scientific reports

Check for updates

OPEN

Deep learning pipeline for accelerating virtual screening in drug discovery

Fatima Noor[1,2], Muhammad Junaid[3], Atiah H. Almalki[4,5], Mohammed Almaghrabi[6], Shakira Ghazanfar[7] & Muhammad Tahir ul Qamar[2]✉

In the race to combat ever-evolving diseases, the drug discovery process often faces the hurdles of high-cost and time-consuming procedures. To tackle these challenges and enhance the efficiency of identifying new therapeutic agents, we introduce VirtuDockDL, which is a streamlined Python-based web platform utilizing deep learning for drug discovery. This pipeline employs a Graph Neural Network to analyze and predict the effectiveness of various compounds as potential drug candidates. During the validation phase, VirtuDockDL was instrumental in identifying non-covalent inhibitors against the VP35 protein of the Marburg virus, a critical target given the virus's high fatality rate and limited treatment options. Further, in benchmarking, VirtuDockDL achieved 99% accuracy, an F1 score of 0.992, and an AUC of 0.99 on the HER2 dataset, surpassing DeepChem (89% accuracy) and AutoDock Vina (82% accuracy). Compared to RosettaVS, MzDOCK, and PyRMD, VirtuDockDL outperformed them by combining both ligand- and structure-based screening with deep learning. While RosettaVS excels in accurate docking but lacks high-throughput screening, and PyRMD focuses on ligand-based methods without AI integration, VirtuDockDL offers superior predictive accuracy and full automation for large-scale datasets, making it ideal for comprehensive drug discovery workflows. These results underscore the tool's capability to identify high-affinity inhibitors accurately across various targets, including the HER2 protein for cancer therapy, TEM-1 beta-lactamase for bacterial infections, and the CYP51 enzyme for fungal infections like Candidiasis. To sum up, VirtuDockDL combines user-friendly interface design with powerful computational capabilities to facilitate rapid, cost-effective drug discovery and development. The integration of AI in drug discovery could potentially transform the landscape of pharmaceutical research, providing faster responses to global health challenges. The VirtuDockDL is available at https://github.com/FatimaNoor74/VirtuDockDL.

Keywords VirtuDockDL, Deep learning (DL), Graph neural network (GNN), Drug discovery, Python, Pipeline

Drug discovery presents a major challenge in the field of biomedical sciences. The screening attrition rate in the current drug discovery protocols suggests that one marketable drug emerges from approximately one million screened compounds[1]. The strong increase in both the number of available compounds as well as molecular targets has caused a fundamental change in the drug discovery process applied at Pharma and Biotech companies during the past two decades[2,3]. Various technologies for assay miniaturization, lab automation, and robotics enable the testing of chemical compounds in biological systems employing high-throughput screening (HTS) and ultra-high-throughput screening (uHTS)[4]. Whereas HTS is defined by the number of compounds tested to be in the range of 10,000–100,000 per day, uHTS is determined by screening numbers over 100,000 data points generated per day[5]. Taken together, the technologies of HTS and uHTS are seen as key elements for filling the drug discovery pipeline in the industry with new chemical compounds and novel modes of action.

[1]Institute of Molecular Biology and Biotechnology, The University of Lahore, Lahore 35000, Pakistan. [2]Integrative Omics and Molecular Modeling Laboratory, Department of Bioinformatics and Biotechnology, Government College University Faisalabad (GCUF), Faisalabad 38000, Pakistan. [3]Institute for Advanced Study, Shenzhen University, Shenzhen 518060, China. [4]Department of Pharmaceutical Chemistry, College of Pharmacy, Taif University, P.O. Box 11099, 21944 Taif, Saudi Arabia. [5]Addiction and Neuroscience Research Unit, College of Pharmacy, Taif University, Al-Hawiah, 21944 Taif, Saudi Arabia. [6]Pharmacognosy and Pharmaceutical Chemistry Department, Faculty of Pharmacy, Taibah University, 30001 Al Madinah Al Munawarah, Saudi Arabia. [7]National Institute for Genomics and Advanced Biotechnology (NIGAB), National Agricultural Research Center (NARC), Islamabad, Pakistan. ✉email: tahirulqamar@gcuf.edu.pk

Despite that, these high-throughput approaches are facing challenges including limited technological advancements and knowledge to identify novel drug targets, lack of drugs, and poor quality of databases in terms of data management[6,7]. To bridge the gap, computational algorithms must be exploited to overcome the limitations of drug discovery. Machine Learning (ML) approaches are becoming popular across all facets of science. Fundamentally, ML is the practice of using algorithms to parse data, learn from it, and then accordingly make a prediction about the future state of any new dataset[8,9]. Advancements in ML have paved the road to the discovery of synergistic drugs[10]. Big data and cutting-edge algorithms are now readily available, which has increased interest and led to substantial advancements in the application of ML in drug discovery[11].

Previous studies provide a realistic picture of the successful implementation of ML in drug discovery pipelines. For example, Machado et al.[12] utilized ML algorithms for screening natural inhibitors against HIV-1 integrase. Similarly, Zhou et al.[13] developed HIV-1 integrase inhibitors using machine learning techniques, and Sun et al.[14] employed the support vector machine (SVM) model for optimizing hyperparameters in virtual screening. Opportunities to apply ML occur in all stages of drug discovery. However, the challenges of applying ML lie primarily with the lack of interpretability and repeatability of ML-generated results, which may limit their application. Regarding this, Deep learning (DL) bears promise for drug discovery, including advanced image analysis, prediction of molecular structure and function, and automated generation of innovative chemical entities with bespoke properties. DL methods have also gained huge popularity due to their ability to extract features from raw data instead of relying on hand-crafted features (required in traditional ML). The ability of DL methods to extract features and analyze data simultaneously is proving them as real problem-solving methods[15,16]. Therefore there is an urgent need to develop automated DL-based algorithms for screening novel drug-like candidates against diseases and disorders.

Following that, the current study introduced VirtuDockDL, an automated DL-based pipeline, for streamlining the process of drug discovery. By integrating molecular graph construction, Graph Neural Network (GNN) modeling, virtual screening, and compound clustering, VirtuDockDL provides a comprehensive framework to automate the drug discovery process. At its foundation, the GNN model, featuring multiple custom layers, employs molecular structure (descriptors and fingerprints) data to predict the drug potential of compounds. After identification, the selected compounds were docked with proteins of interest for predicting their binding affinity. Thus, VirtuDockDL marks a pioneering effort in merging deep learning strategies with scientific computation to tackle intricate challenges in drug discovery and molecular dynamics. By streamlining and enhancing the analysis of molecular interactions and properties, our solution facilitates the more efficient identification of drug candidates and enhances the understanding of biological mechanisms. Our methodology showcases the potential of combining AI with scientific research, opening up new avenues for innovation and exploration in the pharmaceutical and biotechnological sectors.

Methods

Architecture of VirtuDockDL

Molecular data processing

Molecular data in this application is represented using SMILES strings, which are processed and transformed into graph structures using the RDKit library. Once molecules are represented as graphs, they can be used as input for machine learning models such as the Graph Neural Network (GNN). This transformation is critical, as graph representations allow the model to capture the structural relationships within molecules, enabling it to make accurate predictions about molecular properties. The application uses PyTorch Geometric to build and train GNN models. These models process molecular graphs and learn patterns in the data that relate to properties such as molecular activity or binding affinity. The data is handled by custom dataset classes like $MoleculeDataset$, which ensures that the molecular data is structured correctly for model training and evaluation.

Graph neural network (GNN) model and feature extraction

A state-of-the-art GNN is exploited to predict the biological activity of compounds based on their corresponding structural data, which are encoded in the form of a molecular graph. Indeed, the architecture has been specifically designed to handle such graphs, including several layers of computation that are crucial in capturing the highly complex, hierarchical structure of a molecule.

GNN model architecture The core of the GNN model is built on specialized GNN layers, particularly designed to process molecular graphs through graph convolution operations[17]. At the heart of each GNN Layer, a sequence of computational steps unfolds, starting with a linear transformation of node features h_v, parameterized by weights W, to yield transformed features $h'_v = W \cdot h_v$. These features undergo batch normalization, a critical step for stabilizing the learning process and enhancing convergence rates, defined mathematically as,

$$\widehat{x} = \frac{x - \mu_\beta}{\sqrt{\sigma_\beta^2 + \in}}$$

where μ_β and σ_β^2 denote the batch mean and variance, respectively, and$\in$ is a small constant to prevent division by zero. To introduce non-linearity, a $ReLU$ (Rectified Linear Unit) activation function is applied to each batch-normalized feature, formulated as:

$$h''_v = \max(0, \widehat{h}'v)$$

A distinctive feature of our architecture is the incorporation of residual connections for layers with matching input and output dimensions, enabling the model to effectively learn identity mappings and mitigate the vanishing gradient problem. The final output of such layers, h'''_v, is the sum of the input and activated features,$h'''_v = h_v + h''_v$, ensuring a smooth gradient flow essential for deep learning. Furthermore, to combat overfitting, a dropout mechanism is employed, randomly deactivating a subset of features during training with a probability *p*, denoted as $h''''_v = \mathrm{Dropout}(h'''_v, p)$.

Complementing the graph-based feature processing, the model adeptly merges graph-derived features with additional molecular descriptors and fingerprints, encapsulating both structural and physicochemical molecule aspects. This fusion is achieved by concatenating the aggregated graph features h_{agg} with engineered features f_{eng}, subsequently passing through a fully connected layer to materialize the combined feature vector,

$$f_{combined} = ReLU\left(W_{combine} \cdot [h_{agg}; f_{eng}]\, b_{combine}\right),$$

where $[;]$ signifies concatenation. This holistic approach, underpinned by rigorous mathematical formalisms, accentuates the model's capacity to distill intricate molecular structures and properties into actionable insights, offering an advanced tool for predicting chemical compounds' biological activities with enhanced precision and reliability.

Feature extraction The feature extraction journey commences with the transformation of SMILES strings into their molecular graph counterparts, a pivotal step conducted using RDKit[18]. This transformation entails identifying atoms and bonds within a molecule, represented as nodes $v$and edges e, respectively, within the graph. Formally, a molecular graph $G$is constructed such that

$$G = (V, E)$$

where $V$is the set of nodes corresponding to atoms, and $E$is the set of edges representing bonds. These graphs are then inputted into our GNN model, which utilizes the rich topological information to learn the complex patterns underlying molecular structures.

Beyond the graph-based features, our methodology extends to the extraction of molecular descriptors and fingerprints, vital for capturing the physicochemical essence of the compounds. For instance, the molecular weight (MolWt) can be mathematically expressed as $\mathrm{MolWt} = \sum_{i=1}^{n} m_i$, where $m_i$is the atomic mass of the i-th atom, and n is the total number of atoms in the molecule. Similarly, topological polar surface area (TPSA) and the octanol-water partition coefficient (MolLogP) are quantified through established computational chemistry methods, providing insights into the molecule's bioavailability and reactivity. In addition to these features, the Number of Hydrogen Donors, Number of Hydrogen Acceptors, LogP, and Number of Rotatable Bonds were also calculated. These features are essential for machine learning models as they provide critical information about the molecular properties that can be used to predict various biological and chemical behaviors. Furthermore, molecular fingerprints, such as Morgan fingerprints and MACCS keys, are binary vectors F where each element $F_i$represents the presence (1) or absence (0) of specific molecular features or substructures, enabling the model to recognize and utilize molecular patterns for accurate activity prediction.

This integrative approach of combining graph-based learning with traditional cheminformatics features, is summarized as, $Features = Graphs\,(G) + Descriptors\,(D) + Fingerprints\,(F)$, empowers our GNN model to achieve a comprehensive understanding of molecular structures and their properties. By leveraging this sophisticated feature extraction and modeling framework, our model boasts high accuracy in predicting the biological activities of chemical compounds. This not only underscores the efficacy of merging graph-theoretical approaches with cheminformatics but also highlights the model's utility in advancing drug discovery and chemical research, paving the way for novel insights into the molecular determinants of biological activity.

GNN Model implementation and evaluation

The study uses GNN with one input layer, 64 hidden layers, and two output layers to predict the biological activity of compounds. The classification-based GNN model uses CrossEntropyLoss, optimized with RMSprop with a learning rate of 0.001 and a weight decay value of $5e^{-4}$. The learning rate was further adjusted by using $ReduceLROnPlateauscheduler$ to improve the generalization. After hypermeter tuning that dataset was divided into a training and test set using a stratified shuffle split to ensure all class activities are equally distributed. This ensures that the model is evaluated and can be generalized into new unseen data.

In evaluating the model's efficacy, the current study employs a range of evaluation metrics such as Accuracy, Precision, Recall, and F1 Score. These metrics collectively provide a comprehensive assessment of the model's classification accuracy, enabling us to refine its predictive capabilities. Additionally, the Area Under the Receiver Operating Characteristic (AUC-ROC) curve serves as a crucial metric, offering insights into the model's capacity to distinguish between classes with a nuanced understanding of true and false positive rates. To mitigate overfitting and improve model training, we incorporate early stopping, a technique that halts training when the validation loss stops decreasing, thereby preserving the model's generalizability. In addition, a learning rate scheduler dynamically adjusts the learning rate based on validation loss trends, optimizing the training process for optimal performance. After model implementation, we used a variety of evaluation metrics to evaluate the model's effectiveness, including Accuracy, Precision, Recall, and F1 Score. Furthermore, the AUC-ROC curve, measures the trade-off between true positive rate and false positive rate[19].

Virtual screening and clustering
Virtual screening is another core feature of the application, allowing users to screen large libraries of molecules for potential drug candidates. The application allows users to upload new libraries of molecules, which are then processed by the trained GNN model to predict their activity or binding affinity. Molecules with high predicted activity are retained for further evaluation, allowing researchers to quickly identify promising candidates. Further clustering is used to group molecules based on their predicted properties.

Clustering using Gaussian mixture model (GMM)

After model implementation, the clustering of compounds was performed based on their predicted probabilities by using the GMM[20], formulated as $P(y = 1|x)$ ($y = 1$ represents the active class). Mathematically, the probability of a compound x given the model parameters θ is expressed as:

$$p\left(x|0\right) = \sum\nolimits_{i=1}^{k} \pi_i N\left(x|\,\mu_i, \sum i\right)$$

Here, k is the number of clusters, $\pi_i$is the mixing coefficient for cluster i, and $N\left(x|\,\pi_i, \sum i\right)$ is the Gaussian distribution with mean $\mu_i$and covariance $\sum i$. The model parameters are optimized to best fit the distribution of the data.

To assess the quality of the clusters formed by the GMM, we compute the Silhouette Score[21] and the Davies-Bouldin Score[22]. The Silhouette Score for a single sample is defined as:

$$S = \frac{b - a}{\max(a, b)}$$

where$a$ is the mean distance between a sample and all other points in the same class, and b is the mean distance between a sample and all other points in the next nearest cluster. The Silhouette Score ranges from − 1 to 1, where a high value indicates that the sample is well matched to its own cluster and poorly matched to neighboring clusters. The Davies-Bouldin Score is calculated as:

$$DB = \frac{1}{K}\sum\nolimits_{i=1}^{k} max_{j\neq i}\left(\frac{\sigma_i + \sigma_j}{d(c_i + c_j)}\right)$$

where k is the number of clusters, σ_i is the average distance of all points in cluster i to the cluster centroid $\mathrm{c_i}$ and $\mathrm{d(c_i + c_j)}$ is the distance between centroids $\mathrm{c_i}$and $\mathrm{c_j}$.A lower Davies-Bouldin Score indicates better clustering.

Protein refinement and docking

The application also provides functionality for protein structure refinement and ligand docking, critical steps in molecular modeling and drug discovery workflows. Protein refinement is handled by the $\mathrm{perform_protein_refinement}\,()$function, which uses OpenMM to perform energy minimization and structural corrections on protein models. This step is essential for preparing protein structures for accurate docking simulations. The refined protein structures are saved in PDB format, and structural analysis tools are used to generate visual outputs, such as Ramachandran plots and SASA (Solvent Accessible Surface Area) plots, which give insight into the quality and properties of the protein structure.

Ligand docking is carried out using AutoDock Vina, a popular tool for predicting how small molecules (ligands) bind to proteins. The $\mathrm{run_docking}\,()$function orchestrates the docking process, converting ligand files into the PDBQT format required by AutoDock Vina and running the docking simulations. The results, including binding affinities and RMSD (Root Mean Square Deviation) values, are saved for further analysis. Users can then download the docking results in various formats, including CSV files and PDBQT files.

Graphical user interface (GUI) of VirtuDockDL

VirtuDockDL is developed using Python and the Flask web framework[23]. This application integrates cutting-edge computational tools such as RDKit[24] for molecular manipulation, OpenMM[25] for protein refinement, and PyTorch[26] Geometric for training GNNs to predict molecular properties. Additionally, AutoDock Vina[27] was employed for ligand docking simulations, enabling efficient virtual screening of large molecular libraries for drug discovery. Through Flask, VirtuDockDL offers a user-friendly interface that simplifies tasks like file uploads, data preprocessing, model training, and result visualization, creating a smooth and effortless experience for researchers and drug developers. Figure 1 shows the workflow of the VirtuDockDL pipeline, which integrates de-novo molecule generation, feature selection, graph neural networks, molecular docking, and benchmarking for virtual drug screening.

Case study and benchmarking

The current study used the VP35 protein of the Marburg virus as a case study to identify and predict compounds with potential drug-like efficacy against MARV disease. The VP35 is a multifunctional target protein of Marburgvirus, which has a key role to play in the not only plays a crucial role in viral RNA synthesis, assembly, and structural integrity VP35 protein binds double-stranded RNA and inhibits alpha/beta interferon production induced by RIG-I signaling. To date, there are no validated small molecule VP35 inhibitors. Thus, it is the need of the hour to develop some novel therapeutic strategies for MARV disease. To bridge this gap, current study also focused to identhe tify novel inhibits against the VP35 protein. Regarding this, the positive dataset (active molecules) was initially prepared by selecting only active molecules of protein of interest from a literature review

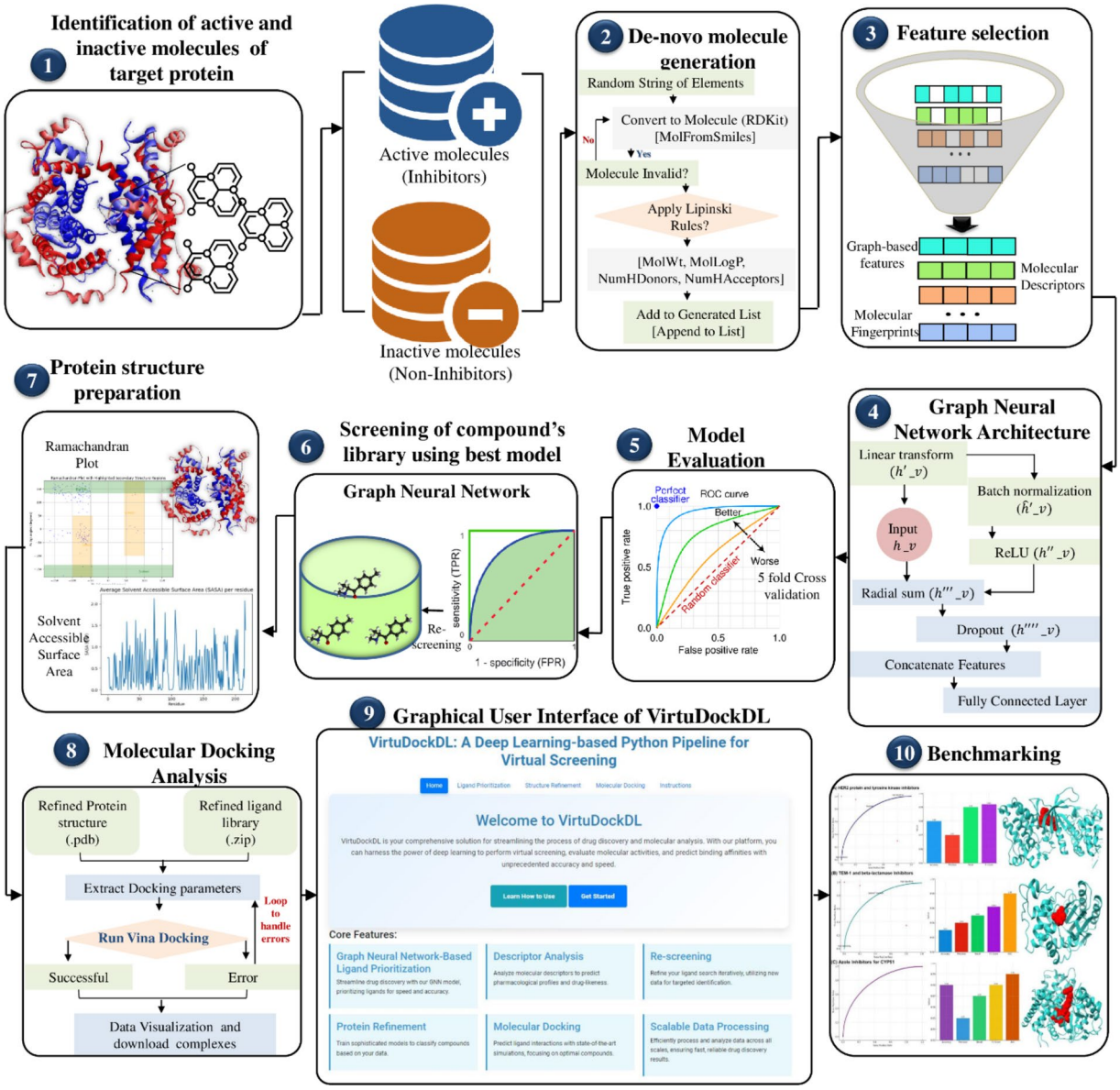

Fig. 1. Graphical synopsis illustrates the workflow of the VirtuDockDL pipeline for virtual screening in drug discovery. It begins with identifying active and inactive molecules for a target protein. De-novo molecules are generated, filtered by drug-likeness rules, and their features are selected based on graph-based features, molecular descriptors, and fingerprints. GNN model is trained and evaluated using metrics like ROC curves. The best model is used to screen a compound library for potential inhibitors. Protein structures are prepared and refined for molecular docking simulations. The results are visualized and benchmarked against experimental data. The VirtuDockDL platform provides a user interface to manage all these steps efficiently.

and databases like BindingDB[28], PubChem[29] and ChEMBL[30]. The complementary negative dataset (inactive/decoy) was constructed using de-novo generation function of VirtuaDockDL as well as from public repositories like DUDE database[31]. The de-novo generation function of VirtuDockDL synthesizes novel molecular structures by randomly assembling chemical elements into SMILES strings, which are then validated using RDKit's 'Chem.MolFromSmiles'. Decoys, by design, are structurally plausible but inactive molecules with respect to the biological target. They do not exhibit any binding affinity to the target, making them critical for training DL models. These decoys help the model learn to differentiate between active and inactive compounds, ultimately enhancing its ability to accurately predict true actives. Tay et al.[32] employed same approach, for filtering out 9,596,585 decoy molecules from a set of 100 million generated structures. This demonstrates the effectiveness of generating diverse decoys to improve model robustness, as employed in the current study. Lastly, the positive and negative datasets were then merged to generate a final dataset for further DL model implementation.

In the next phase of the study, the automated backend of VirtuDockDL employed a GNN model to predict the activity of molecules against the VP35 protein of the Marburg virus. The model utilized Graph Convolutional Layers (GCNConv) to process molecular graphs generated from SMILES strings, allowing it to learn intricate

structural and chemical patterns within the dataset. Active and decoy molecules were automatically split into training and testing sets, and the GNN was trained with Cross-Entropy Loss and RMSprop optimization for accurate classification of active and inactive compounds.

The backend further automated feature extraction using RDKit, calculating molecular descriptors such as molecular weight, logP, and topological polar surface area. The model's performance was evaluated through metrics like accuracy, precision, recall, F1-score, and AUC (Area Under Curve), giving a clear assessment of its predictive accuracy. Once training was complete, protein refinement was automated using OpenMM for energy minimization of the VP35 protein structure. The refined structure was then used for ligand docking simulations in AutoDock Vina, where active molecules identified by the GNN model were docked against the protein's binding site. Docking scores and binding affinities were calculated to identify potential VP35 inhibitors. All results, including molecular structures, docking poses, and cluster plots, were automatically visualized and stored for further analysis.

Results and discussion

GUI of VirtuDockDL

At the core of the application is the Flask web framework, which facilitates interactions between users and the system. Users access the system through a web interface, where they can upload molecular files, initiate tasks, and download results. The graphical user interface of VirtuDockDL exemplifies a user-centric design, offering an intuitive and efficient workflow for the complex process of virtual screening in drug discovery (Fig. 2). Users are welcomed by a clean and organized layout, with primary functions segmented into dedicated tabs including Ligand Prioritization, Structure Refinement, and Molecular Docking. For example, the CSV upload tabs allow for easy submission of molecular data, while maintaining an uncluttered aesthetic that promotes focus and ease of use. Under the Rescreening page, uploading and analyzing compound SMILES notations is made easier with a streamlined upload interface, a 'Analysis' button, and an tabular display of the compound clustering along with probability value. The results tabs also provide Ramachandran and SASA plots provide independent methods to evaluate the conformational quality of protein structures. The next tab, Molecular Docking will help the users to tune the docking process through customizable parameter set and docking results will be shown in an easy-to-understand diagram. Navigating to the Docking tab, users are able to customize the docking procedure using an adjustable parameter set, and the result is displayed in form of bar plots, demonstrating the binding affinity of proteins and ligands. This design philosophy embodies the concept of 'aesthetic simplicity with functional sophistication' in VirtuDockDL, making it a friendly interface for the complex task of virtual screening with deep learning.

GNN model implementation

Initially, active and inactive molecules of VP35 protein were collected using BindingDB and DUDE databases. The denovo molecule generation function of VirtuDockDL was also employed to generate new molecules. Subsequently, feature extraction was carried out on the compiled dataset to facilitate the differentiation between active and inactive compounds (Supplementary Data 1, Fig. 3). The dataset was then divided into training and test sets by 80:20 ratio, enabling the application of machine learning models to predict the drug-like potential of compounds against MARV disease (Supplementary Data 2). This systematic approach combines the utilization of bioinformatics resources and machine learning techniques to identify promising candidates for further experimental investigation.

Model evaluation

The performance of the VP35 protein–focused predictive model was evaluated using different performance measures. GNN model achieved the test Accuracy of 0.9779, reflecting a high level of reliability in the model's predictions. The Precision value was recorded to be 0.9688, indicating the accuracy of model in identifying true positive instances. In addition, the Recall and F1 Score was reached to value of 0.9841 and 0.9764 respectively, indicating a balanced performance of the model. Lastly, the AUC score was found to be 0.9972, demonstrating outstanding capability of the model to discriminate various classes. Furthermore, cluster analysis revealed Silhouette Score of 0.9355 and a Davies-Bouldin Score of 0.0201. These scores implied that the clusters were well defined and the model could successfully group compounds that show similar drug like properties.

Rescreening with library of non-covalent inhibitors

The trained model was used to re-screen a library of non-covalent inhibitors. The non-covalent inhibitors have high selectivity and reduce possible toxicity and off-target effects, which are pivotal for targeting viral proteins. The reversible binding of non-covalent inhibitors enables for controlled modulation of viral activity. A library of non-covalent inhibitors were retrieved from ZINC and PubChem databases and rescreened using trained GNN model for analyzing their drug-like potential. A total of 146 molecules were found to be active and have drug like potential against target proteins. Finally, the folder containing the compounds in SDF form along with their official name was downloaded in zip file format. The folder were then further considered for molecular docking analysis.

Protein structure refinement and molecular docking analysis

The 3D structure of VP35 protein (PDB ID: 4GH9) was downloaded from RCSB PDB. The structure was uploaded to the structure refinement tab of ViruDockDL. All the solvent and ligand molecules were initially removed from the 3D structure. Later, the missing atoms in amino acid residues, incorrect bond orders, and other structural anomalies were corrected. After that the structure was energy minimized by adjusting the

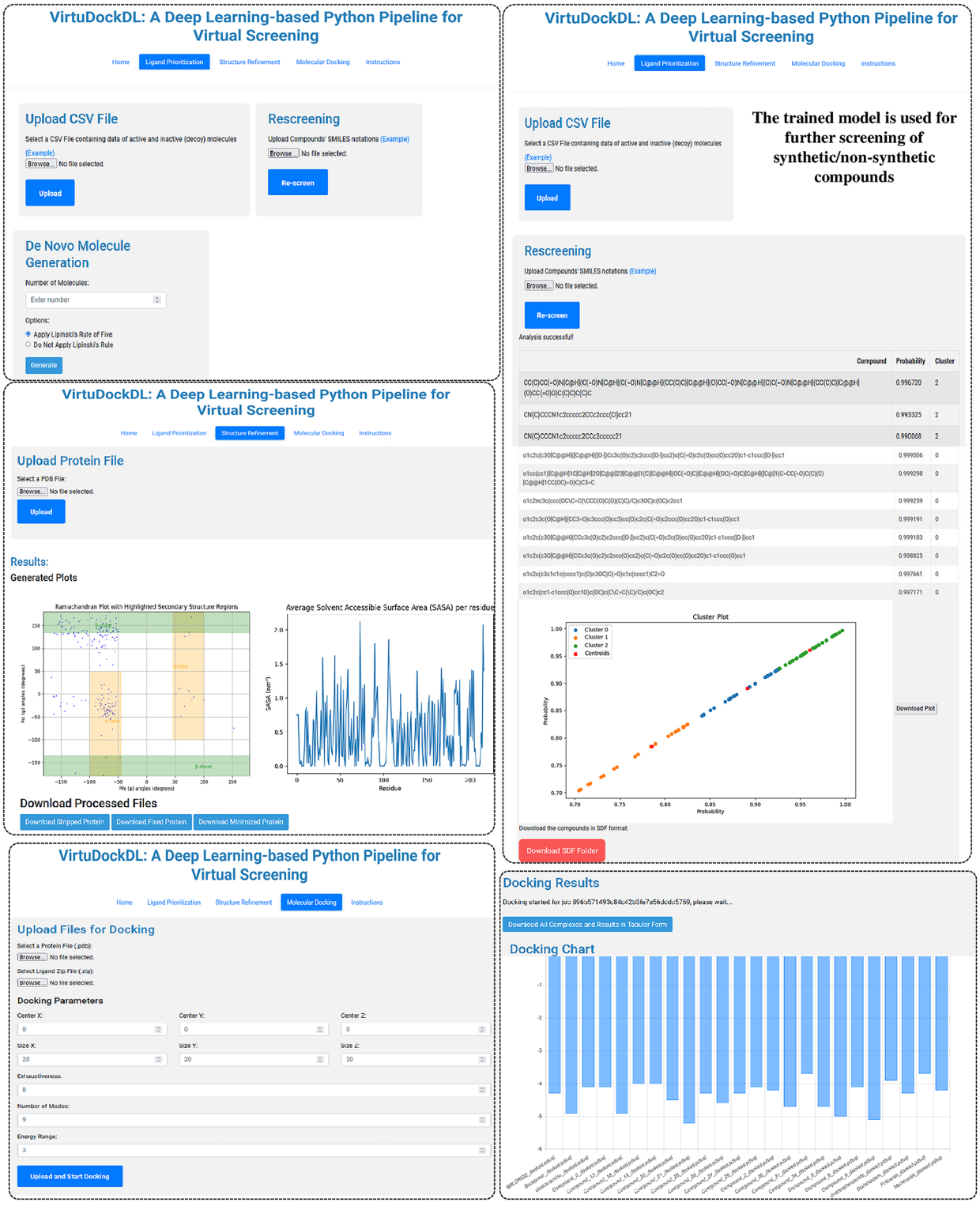

Fig. 2. Graphical user interface of VirtuDockDL.

positions of atoms in the protein structure. The final refined protein structure was later downloaded using the download option available in VirtuDockDL (Fig. 4).

Lastly, the molecular docking analysis was performed among non-covalent inhibitors and refined VP35 protein in order to predict binding affinity between protein and non-covalent inhibitors. The molecular docking analysis revealed that the selected inhibitors has highest binding affintity with VP35 protein. Clomipramine and Mrizomib were particularly to have highest binding affinity with VP35 protein (Fig. 5). Previous studies provide shred of shreds of evidence that Clomipramine and marizomib play distinct roles in combating viral infections.

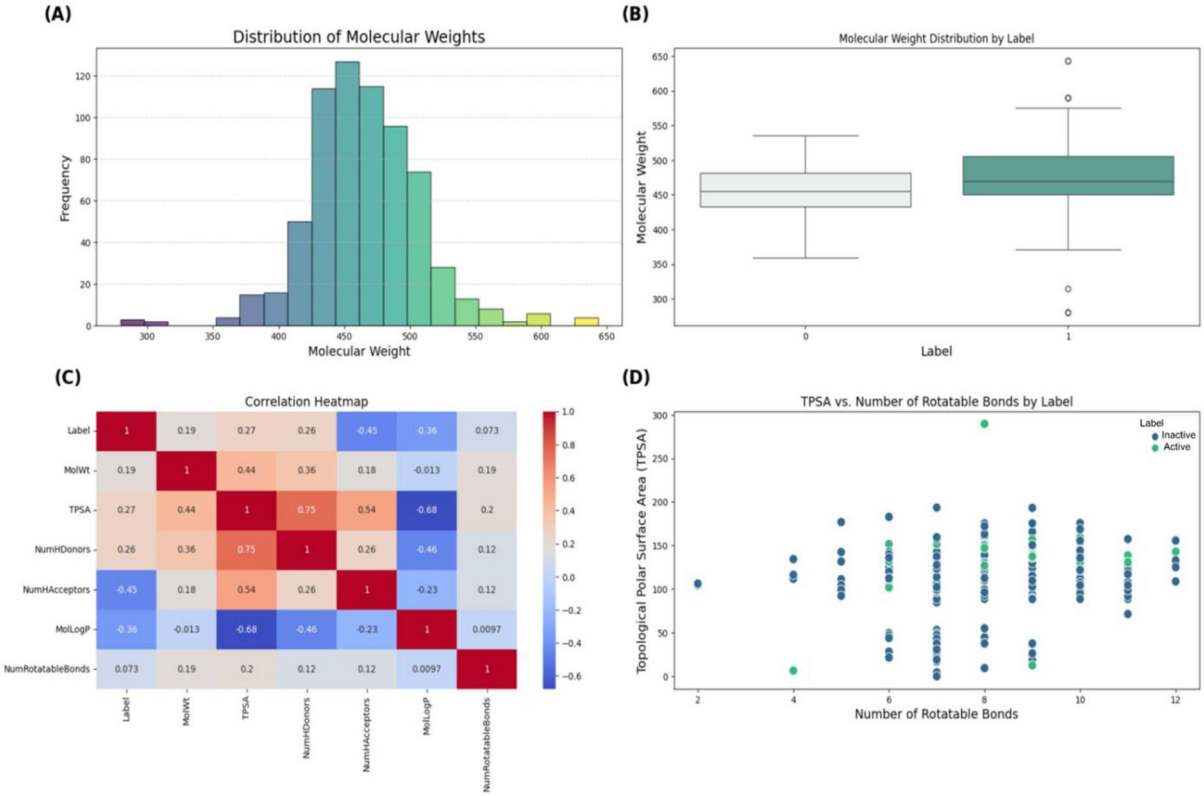

Fig. 3. Comprehensive visualization of molecular data characteristics. (**A**) Distribution of molecular weights across the dataset, highlighted with a gradient color scheme to enhance visibility. (**B**) Boxplot illustrating the variance and distribution of molecular weights segmented by label, facilitating comparison between different groups. (**C**) Heatmap of the correlation matrix displaying the relationships between numerical features within the data, annotated for clarity. (**D**) Scatter plot showing the relationship between the number of rotatable bonds and the topological polar surface area (TPSA) with data points colored by label, providing insights into molecular flexibility and polar surface characteristics.

Clomipramine has been shown to suppress ACE2-mediated SARS-CoV-2 entry, potentially hindering viral infection. On the other hand, marizomib, a proteasome inhibitor, exhibits a prolonged and broader proteasome inhibition profile compared to bortezomib, offering the potential to target viral proteins through a unique mechanism.

Benchmarking

VirtuDockDL was compared with other popular virtual screening tools like PyRMD[33], RosettaVS[34], and MzDOCK[35], each offering different capabilities in drug discovery. PyRMD is an AI-powered, ligand-based screening tool but lacks structure-based docking and protein-ligand interaction capabilities. In contrast, VirtuDockDL combines both ligand-based and structure-based screening with deep learning via GNNs, enabling accurate predictions of biological activity and docking outcomes. RosettaVS excels at highly accurate structure-based docking and binding affinity predictions but is not designed for large-scale compound screening. VirtuDockDL, on the other hand, automates the entire virtual screening process, including docking and clustering, making it more efficient for high-throughput applications. MzDOCK offers a user-friendly GUI for molecular docking, but it relies on traditional docking algorithms without AI integration, limiting its predictive accuracy compared to VirtuDockDL's deep learning approach. VirtuDockDL's AI-driven automation, combining GNNs with traditional docking, allows for faster and more accurate screening of large datasets, making it a more powerful and versatile tool than PyRMD, RosettaVS, or MzDOCK. Its ability to handle both ligand- and structure-based virtual screening in a fully automated pipeline positions it as an ideal solution for modern drug discovery.

The performance of VirtuDockDL was further compared against both machine learning-based virtual screening tools and traditional docking software. Regarding this, we compared VirtuDockDL to DeepChem[36], Chemprop[37], and MolDQN[38], which are commonly used for ligand-based virtual screening but lack integration with structure-based docking. On a dataset of HER2 protein-ligand interactions (PDB ID: 3PP0), VirtuDockDL achieved superior performance, with a test accuracy of 99%, an F1 score of 0.992, and an AUC of 0.99. In comparison, DeepChem achieved an accuracy of 89%, with an F1 score of 0.89 and an AUC of 0.90, illustrating the enhanced predictive power of VirtuDockDL's GNN-based approach.

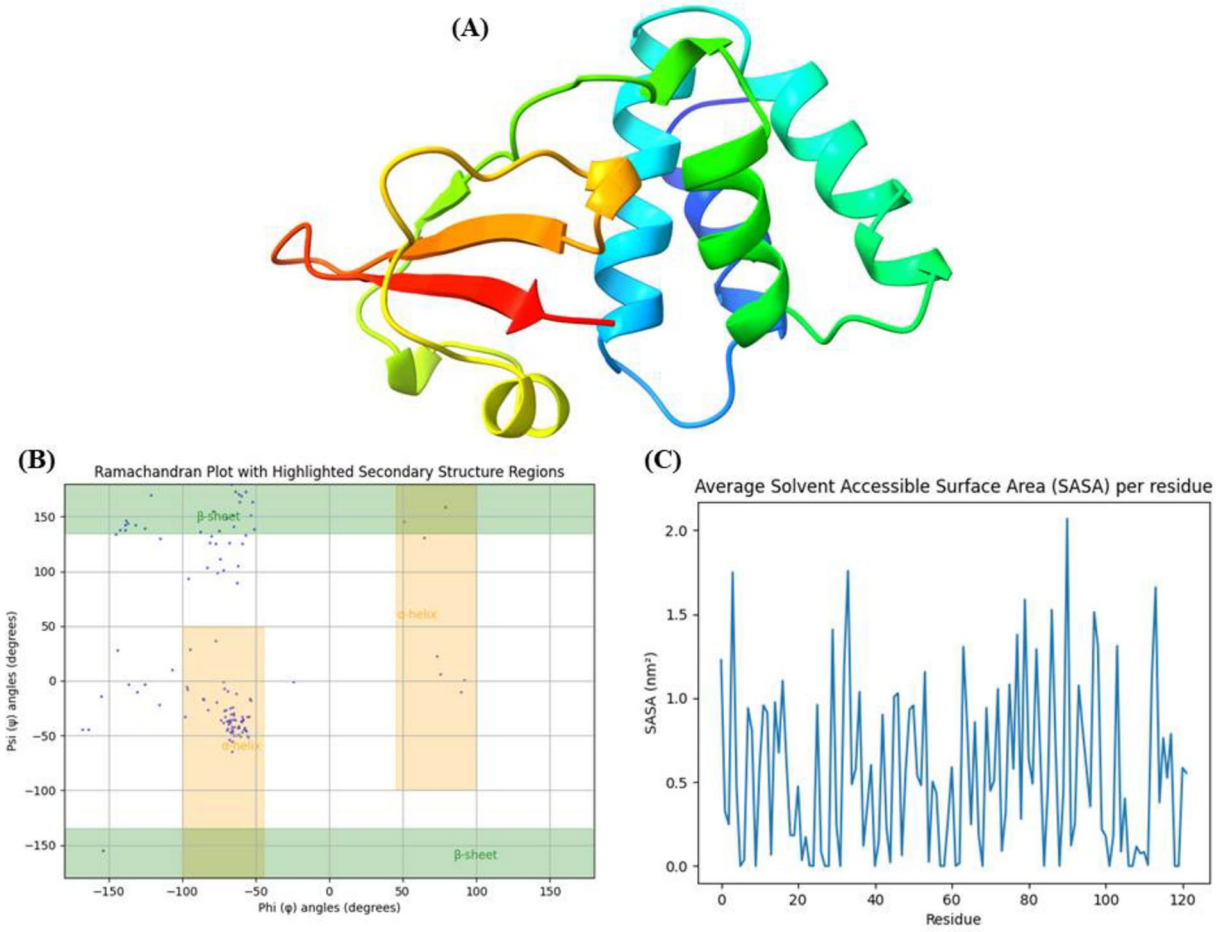

Fig. 4. (**A**) Three-dimensional structure of the VP35 protein of the Marburg virus, showing the arrangement of secondary structures. (**B**) Ramachandran plot indicating the phi (φ) and psi (ψ) angles of amino acid residues, with regions corresponding to α-helices and β-sheets highlighted. (**C**) Graph of the Average Solvent Accessible Surface Area (SASA) per residue, indicating the extent of protein surface exposure to solvent molecules.

We also benchmarked VirtuDockDL against AutoDock Vina and Glide, widely used docking tools. When tested on the HER2 dataset, VirtuDockDL outperformed AutoDock Vina, which predicted binding affinities with an accuracy of 82% and an F1 score of 0.84, compared to VirtuDockDL's 99% accuracy. Glide, while more accurate in docking small datasets, struggled with larger datasets, whereas VirtuDockDL maintained high performance, demonstrating binding affinity predictions for key inhibitors, including Dasatinib (− 7.12 kcal/mol) and Nilotinib (− 6.89 kcal/mol).

Additionally, TEM-1 beta-lactamase (PDB ID: *Escherichia coli* infection treatment strategies (1xpb) was targeted for *E. coli* (Fig. 6). After the training process of the GNN model, the Test Accuracy achieved was 0.93. This accuracy when compounded with a Precision of 0.94 and Recall of 0. 95, captures the ability of the model to correctly predicting active inhibitors. The F1 Score was at 0. 962, and an AUC of 0.98, supports overall validity of the model and confirms its ability to separate between active and inactive compounds rather successfully. Further structure preparation and docking analysis of a large set of BLI molecules revealed Durlobactam, ETX2514, Taniborbactam, Sulbactam, and BLI-489 as the strongest binders. These results emphasize the accuracy of VirtuDockDL's docking pipeline and its applicability for rapidly advancing the search for new beta-lactamase inhibitors.

Regarding Candidiasis (fungal infection), the GNN model was trained and showcased a high accuracy in segmenting active inhibitors with a test accuracy of 0. The model also obtained an accuracy of 0. 95 and a recall of 0. As shown in score 97, the model has a high level of accuracy in identifying the right active compounds. The computed F1 score is 0. 98 and an AUC of 0. 99 also gives more support to the model and its ability to accurately classify between active and inactive compounds. After the model training, structure preparation, and docking analysis were carried out on a large database of azole antifungal agents using VirtuDockDL docking pipeline. From the compounds, Bifonazole, Butaconazole, Penconazole, Tebuconazole, and Epoxiconazole were found to have the best binding affinities to the CYP51 enzyme. These observations demonstrate the effectiveness of the VirtuDockDL pipeline and the possibility of reducing the time required to identify new azole inhibitors.

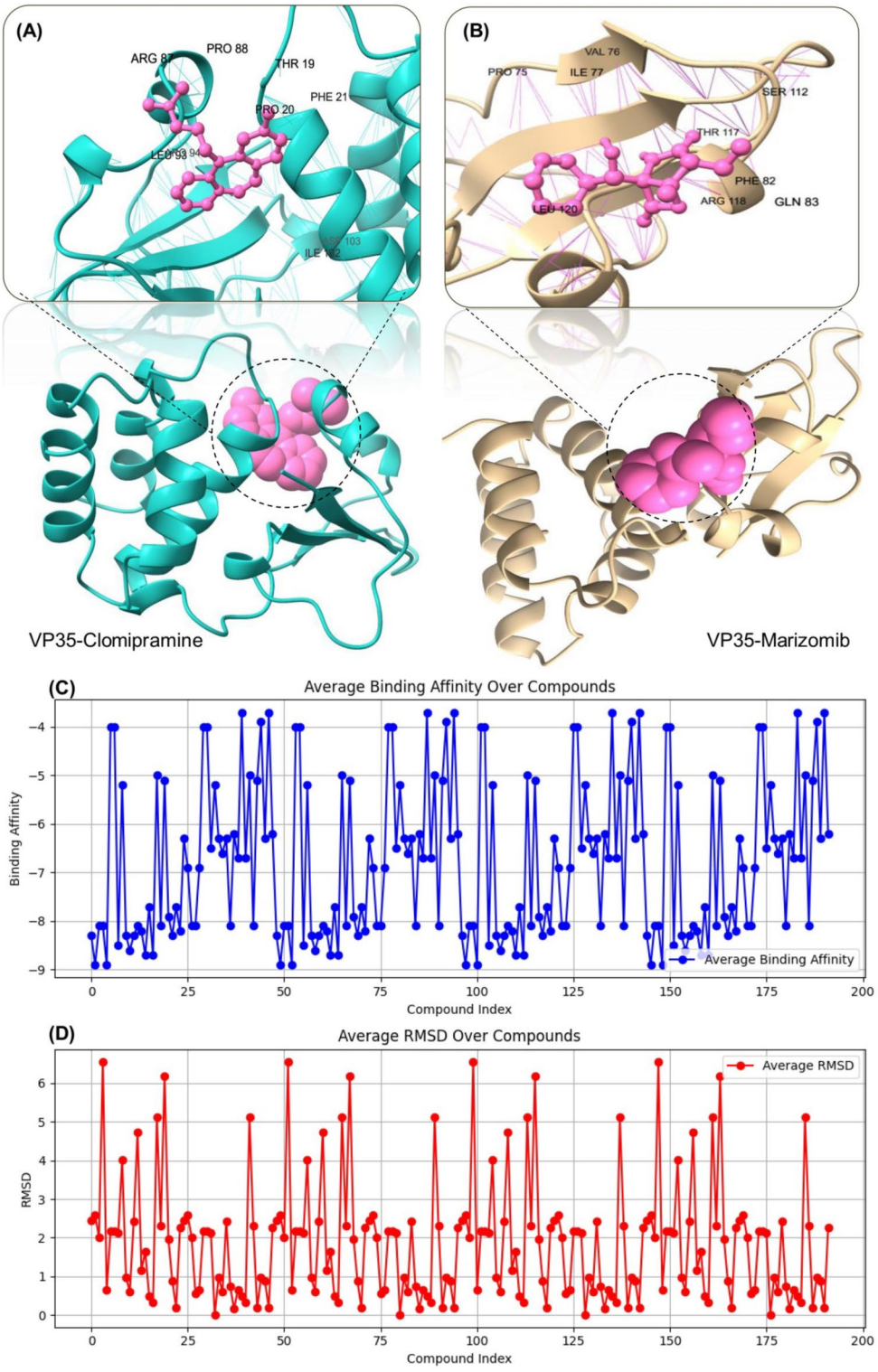

Fig. 5. (**A**) Illustrates a molecular docking pose of the compound Clomipramine (depicted in pink spheres) within a binding site of the target protein VP35 (rendered in cyan ribbon diagram). Key interacting residues are highlighted and labeled. (**B**) Shows a similar docking pose for the compound Marizomib (again in pink spheres) within a binding pocket of the protein VP35 (this time in a wheat-colored ribbon diagram). Specific interactions with residues are detailed. (**C**) A line graph represents the average binding affinity of a series of compounds, indexed numerically along the x-axis. The data points are marked in blue, with lines extending vertically to indicate the variance or range of binding affinity measured for each compound. (**D**) Another line graph displays the average RMSD (Root Mean Square Deviation) for the same series of compounds. Each compound's RMSD value is marked in red, with vertical lines illustrating the variability or precision of the docking pose in relation to a reference pose.

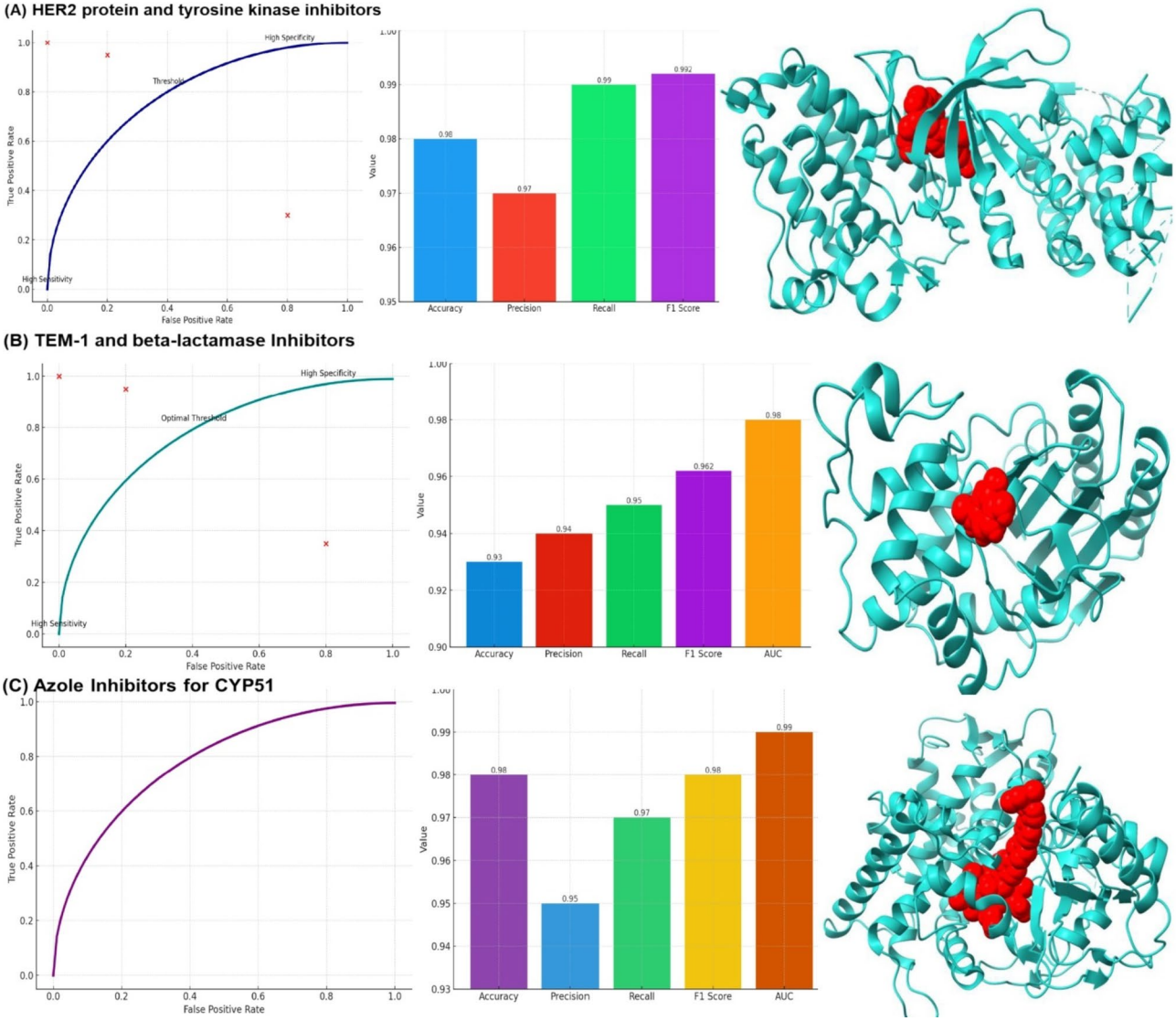

Fig. 6. Performance and binding analyses of predictive models for different inhibitors. (**A**) HER2 tyrosine kinase, (**B**) TEM-1 beta-lactamase, and (**C**) CYP51 azole inhibitors, showcasing ROC curves, key metrics, and molecular docking visualizations. The Docked complexes were visualized using Discovery studio.

VirtuDockDL was benchmarked against other virtual screening tools, including PyRMD, RosettaVS, and MzDOCK. Additionally, VirtuDockDL was tested against machine learning-based platforms such as DeepChem and docking tools like AutoDock Vina and Glide on the HER2 protein-ligand dataset, where it demonstrated superior predictive performance and efficiency for large-scale virtual screening tasks.

Conclusion

In conclusion, VirtuDockDL stands as a novel innovation in computational drug discovery. So, this web application based on Python, which uses the principles of deep learning, has shown a high potential for the virtual screening of pharmaceuticals. Combined with a complex GNN model and accurate molecular docking simulation, the platform can effectively help to find potential inhibitors with high binding affinity and specificity to the target biological macromolecules. As observed in our comparative validation studies, VirtuDockDL has further provided an excellent performance profile across different datasets such as the identification of potent tyrosine kinase inhibitors for HER2 protein and beta-lactamase inhibitors for TEM-1 for cancer therapy and bacterial infection treatment respectively. Furthermore, the platform effectively aimed at identifying potential azole inhibitors for Candidiasis targeting the CYP51 enzyme. These outcomes not only corroborate the predictive accuracy of VirtuDockDL but also support its applicability to enhance the initial stages of drug design. In addition, the contribution of the application in Marburg virus research through the discovery of non-covalent inhibitors against the VP35 protein shows the possibility of fighting virulent pathogens. The findings shown here offer a new direction for antiviral therapy and support the efficiency of the application in the context of the present acute virus outbreak. The continued development and expansion of VirtuDockDL is expected to

significantly impact the drug discovery field, enhancing the efficiency and accuracy of identifying potential drug candidates. As the link between theoretical computation and practical experiment, VirtuDockDL remains the model for progress, optimizing the process of transferring innovation from laboratory to clinic.

Data availability

Data is provided within the manuscript or supplementary information files. The codes are available online at https://github.com/FatimaNoor74/VirtuDockDL.

Received: 18 July 2024; Accepted: 12 November 2024
Published online: 16 November 2024

References

1. Atanasov, A. G., Zotchev, S. B., Dirsch, V. M. & Supuran, C. T. Natural products in drug discovery: advances and opportunities. *Nat. Rev. Drug Discov.* **20**, 200–216 (2021).
2. Young, R. J. et al. The time and place for nature in drug discovery. *Jacs Au* **2**, 2400–2416 (2022).
3. Galandrin, S., Oligny-Longpré, G. & Bouvier, M. The evasive nature of drug efficacy: implications for drug discovery. *Trends Pharmacol. Sci.* **28**, 423–430 (2007).
4. Wölcke, J. & Ullmann, D. Miniaturized HTS technologies–uHTS. *Drug Discov. Today* **6**, 637–646 (2001).
5. Mayr, L. M. & Fuerst, P. The future of high-throughput screening. *SLAS Discov.* **13**, 443–448 (2008).
6. REN, Y. Research progress and challenges of network pharmacology in field of traditional Chinese medicine. *Chin. Tradit. Herb. Drugs*, 4789–4797 (2020).
7. Li, S. Network pharmacology evaluation method guidance-draft. *World J. Tradit. Chin. Med.* **7**, 146 (2021).
8. Jordan, M. I. & Mitchell, T. M. Machine learning: Trends, perspectives, and prospects. *Science (New York N Y).* **349**, 255–260. https://doi.org/10.1126/science.aaa8415 (2015).
9. Patel, L., Shukla, T., Huang, X., Ussery, D. W. & Wang, S. Machine learning methods in drug discovery. *Molecules (Basel Switzerland)* **25**. https://doi.org/10.3390/molecules25225277 (2020).
10. Jiménez-Luna, J., Grisoni, F., Weskamp, N. & Schneider, G. Artificial intelligence in drug discovery: recent advances and future perspectives. *Expert Opin. Drug Discov.* **16**, 949–959. https://doi.org/10.1080/17460441.2021.1909567 (2021).
11. Zhu, H. Big data and artificial intelligence modeling for drug discovery. *Annu. Rev. Pharmacol. Toxicol.* **60**, 573–589. https://doi.org/10.1146/annurev-pharmtox-010919-023324 (2020).
12. Machado, L. A., Krempser, E. & Guimarães, A. C. R. A machine learning-based virtual screening for natural compounds capable of inhibiting the HIV-1 integrase. *Front. Drug Discov.* **2**, 954911 (2022).
13. Zhou, J. et al. Classification and design of HIV-1 integrase inhibitors based on machine learning. *Comput. Math. Methods Med.* **2021** (2021).
14. Sun, H. et al. Constructing and validating high-performance MIEC-SVM models in virtual screening for kinases: a better way for actives discovery. *Sci. Rep.* **6**, 24817 (2016).
15. Alzubaidi, L. et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *J. Big Data*. **8**, 53. https://doi.org/10.1186/s40537-021-00444-8 (2021).
16. Mamoshina, P., Vieira, A., Putin, E. & Zhavoronkov, A. Applications of deep learning in biomedicine. *Mol. Pharm.* **13**, 1445–1454. https://doi.org/10.1021/acs.molpharmaceut.5b00982 (2016).
17. Jiang, D. et al. Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. *J. Cheminform.* **13**, 1–23 (2021).
18. Bento, A. P. et al. An open source chemical structure curation pipeline using RDKit. *J. Cheminform.* **12**, 1–16 (2020).
19. Noor, F., Asif, M., Ashfaq, U. A., Qasim, M. & Tahir Ul Qamar, M. Machine learning for synergistic network pharmacology: a comprehensive overview. *Brief. Bioinform.* **24**, bbad120 (2023).
20. Reynolds, D. A. Gaussian mixture models. *Encyclopedia Biometrics* 741 (2009).
21. Shahapure, K. R. & Nicholas, C. In *IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)* 747–748 (IEEE, 2020).
22. Vergani, A. A. & Binaghi, E. In *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE).* 1–8 (IEEE, 2018).
23. Ghimire, D. Comparative study on Python web frameworks: Flask and Django (2020).
24. Bento, A. P. et al. An open source chemical structure curation pipeline using RDKit. **12**, 1–16 (2020).
25. Eastman, P. et al. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput. Biol.* **13**, e1005659 (2017).
26. Imambi, S., Prakash, K. B. & Kanagachidambaresan, G. R. Programming with TensorFlow: solution for edge computing applications, PyTorch, 87–104 (2021).
27. Huey, R., Morris, G. M. & Forli, S. Using AutoDock 4 and AutoDock Vina with AutoDockTools: A Tutorial (2012).
28. Gilson, M. K. et al. BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.* **44**, D1045–D1053 (2016).
29. Kim, S. et al. PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.* **47**, D1102–D1109 (2019).
30. Gaulton, A. et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **40**, D1100–D1107 (2012).
31. Mysinger, M. M., Carchia, M., Irwin, J. J. & Shoichet, B. K. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J. Med. Chem.* **55**, 6582–6594 (2012).
32. Tay, D. W. P., Yeo, N. Z. X., Adaikkappan, K., Lim, Y. H. & Ang, S. J. 67 million natural product-like compound database generated via molecular language processing. *Sci. data*. **10**, 296. https://doi.org/10.1038/s41597-023-02207-x (2023).
33. Amendola, G. & Cosconati, S. PyRMD: a new fully automated Ai-powered ligand-based virtual screening tool. *J. Chem. Inf. Model.* **61**, 3835–3845 (2021).
34. Zhou, G. et al. An artificial intelligence accelerated virtual screening platform for drug discovery. *Nat. Commun.* **15**, 7761 (2024).
35. Kabier, M. et al. MzDOCK: A free ready-to-use GUI-based pipeline for molecular docking simulations (2024).
36. Ramsundar, B. *Molecular Machine Learning with DeepChem* (Stanford University, 2018).
37. Heid, E. et al. Chemprop: a machine learning package for chemical property prediction. *J. Chem. Inf. Model.* **64**, 9–17 (2023).
38. Zhou, Z., Kearnes, S., Li, L., Zare, R. N. & Riley, P. Optimization of molecules via deep reinforcement learning. *Sci. Rep.* **9**, 10752 (2019).

Acknowledgements

The authors extend their appreciation to Taif University, Saudi Arabia, for supporting this work through project number (TU-DSPP-2024-57).

Author contributions

M.T.Q. and F.N. conceptualized and designed the project. F.N. implemented the models, conducted the experiments, analyzed the data, and drafted the manuscript. M.J., A.H.A., M.A., and S.G. contributed to data analysis and result validation. M.T.Q. provided analytical support, conducted external validation, and supervised the study. M.T.Q., A.H.A., and M.A. secured funding and resources. All authors reviewed, revised, and approved the final manuscript.

Funding

This research was funded by Taif University, Saudi Arabia, Project No. (TU-DSPP-2024-57).

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-79799-w.

Correspondence and requests for materials should be addressed to M.T.u.Q.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.