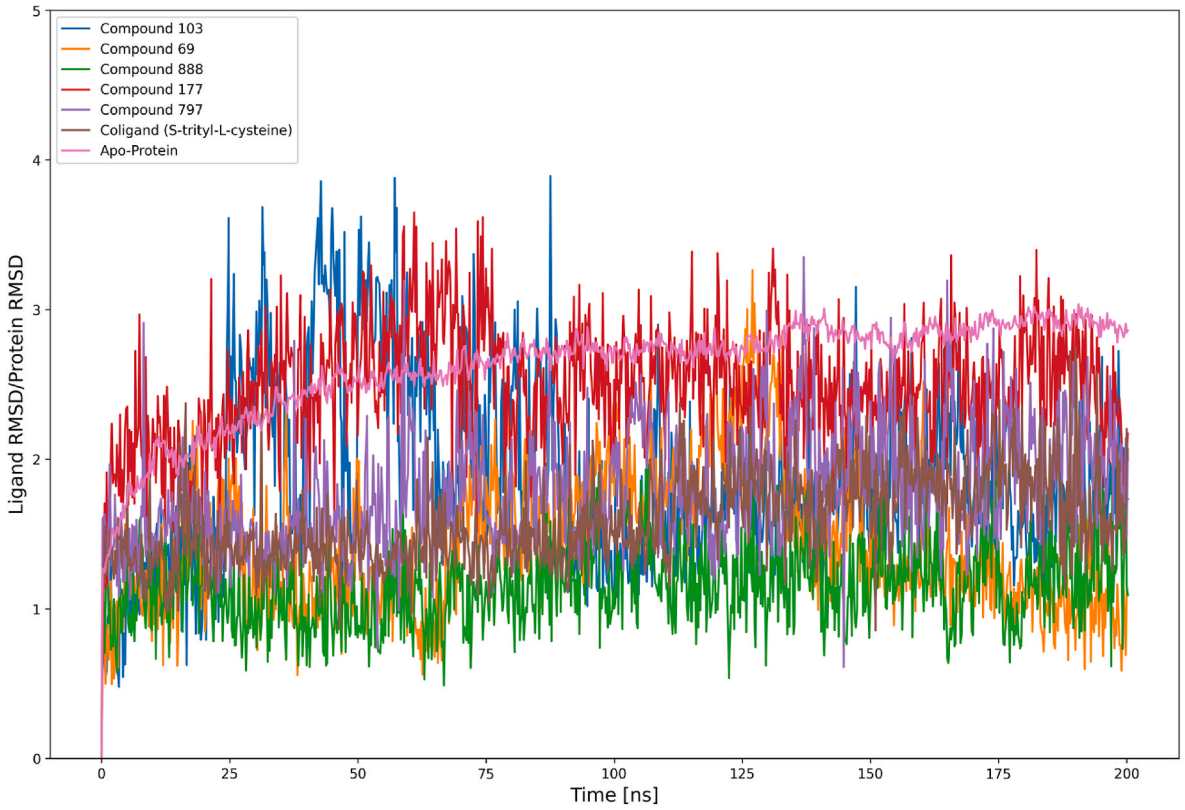


Fig. 13. Root Mean Square Deviation (RMSD) plot.

3.8.2. Root Mean Square Fluctuation (RMSF)

The RMSF is employed as a means to describe the variations in flexibility across residues in relation to the average MD simulation conformation [43]. There is a correlation between RMSF values and the degree of flexibility; the greater the value, the more flexible [44]. Higher values (peaks) indicated the presence of secondary structures like alpha-helices, beta-sheets, and -helices, which indicated structure stability. In comparison, lower values showed the presence of twists, loose bonding loops, terminal ends, and loose bonding, indicating structural flexibility [40]. The RMSF of a protein plotted against the atom index is

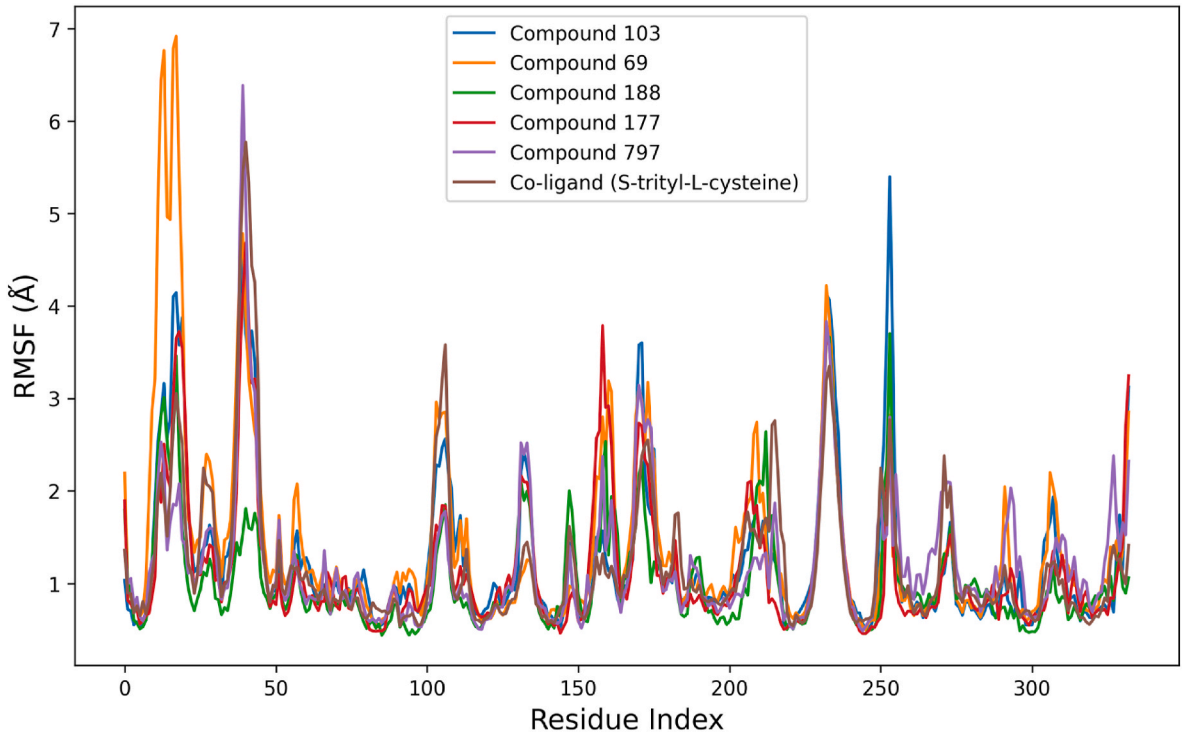


Fig. 14. Root Mean Square Fluctuation (RMSF) plot.

shown in Fig. 14(a–f). The protein's atom index is shown on the X-axis. Every number in the protein sequence represents a distinct atom. The RMSF values in Angstroms (Å), a unit of length frequently used to quantify atomic-scale distances, are displayed on the Y axis. The fluctuating RMSF values across different atom indices indicate varying degrees of flexibility.

Fig. 14 shows the plot of compound 103 with significant peaks in residue indices around 40 and 250. These significant peaks suggest that the ligand could adapt to the protein's binding pocket because they show areas involved in binding or unbinding processes [45]. The peaks show relatively high RMSF values, with residue 40 at approximately 4.6 Å and residue 250 at approximately 5.4 Å, indicating highly flexible regions, likely in a loop or unstructured part of the protein. However, most of the residues show a moderate level of flexibility with RMSF values ranging from 1.2 to 2.4 Å, which suggests minimal deviation in the positions of these residues when bound with Compound 103. The protein in complex with compound 69 exhibited a very high flexible region at residue index 25 with its protein RMSF value above 6.4 Å. While this particular region may indicate less stability, the majority of the peaks observed show a moderate level of flexibility with RMSF values ranging from 1.6A to 4.8 Å, which suggests moderate stability. On the other hand, the protein in complex with compound 888 showed a significant peak around 4.5 Å at residue index 270, which is slightly lower compared to significant peaks observed in plots of compound 103 and 69. This suggests that the decreased flexibility compared to compounds 103 and 69 might indicate better binding stability or reduced structural dynamics in the presence of compound 888 (Fig. 14).

Fig. 14 also illustrates the interaction of the protein in complex with compound 177, which shows a highly flexible region with an RMSF value of 4.8 Å for residue 50. However, other peaks were observed around indices 25, 150, and 250. Additionally, there were more moderate flexible regions at residue 30, 60, 100, 130, 200, and 300. The presence of more moderate flexible regions indicates relatively stable interaction within this region of the protein-ligand complex. The fluctuation ranges from 0.6 to 4.8 Å, suggesting a balance between protein flexibility and stability, with moderate flexibility near key functional regions. Moreover, the plot of the protein-ligand complex of compound 797 shows a sharp peak at 50 with RMSF values surpassing 5.4 Å. This highly flexible region is similarly observed in the plot of compound 69, which indicates less stability in the protein structure. The fluctuation ranges from 0.6 to >5.4 Å; a high peak at residue 50 might indicate interaction-induced flexibility, which could influence the binding of compound 797 with the target protein. The plot of the co-crystallized ligand showed a similar pattern with compound 797 with a high peak at residue 50; this suggests that the co-crystallized ligand has similar binding compatibility with compound 797. Additionally, the flexibility at 50 and 270 indicates these regions might be significant for ligand binding or structural dynamics.

3.9. Protein-ligand contacts

Protein-ligand contacts highlight the interactions between specific amino acid residues and ligands, revealing interactions that help stabilize the ligand within the protein's binding pocket. Fig. 9 shows the amino acid interactions of the Eg5 target protein with the top-ranked compounds (A–E) and the co-crystallized ligand (Compound F). The observed interactions include various types, such as water bridges, ionic bonds, hydrogen bonds, and hydrophobic interactions.

The histogram in Fig. 15A shows the active site amino acids (primarily GLU-116, GLY-117, GLU-118, and ARG-221) forming hydrogen bonds with Compound 103. In the initial docked pose of Compound 69, hydrogen bonds with residues THR-112, GLY-117, GLU-118, and ARG-221 were significant but shifted to include hydrophobic bonds and water bridges during the simulation. For the Eg5-Compound 888 complex, hydrogen bonding interactions with GLY-117 and GLU-118 were less prominent, persisting for less than 30 % of the simulation time. In contrast, for Compounds 177 and 797, key residues such as GLU-116, GLY-117, GLU-118, and ARG-221 were observed to interact in both complexes, though the duration of contacts varied. Notably, Compound 797 maintained interactions for more than 30 % of the simulation time, indicating stronger and more sustained binding.

There are three different kinds of hydrophobic contacts, including p-cation, pi-pi, and additional non-specific interactions. These interactions take place between an aromatic or aliphatic group on the ligand and a

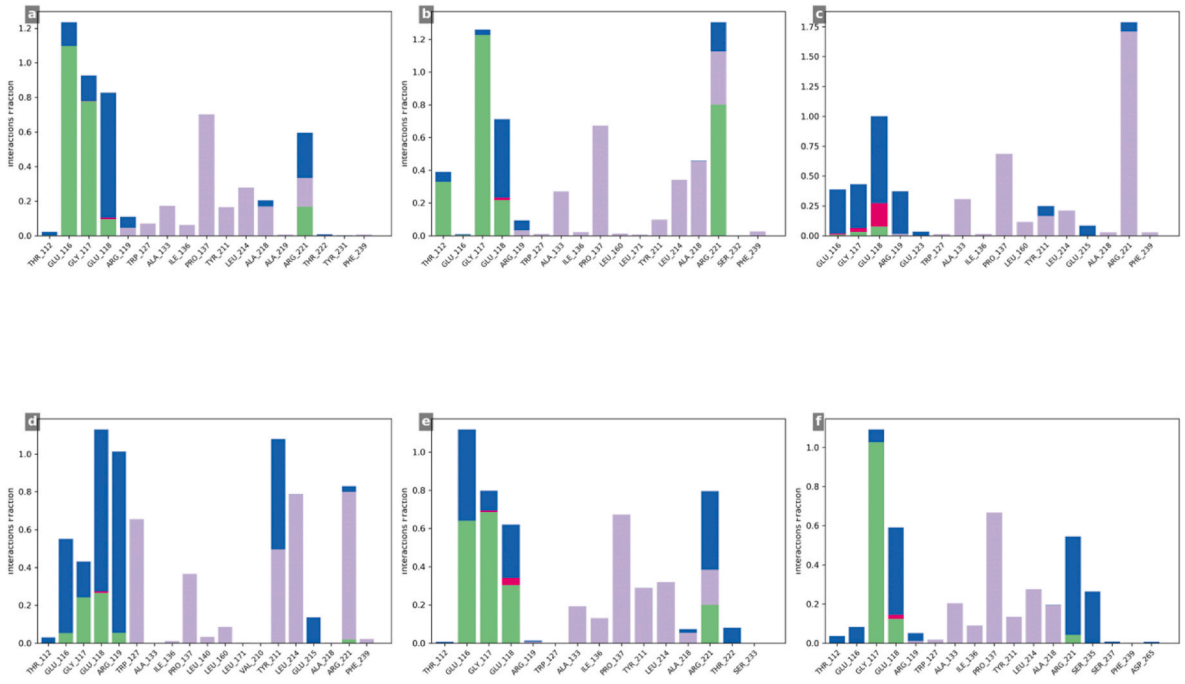


Fig. 15. Protein-ligands contacts monitored throughout 200 ns simulation run time of Eg5 complex with the lead compounds: (A) Compound 103; (B). Compound 69 (C). Compound 888 (D). Compound 177 (E). Compound 797 (F). Co-crystallized ligand.

hydrophobic amino acid [46]. Hydrophobic interactions were also observed with residues ARG-119, TRP-117, ALA-133, ILE-136, PRO-137, TYR-111, LEU-214, ALA-218, ALA-219, and ARG-221 in Eg5-Compound 888 complex. The plot for Eg5-Compound 888 complex demonstrate dominant hydrophobic interactions amongst other complexes; the prevalence of these interactions suggest that compound 888 is anchored in non-polar regions of the binding site. The ligand's persistent binding in the protein's hydrophobic pockets is facilitated by these interactions, which frequently result in higher binding affinities based on frequency. Water bridged interactions were also observed with key amino acid residues such as THR 112, GLU 116, GLY 117, GLU 118, and ARG 119 (See Supplementary Fig. 7). Water bridges provide the interaction network flexibility, enabling the ligand to change its shape. The histogram in Fig. 15 demonstrates strong ionic interactions in compound 888 and 797; this is likely due to charged functional groups in the ligands. Certain subplots such as compounds 103, 69 and 177 have little to no ionic interactions, which suggests that they are dependent on particular residues or circumstances.

4. Conclusion

In this study, we identified new and promising inhibitors of Mitotic Kinesin Eg5 using an LSTM network to support breast cancer treatment. The top compound, Compound 103, is a bioisostere of the co-crystallized ligand, a known Eg5 inhibitor, and showed good predicted binding affinity and free energy. However while LSTMs are useful, they have some limitations, especially with the development of newer methods like Transformers and state-space models (SSM). In future work, we plan to use these advanced methods, including the updated xLSTM framework, to explore chemical space further and discover even better molecules. We also acknowledge the limitation of this study, as it is fully computational. We propose further *in vitro* and *in vivo* experiments to help validate the biological relevance of the top compounds identified.

CRediT authorship contribution statement

Damilola Samuel Bodun: Writing – review & editing, Visualization, Supervision, Data curation, Writing – original draft, Validation, Investigation, Conceptualization. **Isiaka O. Ibrahim:** Writing – review & editing, Project administration, Investigation, Writing – original draft, Methodology, Conceptualization. **Mujeebat Bashiru:** Writing – original draft, Funding acquisition, Writing – review & editing, Software, Data curation. **Rachael Oluwakamiye Abolade:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization, Writing – original draft, Software, Data curation. **Wilberforce K. Ndarawit:** Writing – review & editing, Visualization, Formal analysis, Writing – original draft, Supervision, Data curation. **Adedoyin John-Joy Owolade:** Writing – original draft, Writing – review & editing, Visualization. **Adegbenro Temitope:** Writing – review & editing, Visualization, Resources, Writing – original draft, Software. **Ibeh John Somtochukwu:** Writing – review & editing, Visualization, Writing – original draft, Validation. **Ezekiel Abiola Olugbogi:** Writing – original draft, Formal analysis, Writing – review & editing, Investigation, Data curation. **Katukoliya Gamage Anuththara Samadhi:** Writing – original draft, Software, Writing – review & editing, Visualization. **Gbolahan Oduselu:** Writing – original draft, Software, Writing – review & editing, Validation. **Suliman Sharif:** Supervision, Resources, Conceptualization, Software, Project administration. **Damilola Alex Omoboyowa:** Writing – review & editing, Data curation, Writing – original draft, Conceptualization. **Olayinka O. Ajani:** Validation, Software, Supervision, Resources.

Data availability statement

The data that support the findings of this study are available from the corresponding author on request.

Ethics statement

Not applicable.

Consent

The authors have nothing to report.

Disclosure

All authors agree to be accountable for all aspects of the work.

Ethical Statement

Not applicable.

Funding

The authors received no funding for this work.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Abbreviations

MPNN	Message Passing Neural Network
LSTM	Long Short Term Memory

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2025.110672.

References

[1] H. Sung, J. Ferlay, R.L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, F. Bray, Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA Cancer J. Clin. 71 (3) (2021) 209–249, https://doi.org/10.3322/caac.21660.

[2] I. Garcia-Saez, D.A. Skoufias, Eg5 targeting agents: from new anti-mitotic based inhibitor discovery to cancer therapy and resistance, Biochem. Pharmacol. 184 (2021) 114364, https://doi.org/10.1016/j.bcp.2020.114364.

[3] N.J. Ayon, High-throughput screening of natural product and synthetic molecule libraries for antibacterial drug discovery, Metabolites 13 (5) (2023), https://doi.org/10.3390/metabo13050625. Article 5.

[4] X. Lin, X. Li, X. Lin, A review on applications of computational methods in drug screening and design, Molecules 25 (6) (2020), https://doi.org/10.3390/molecules25061375. Article 6.

[5] H. Nguyen, De Novo Molecular Design Using Deep Learning, Open Access Te Herenga Waka-Victoria University of Wellington, 2024, https://doi.org/10.26686/wgtn.25044362 [Thesis.

[6] R.R. Kotkondawar, S.R. Sutar, A.W. Kiwelekar, H.S. Wankhede, Performability of deep recurrent neural networks for molecular sequence data, Int J Comput Dig Syst 13 (1) (2023) 1317–1327.

[7] B. Zdrazil, E. Felix, F. Hunter, E.J. Manners, J. Blackshaw, S. Corbett, M. de Veij, H. Ioannidis, D.M. Lopez, J.F. Mosquera, M.P. Magarinos, N. Bosc, R. Arcila, T. Kizilören, A. Gaulton, A.P. Bento, M.F. Adasme, P. Monecke, G.A. Landrum, A. R. Leach, The ChEMBL database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods, Nucleic Acids Res. 52 (D1) (2024) D1180–D1192, https://doi.org/10.1093/nar/gkad1004.

[8] Y. Yu, X. Si, C. Hu, J. Zhang, A review of recurrent neural networks: LSTM cells and network architectures, Neural Comput. 31 (7) (2019) 1235–1270.

[9] J. Schmidhuber, S. Hochreiter, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.

[10] M.V. Santana, F.P. Silva-Jr, De novo design and bioactivity prediction of SARS-CoV-2 main protease inhibitors using recurrent neural network-based transfer learning, BMC chem. 15 (1) (2021) 8.

[11] X. Hu, G. Liu, Q. Yao, Y. Zhao, H. Zhang, Hamiltonian diversity: effectively measuring molecular diversity by shortest hamiltonian circuits, J. Cheminf. 16 (1) (2024) 94.

[12] Laurens Van der Maaten, Geoffrey Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (2008) 11.
[13] D.S. Bodun, A.A. Abdulwahab, T.I. Adegbenro, E.O. Chidinma, S.A. Osunnaya, M. O. Alonge, Z.O. Diyaolu, M.T. Abdulazeez, R.O. Abolade, Machine learning-based drug repositioning of novel human aromatase inhibitors utilizing molecular docking and molecular dynamic simulation, J. Comput. Biophys. Chem. 24 (3) (2025) 371–387.
[14] E. Heid, K.P. Greenman, Y. Chung, S.C. Li, D.E. Graff, F.H. Vermeire, C.J. McGill, Chemprop: a machine learning package for chemical property prediction, J. Chem. Inf. Model. 64 (1) (2023) 9–17.
[15] Jonathan B. Baell, J. Willem M. Nissink, Seven year itch: pan-assay interference compounds (PAINS) in 2017□ Utiliuy and Limitltions, ACS Chem. Biol. 13 (1) (2018) 36–44.
[16] 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: Chem. Enabling Drug Discover. 3 (3) (2008) 435–444.
[17] D.S. Bodun, D.A. Omoboyowa, O.I. Omotuyi, E.A. Olugbogi, T.A. Balogun, C. J. Ezeh, E.S. Omirin, QSAR-Based virtual screening of traditional Chinese medicine for the identification of mitotic kinesin Eg5 inhibitors, Comput. Biol. Chem. 104 (2023) 107865.
[18] D.A. Omoboyowa, D.S. Bodun, J.A. Saliu, Structure-based in silico investigation of antagonists of human ribonucleotide reductase from Annona muricata, Inform. Med. Unlocked 38 (2023) 101225.
[19] H.B. El-Nassan, Advances in the discovery of kinesin spindle protein (Eg5) inhibitors as antitumor agents, Eur. J. Med. Chem. 62 (2013) 614–631.
[20] H. Ruan, A comparison of long short-term memory, convolutional neural network, transformer. Mamba Models for Sentiment Analysis, Doctoral dissertation, University of Victoria, 2024.
[21] A.F. Ganai, F. Khursheed, Predicting next word using RNN and LSTM cells: stastical language modeling, in: 2019 Fifth International Conference on Image Information Processing (ICIIP), IEEE, 2019, November, pp. 469–474.
[22] J.B. Baell, Feeling nature's PAINS: natural products, natural product drugs, and pan assay interference compounds (PAINS), J. Nat. Prod. 79 (3) (2016) 616–628.
[23] E. Gilberg, D. Stumpfe, J. Bajorath, Activity profiles of analog series containing pan assay interference compounds, RSC Adv. 7 (57) (2017) 35638–35647.
[24] D. Ramírez, J. Caballero, Is it reliable to use common molecular docking methods for comparing the binding affinities of enantiomer pairs for their protein target? Int. J. Mol. Sci. 17 (4) (2016) 525.
[25] R.A. Friesner, R.B. Murphy, M.P. Repasky, L.L. Frye, J.R. Greenwood, T.A. Halgren, D.T. Mainz, Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein– ligand complexes, J. Med. Chem. 49 (21) (2006) 6177–6196.
[26] F. Godschalk, S. Genheden, P. Söderhjelm, U. Ryde, Comparison of MM/GBSA calculations based on explicit and implicit solvent simulations, Phys. Chem. Chem. Phys. 15 (20) (2013) 7731–7739.
[27] F. Humayun, F. Khan, A. Khan, A. Alshammari, J. Ji, A. Farhan, D.Q. Wei, De novo generation of dual-target ligands for the treatment of SARS-CoV-2 using deep learning, virtual screening, and molecular dynamic simulations, J. Biomol. Struct. Dyn. 42 (6) (2024) 3019–3029.
[28] M.A.A. Noor, M.M. Haq, M.A.R. Chowdhury, H. Tayara, H. Shim, K.T. Chong, In silico exploration of novel EGFR kinase mutant-selective inhibitors using a hybrid computational approach, Pharmaceuticals 17 (9) (2024) 1107, https://doi.org/10.3390/ph17091107.
[29] J.A. Arnott, S.L. Planey, The influence of lipophilicity in drug discovery and design, Expet Opin. Drug Discov. 7 (10) (2012) 863–875, https://doi.org/10.1517/17460441.2012.714363.
[30] M.J. Waring, Lipophilicity in drug discovery, Expet Opin. Drug Discov. 5 (3) (2010) 235–248, https://doi.org/10.1517/17460441003605098.
[31] 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. 64 (2012) 4–17.
[32] C.J. Ononamadu, A. Ibrahim, Molecular docking and prediction of ADME/drug-likeness properties of potentially active antidiabetic compounds isolated from aqueous-methanol extracts of Gymnema sylvestre and Combretum micranthum, BioTechnologia 102 (1) (2021) 85–99, https://doi.org/10.5114/bta.2021.103765.
[33] R. Farid, T. Day, R.A. Friesner, R.A. Pearlstein, New insights about HERG blockade obtained from protein modeling, potential energy mapping, and docking studies, Bioorg. Med. Chem. 14 (9) (2006) 3160–3173, https://doi.org/10.1016/j.bmc.2005.12.032.
[34] W. Sherman, T. Day, M.P. Jacobson, R.A. Friesner, R. Farid, Novel procedure for modeling ligand/receptor induced fit effects, J. Med. Chem. 49 (2) (2006) 534–553, https://doi.org/10.1021/jm050540c.
[35] A. Daina, O. Michielin, V. Zoete, SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules, Sci. Rep. 7 (1) (2017), https://doi.org/10.1038/srep42717.
[36] M.K. Gupta, K. Misra, Atom-based 3D-QSAR, molecular docking and molecular dynamics simulation assessment of inhibitors for thyroid hormone receptor α and β, J. Mol. Model. 20 (6) (2014) https://doi.org/10.1007/s00894-014-2286-5.
[37] C.W. Coley, D.A. Thomas III, J.A. Lummiss, J.N. Jaworski, C.P. Breen, V. Schultz, K. F. Jensen, A robotic platform for flow synthesis of organic compounds informed by AI planning, Science 365 (6453) (2019) eaax1566.
[38] A.K. Padhi, M. Janežič, K. Zhang, Molecular Dynamics Simulations: Principles, Methods, and Applications in Protein Conformational Dynamics, Elsevier EBooks, 2022, pp. 439–454, https://doi.org/10.1016/b978-0-323-90264-9.00026-x.
[39] A. Owoloye, F.C. Ligali, O.A. Enejoh, A.Z. Musa, O.O. Aina, E.T. Idowu, K. Oyebola, Molecular docking, simulation and binding free energy analysis of small molecules as PfHT1 inhibitors, PLoS One 17 (8) (2022) e0268269, https://doi.org/10.1371/journal.pone.0268269.
[40] A. Bouzina, Y.O. Bouone, O. Sekiou, M. Aissaoui, T.-S. Ouk, A. Djemel, R. Mansouri, M. Ibrahim-Ouali, Z. Bouslama, N.-E. Aouf, In vitro antitumor activity, molecular dynamics simulation, DFT study, ADME prediction, and Eg5 binding of enastron analogues, RSC Adv. 13 (28) (2023) 19567–19584, https://doi.org/10.1039/d3ra02904b.
[41] K. Liu, H. Kokubo, Exploring the stability of ligand binding modes to proteins by molecular dynamics simulations: a cross-docking study, J. Chem. Inf. Model. 57 (10) (2017) 2514–2522, https://doi.org/10.1021/acs.jcim.7b00412.
[42] A. Kumar Gupta, S. Kumar Jain, Designing and docking studies of aryl bioisosteres of enzalutamide for prostate cancer therapy, in: Futuristic Trends in Chemical Material Sciences & Nano Technology, vol. 3, Iterative International Publishers, Selfypage Developers Pvt Ltd, 2024, pp. 167–179, https://doi.org/10.58532/V3BECS12P2CH5. Book 12.
[43] K.M. Kumar, A. Anbarasu, S. Ramaiah, Molecular docking and molecular dynamics studies on β-lactamases and penicillin binding proteins, Mol. Biosyst. 10 (4) (2014) 891–900, https://doi.org/10.1039/c3mb70537d.
[44] S. Suganya, B. Nandagopal, A. Anbarasu, Natural inhibitors of HMG-CoA Reductase-An insilico approach through molecular docking and simulation studies, J. Cell. Biochem. 118 (1) (2016) 52–57, https://doi.org/10.1002/jcb.25608.
[45] J. Sharma, V.K. Bhardwaj, R. Singh, V. Rajendran, R. Purohit, S. Kumar, An in-silico evaluation of different bioactive molecules of tea for their inhibition potency against non structural protein-15 of SARS-CoV-2, Food Chem. 346 (1) (2021), https://doi.org/10.1016/j.foodchem.2020.128933, 128933–128933.
[46] S. Beura, P. Chetti, In-silico strategies for probing chloroquine based inhibitors against SARS-CoV-2, J. Biomol. Struct. Dyn. 39 (10) (2020) 1–13, https://doi.org/10.1080/07391102.2020.1772111.