RSC
Medicinal Chemistry

RESEARCH ARTICLE

View Article Online
View Journal

Cite this: DOI: 10.1039/d5md00361j

EGFR^{AP}: a predictive machine learning model for assessing small molecule activity against the epidermal growth factor receptor

Ashish Gupta,[a] Amarinder S. Thind[c] and Rituraj Purohit *[ab]

Received 26th April 2025,
Accepted 8th July 2025

DOI: 10.1039/d5md00361j

rsc.li/medchem

Epidermal growth factor receptor (EGFR) is a membrane-bound protein that interacts with epidermal growth factor, triggering receptor dimerization and tyrosine autophosphorylation, subsequently promoting cell proliferation. EGFR-associated pathways regulate cell housekeeping functions like growth, division, and apoptosis. However, the mutations/overexpression of EGFR cause unrestrained cell differentiation, leading to tumorigenesis. This study proposes a machine-learning-based tool, EGFR^{AP}, to compute novel molecules' biological activities (pIC$_{50}$) against EGFR. The tool is based on a robust quantitative structure–activity relationship (QSAR) model, trained on a large dataset of existing EGFR inhibitors using multiple machine learning algorithms. The extra trees regressor (ET) model showed promising results for the training dataset with an R^2 value of 0.99, an RMSE value of 0.07 and an MAE of 0.009. The Pearson correlation between the observed and predicted pIC$_{50}$ values of the training set inhibitors was also very substantial, *i.e.* 0.99. The model was then validated using a test dataset, and the findings were satisfactory. An R^2 value of 0.67, an RMSE of 0.89 and an MAE of 0.61 were detected for the test dataset, and the Pearson correlation coefficient of observed/predicted pIC$_{50}$ values was 0.82. The model was probed for overfitting using 10-fold cross-validation, and a series of structure-based drug design experiments were performed to validate the tool's predictions. The findings backed up the model's performance. This tool will be of significant importance to medicinal chemists in identifying promising EGFR inhibitors.

1. Introduction

The epidermal growth factor receptor (EGFR), a transmembrane tyrosine kinase protein, acts as a receptor for EGF family proteins, regulating key processes like human epithelial cell growth, spread, invasion, and cell death.[1] It mediates the signalling pathways responsible for cell proliferation, angiogenesis, and apoptosis.[2] It is also involved in tumour metastasis, making it a crucial drug target for curing malignancies.[3,4] There are four members in the EGFR family, which include EGFR (HER1/ErbB1), ErbB2 (HER2/neu), ErbB3 (HER3), and ErbB4 (HER4).[5] These receptors are activated upon ligand binding by dimerization. The activated receptors undergo autophosphorylation near the structural domain to initiate downstream signalling pathways like RAS/RAF/MEK and STAT.[6] Mutations in EGFR lead to abnormal signalling, causing tumorigenesis and overexpression of EGFR in various cancers.[7] The overexpression of EGFR in multiple cancers results in resistance to traditional radiotherapy and hormonal therapy.[8] Thus, the international guidelines recommend using anti-EGFR medications as the initial treatment for patients with high EGFR mutations due to their improved efficacy and safety compared to standard chemotherapy.[9,10]

Monoclonal antibodies and tyrosine kinase inhibitors are widely used to target EGFR.[11] Cetuximab and panitumumab are the commonly used anti-EGFR monoclonal antibodies to treat certain cancers.[12] These antibodies bind to EGFR's extracellular domain and prevent endogenous ligands from interacting with EGFR. As a result, the EGFR signalling cascade is disrupted, and EGFR tyrosine kinase activation is suppressed.[13] The tyrosine kinase inhibitors, on the other hand, target EGFR by inhibiting its phosphorylation and thus blocking the downstream signalling cascades. These medications offer an advantage over monoclonal antibodies as they are also effective in treating cancers associated with mutations in EGFR.[14–16] EGFR is an important drug target to cure cancer, and efforts are still ongoing to come up with novel and more effective anti-EGFR inhibitors or medications.[17–20]

[a] *Structural Bioinformatics Lab, Biotechnology division, CSIR-Institute of Himalayan Bioresource Technology (CSIR-IHBT), Palampur, HP, 176061, India. E-mail: rituraj.purohit@csir.res.in, riturajpurohit@gmail.com*
[b] *Academy of Scientific & Innovative Research (AcSIR), Ghaziabad-201002, India*
[c] *Illawarra Shoalhaven Local Health District (ISLHD), Wollongong, NSW, Australia*

Computer-aided drug discovery (CADD) strategies are essential in identifying and developing novel drug-like molecules. CADD techniques can be combined with other computational methods for more accurate findings.[21,22] Chang *et al.* applied a machine learning-based virtual screening setup to identify a new generation of EGFR tyrosine kinase inhibitors. They synthesized novel small molecules from the knowledge obtained from the structural molecular descriptors. Their identified molecules were active against some mutated kinases.[23] In another report by Nada *et al.*, novel small molecules were reported as potent candidates for treating breast cancer by inhibiting EGFR. They used rational drug design strategies coupled with artificial intelligence applications. The identified compounds showed promising anti-proliferative activity in cancer cell lines.[24] Huo *et al.* developed a ligand-based virtual screening protocol and obtained active structures by screening a library of ∼5 million compounds. Structure–activity relationship (SAR) and quantitative structure–activity relationship (QSAR) models developed using an SVM-based approach further filtered the most active EGFR inhibitors, which showed potential anti-EGFR activity in the experimental validations.[25] Eissa *et al.* also reported a novel theobromine derivative as a potent EGFR inhibitor based on the findings of a computational investigation. The identified compound was initially tested *in silico* and showed tremendous potential. The lead compound was validated for its anti-EGFR activity using *in vitro* experiments. It showed very high activity against cancer cell lines and low potency against regular ones. The identified compound had optimal ADMET properties and was selective against EGFR.[26]

EGFR has a huge role in effective cancer management, and many scientific investigations are conducted to develop novel EGFR inhibitors. The present work introduces a machine learning-based tool that depicts small molecules' biological activities (pIC_{50}) against EGFR. A set of 2D molecular descriptors was calculated for the previously reported EGFR inhibitors. The inhibitors were segregated based on their activity against EGFR and were used to prepare predictive regression models. The models were built using the extra trees (ET) regression algorithms in Python's machine learning modules. The files needed to run the app on any local machine can be accessed publicly at **https://github.com/amarinderthind/EGFR-ap**. The knowledge derived from this tool can be decisive in developing potent inhibitors targeting EGFR.

2. Materials and methods

2.1. Dataset preparation

For the present work, 16 715 EGFR inhibitors (File S1[27]) were initially retrieved from the ChEMBL database (**https://www.ebi.ac.uk/chembl/**).[28] ChEMBL is a meticulously curated database housing bioactive molecules exhibiting drug-like characteristics. It amalgamates chemical, bioactivity, and genomic data, facilitating the seamless translation of genomic insights into innovative pharmaceuticals.[29] The retrieved file contained various features regarding these inhibitors, such as ChEMBL IDs, canonical smiles, activities (IC_{50}), references, *etc.* The raw file was cleaned, the duplicate molecules and molecules with approximate activities were removed, and only relevant information was retained. The remaining inhibitor dataset (9508 EGFR inhibitors) was divided into three categories based on the IC_{50} values. These categories were active ($IC_{50} \leq 1000$ nM), intermediately active ($IC_{50} > 1000$ nM $\leq 10\,000$ nM), and inactive ($IC_{50} \geq 10\,000$ nM) EGFR inhibitors (File S2[27]). Only the active and inactive inhibitors (8102 inhibitors) were then taken up for further analysis. The IC_{50} values of the inhibitors were converted to pIC_{50} values (negative logarithm of IC_{50} values), which facilitates the comparison of different inhibitors at the same molar levels (File S3[27]).

2.2. Molecular descriptor computation

2D molecular descriptors were calculated using the PaDELPy module in Python. This module aids in calculating a variety of fingerprints and 2D descriptors for small molecules. PaDELPy is a Python wrapper for the PaDEL-Descriptor molecular descriptor calculation software.[30] It provides seamless Python access to the PaDEL-Descriptor command-line interface, enabling direct utilization.

2.3. Dataset curation and feature selection

The activity (pIC_{50}) of the EGFR inhibitors was considered a dependent variable, and the descriptors were independent variables. Data curation and feature selection were performed for independent variables, improving the feature-to-sample ratio. Feature selection allows the choice of essential features for ML model-building, which reduces the complexity of ML models while avoiding overfitting.[31] Firstly, the rows containing "missing" or "infinite" values were deleted. The next step was to remove the empty descriptor columns from the dataset. This was followed by removing the descriptors with at least 25% missing observations. After initial data curation, the dataset underwent exhaustive filtering to reduce the number of molecular descriptors for final model training. Next, the correlation between the dependent and independent variables was computed, and descriptors with a correlation of less than 0.25 with the dependent variable were subsequently removed from the dataset to ensure that only features with a meaningful relationship with the target were retained. To mitigate multicollinearity, we removed highly correlated descriptors. A correlation matrix of the molecular descriptors among themselves was generated. Redundant descriptors were identified by detecting correlations of 0.8 or higher among themselves. Only one representative molecular descriptor was kept for such high correlations, while others were dropped from the dataset. Further, the dataset was investigated to detect outliers, and capping was performed to address extreme values.

2.4. Model building

After feature selection, the dataset was scaled to perform multivariate modelling. The dataset was randomly split into training and test sets for model building. 80% of the data was treated as a training set, and the remaining 20% was treated as a test set. Forty-two machine learning linear regression algorithms were used for model building and evaluated using root mean squared error (RMSE) and R-squared (R^2) values (Fig. 1). This was achieved using the lazy regressor module of the Sci-kit learn (sklearn) Python library. The algorithms were compared, and the algorithm performing well on these parameters was selected for final model building.

2.5. Model validation

The trained models were further tested for their accuracy and fitness. Statistical features like training R^2 score, testing R^2 score, RMSE training, RMSE testing, and correlation between the experimental and predicted activities (pIC_{50}) were performed on the developed models. Initially, the model was validated using a test set constituting 20% of the total dataset. The model was checked for overfitting using the k-fold (k =10) cross-validation technique. Finally, the correlation between the experimental and predicted pIC_{50} values was also calculated to establish the robustness of the cross-validation run.

2.6. *In silico* validation

After validating the model predictions, the app was deployed and used to predict the pIC_{50} of 10 standard EGFR drugs. The predictions from the app were then compared with the experimental activities of these EGFR drugs to get insights into the efficiency of the projections. The top 2 active EGFR drugs, almonertinib and gefitinib (based on the app's predictions), were prepared for further structure-based drug design experiments. The experimental setup included two other molecules from the CHEMBL (Comp1 and Comp2) database reported as active EGFR inhibitors from the model validation stage (Fig. 2).

2.6.1. Molecular docking. Molecular docking is a well-established structural drug design strategy widely used.[32] A high-resolution (1.98 Å) 3D structure of EGFR (PDB ID: 8SC7)[33] was obtained from the protein data bank.[34] The initial inhibitor structures were prepared using ChemDraw software and were optimized using a DFT-based protocol[35] in the BIOVIA Discovery Studio suite[36] before commencing the structure-based experiments. DFT-based methods perform structural optimization of the small molecules by correcting their molecular geometries and electronic structures.[37–42]

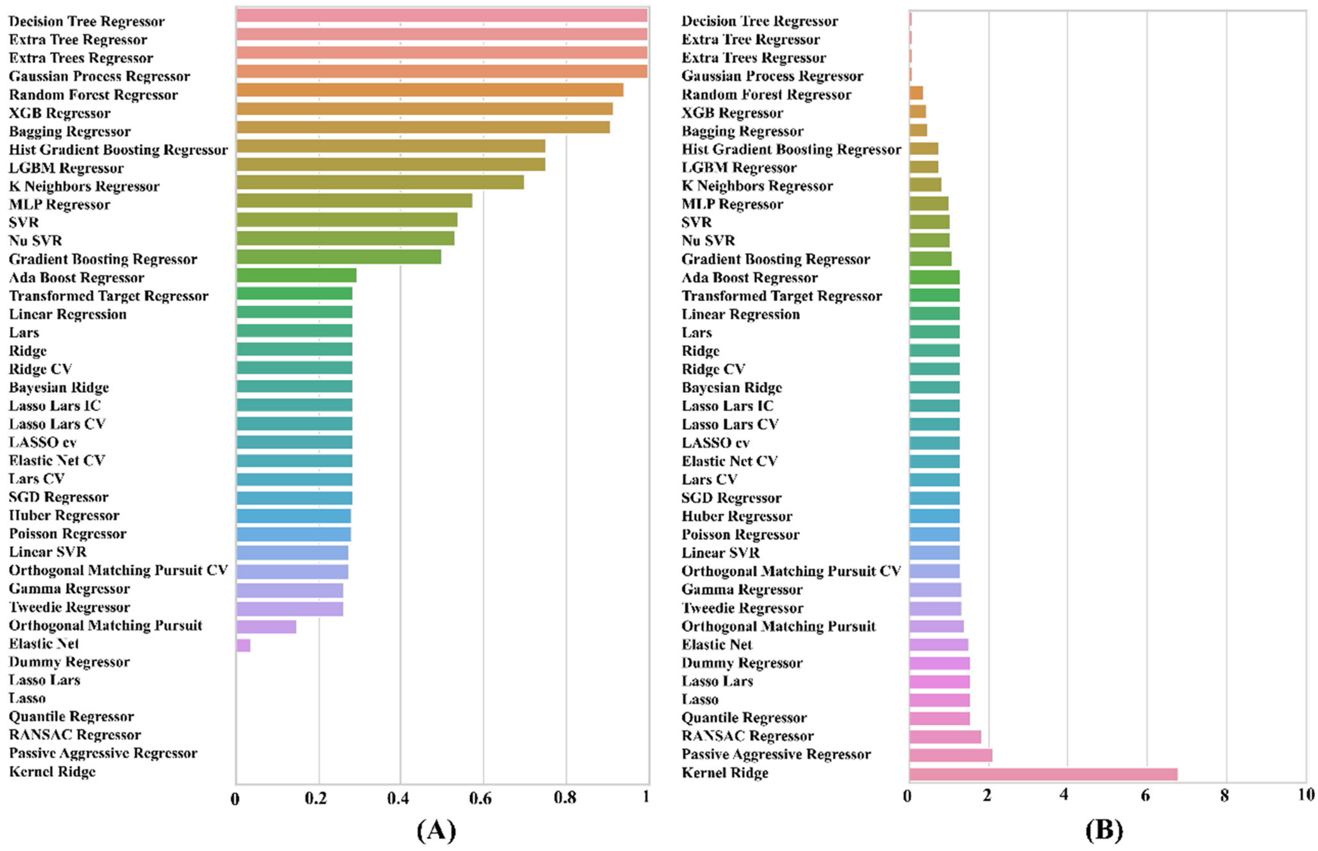

Fig. 1 Plots showing (A) R-square (R^2), and (B) RMSE to complete the computation by different ML algorithms.

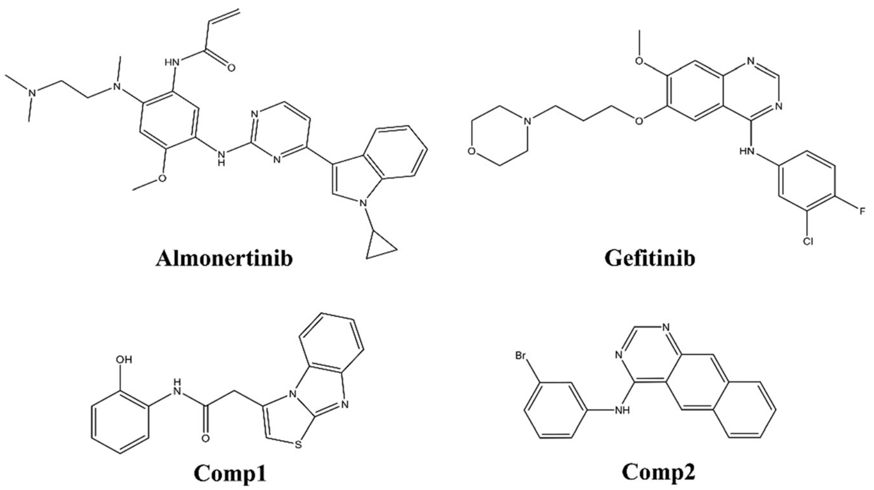

Fig. 2 2D chemical structures of the EGFR inhibitors considered for the *in silico* validation.

After structural optimization, these ligands were docked in the EGFR active site, which was defined using the co-crystallized ligand in the EGFR receptor crystal structure as a reference. The default parameters in Discovery Studio's CDOCKER utility were implemented for the molecular docking experiment.[43] The ligands were treated as flexible molecules, while the protein structure was pre-set to be rigid. The protein–ligand complexes were later subjected to detailed interaction analysis to explore the binding patterns and to identify the key interacting amino acids at the EGFR active site.

2.6.2. Molecular dynamics simulations. The best ligand conformations were selected for more refined classical molecular dynamics-based experiments using the GROMACS package.[44–46] The protein topology was generated using the CHARMM36 force field,[47] and ligand topology was obtained using a web-based tool called CgenFF.[43] The TIP3P water model was used for the solvation of the protein–ligand complexes, followed by the addition of counter ions to neutralize the systems. The steepest descent algorithm was used for energy minimization, followed by a 1000 ps equilibration step utilizing the NVT and NPT ensembles. During the equilibration process, the Parrinello–Rahman method maintained a constant pressure of 1 bar. At the same time, the Berendsen thermostat was applied to regulate the temperature at 300 K. The Energy was minimized, and equilibrated protein–ligand systems were then subjected to a 500 ns production MD run. The particle mesh Ewald (PME) algorithm, widely used for long-range electrostatic interactions, was employed in the calculations. Real-space interactions were considered within a 1.0 nm cutoff, while reciprocal-space interactions were determined using a Fourier transform with optimized grid spacing and fourth-order B-spline interpolation.[48] Van der Waals interactions, classified as short-range with a 1 nm cutoff, were computed using the Lennard-Jones potential. Meanwhile, the linear constraint algorithm (LINCS) was utilized to constrain all covalent bond lengths, including hydrogen bonds.[49] The simulation trajectories were analyzed using built-in GROMACS scripts.

2.6.3. Per-residue decomposition energy (PRDE) analysis. After analyzing the simulated protein–ligand systems for structural stability, the key amino acids involved in intermolecular interactions with the EGFR inhibitors were identified through detailed interaction analysis. The per-residue decomposition energy (PRDE) analysis was done to understand the energetic contribution of the interacting amino acids towards the binding energy. In PRDE analysis, the individual energetic contribution of the amino acid residues is computed by breaking the overall binding Energy of a protein–ligand system into van der Waals, electrostatic and non-polar solvation energy terms.[50] It provides a quantitative account of the energetics of ligand–amino acid binding in terms of the energy contribution of the amino acids along with their backbone and side chains. This is an established method to understand the binding mechanism of protein–ligand and protein–protein systems.[51]

3. Results and discussion

Here, we present a machine learning-based tool that can be important in identifying novel leads and optimizing existing medication strategies to treat cancers targeting EGFR. Various machine learning models of Python were explored to prepare this tool and further validate it before it was operational. The cleaned dataset consisted of 8102 EGFR inhibitors. This dataset was subjected to a series of operations to achieve the proposed objective.

A set of 2D molecular descriptors was computed for the inhibitor dataset. A total of 1444 2D descriptors were obtained for the inhibitors. The raw data was multidimensional and was cleaned before model building. The data curation and feature selection were automated using various machine-learning modules available in Python. We carefully examined the raw data and found a few rows containing only “missing” or “infinite” values in the cells.

These were removed in the first step of the feature selection procedure, and we were left with 8098 EGFR inhibitors in the dataset (File S4[27]). We performed a Murcko scaffold analysis on the remaining dataset to understand chemical diversity. This analysis revealed that the dataset contains more than 2000 unique Murcko scaffolds, suggesting rich chemical diversity. The Supplementary Figure file[27] presents the top 10 most frequent MURCKO scaffolds.

Next, the descriptor columns were inspected for any meaningless data. Few descriptors had only missing values in the observation cells and thus were removed from the data. After this step, we were left with 1435 molecular descriptors subjected to the next feature selection round (File S5[27]). Later, the data was again checked for any missing values, and it was seen that some columns had nil observations in many cells, which needed to be taken care of. So, in the next phase, we selected the complete dataset and explored it to highlight and remove the molecule descriptor columns where at least 25% of the observations were missing. Eliminating columns with missing values is recommended to enhance data quality and relevance. It helps reduce noise and data sparsity, thereby improving the statistical integrity of the dataset. Simplifying models by removing such columns prevents overfitting and boosts computational efficiency, ultimately improving model performance.[52,53] After this feature selection procedure, we were left with 900 molecular descriptors (File S6[27]).

Next, we checked for the relationship between the remaining molecular descriptors and the pIC_{50} of the EGFR inhibitors. A correlation matrix was prepared, and the descriptors showing a correlation of less than 0.25 with the pIC_{50} of the inhibitors were eliminated. The whole motive behind this step was to remove the molecular descriptors that did not direct the EGFR inhibitors' inhibitory potential to improve our tool's predictions. After eliminating such descriptors, we were left with 41 features in our dataset (File S7[27]). In addition, multicollinearity among independent variables (features) can exacerbate model complexity without substantially augmenting its informational value. To deal with multicollinearity, we prepared a correlation matrix depicting the correlation among the remaining independent features. On analyzing the matrix, features showing a correlation of 0.8 or more among each other were reported, and only one feature was kept in the dataset as their representative. In data analysis, choosing a single feature from a set of highly correlated features is advised to minimize redundancy, avoid multicollinearity, and streamline models. This approach improves model interpretability and boosts computational efficiency.[54–56] Finally, our dataset consisted of only 15 molecular descriptors for model building (File S8[27]). The cleaned data file was inspected to detect and treat outliers so they did not affect the predictions. Also, the distribution plots of the untreated features were plotted to unravel their distribution patterns (Fig. 3).

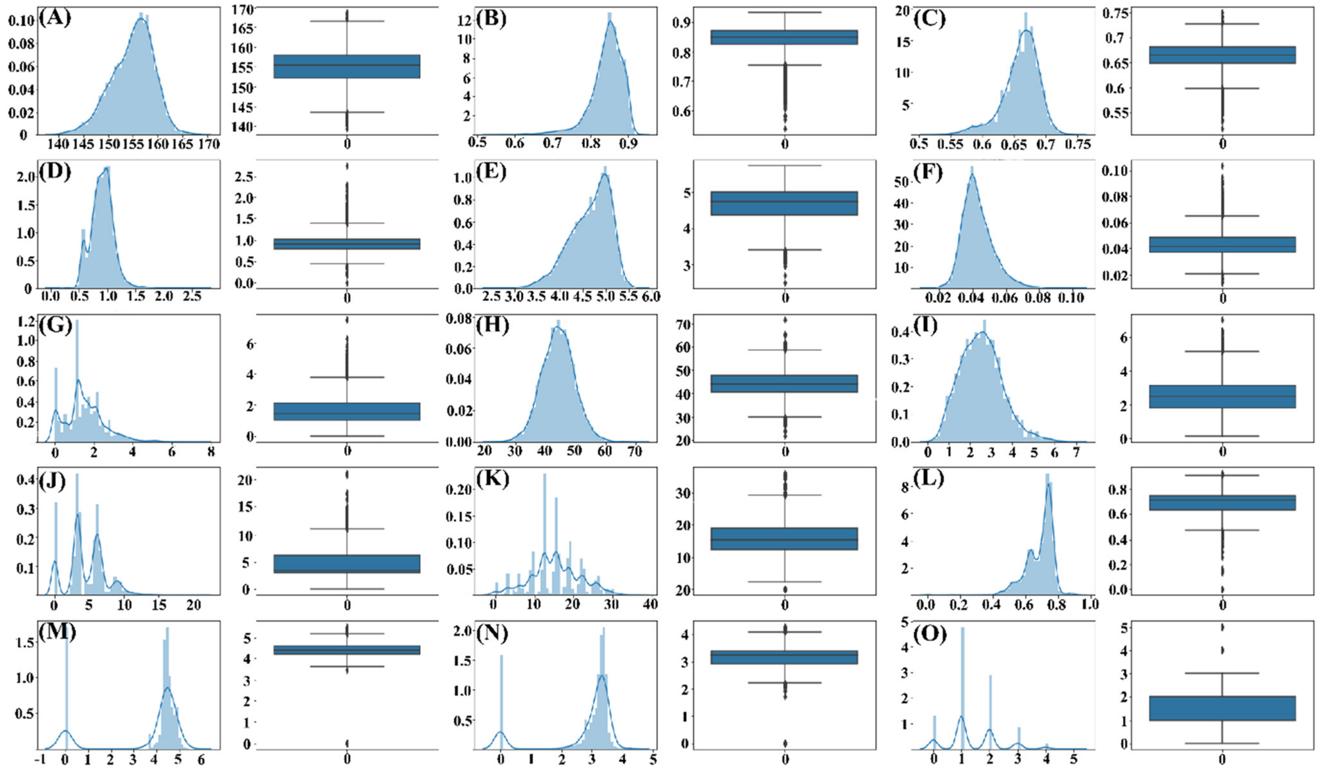

Fig. 3 Distribution and box plots to show the distribution pattern of the uncapped data. (A) AATS2i, (B) BIC3, (C) ETA_BetaP_s, (D) GATS6m, (E) IC2, (F) JGI3, (G) MDEN-22, (H) MIC2, (I) MLFER_BO, (J) SssNH, (K) WTPT-5, (L) maxHother, (M) maxaaN, (N) maxssNH, and (O) nT6HeteroRing.

The first (Q1) and third quartiles (Q3) and the interquartile range (IQR) for the variables in the dataset were calculated initially. The values falling below the lower limit (Q1 − 1.5 × IQR) and the upper limit (Q3 + 1.5 × IQR) were identified as outliers for each molecular descriptor. These outliers were then treated using the data capping technique. The outliers were then replaced with the values of the Q1 and Q3 for the lower and higher outliers, respectively (File S9[27]).

The capped values were plotted again to compare the capped and untreated data distributions (Fig. 4). The figures show a smooth distribution in the case of the capped data; thus, we proceeded with the model-building process.

The data now consisted of 1 dependent variable (pIC_{50}) and 15 independent variables, which included AATS2i, BIC3, ETA_BetaP_s, GATS6m, IC2, JGI3, MDEN-22, MIC2, MLFER_BO, SssNH, WTPT-5, maxHother, maxaaN, maxssNH, nT6HeteroRing. These variables represent various 2D molecular descriptors utilized for model building. Here, AATS2i and GATS6m are the autocorrelation descriptors based on molecular graph theory and describe the interdependence of atomic properties and similarity among molecules. IC2, MIC2, and BIC3 are information content descriptors representing information about chemical structure bonds and symmetry. JGI3 is a topological charge representing molecular descriptors, and MDEN-22 represents molecular distance descriptors. WTPT-5 is the weighted path descriptor, and it is used to calculate the molecular branching of the chemical structures.[57] These descriptors are primarily based on graph theory and compute the physicochemical properties of the molecules by considering the molecule structure as a graph and the atoms as nodes and vertices.[58,59]

The dataset was scaled so that the final variables were on a similar scale (File S10[27]). The dataset was segregated into a training and a test set for model building. The different machine learning regression algorithms were then evaluated for their performance to check which algorithms give the finest predictions for the dataset (the detailed information is provided in Files S11 and S12[27]). The extra trees regressor (ET) algorithm showed the best results for all these parameters for the training at the test set and was selected for future model building. Extra trees, short for extremely randomized trees, constitutes an ensemble machine learning technique rooted in decision trees. This approach involves generating numerous unpruned decision trees from the provided training dataset.

The regression predictions are determined by averaging the outcomes of these decision trees. The aggregated predictions and arithmetic averaging of the trees determine the regression predictions.[60–62] The ET algorithm showed an R^2 value of 0.997, an RMSE value of 0.08, and the time taken to complete the prediction was 0.036 seconds for the training dataset. The test set R^2 value was observed to be 0.62, the RMSE was 0.97, and the time taken was 2.67 seconds. The final model-building protocol was then initiated. After splitting up the data, the essential statistical parameters were

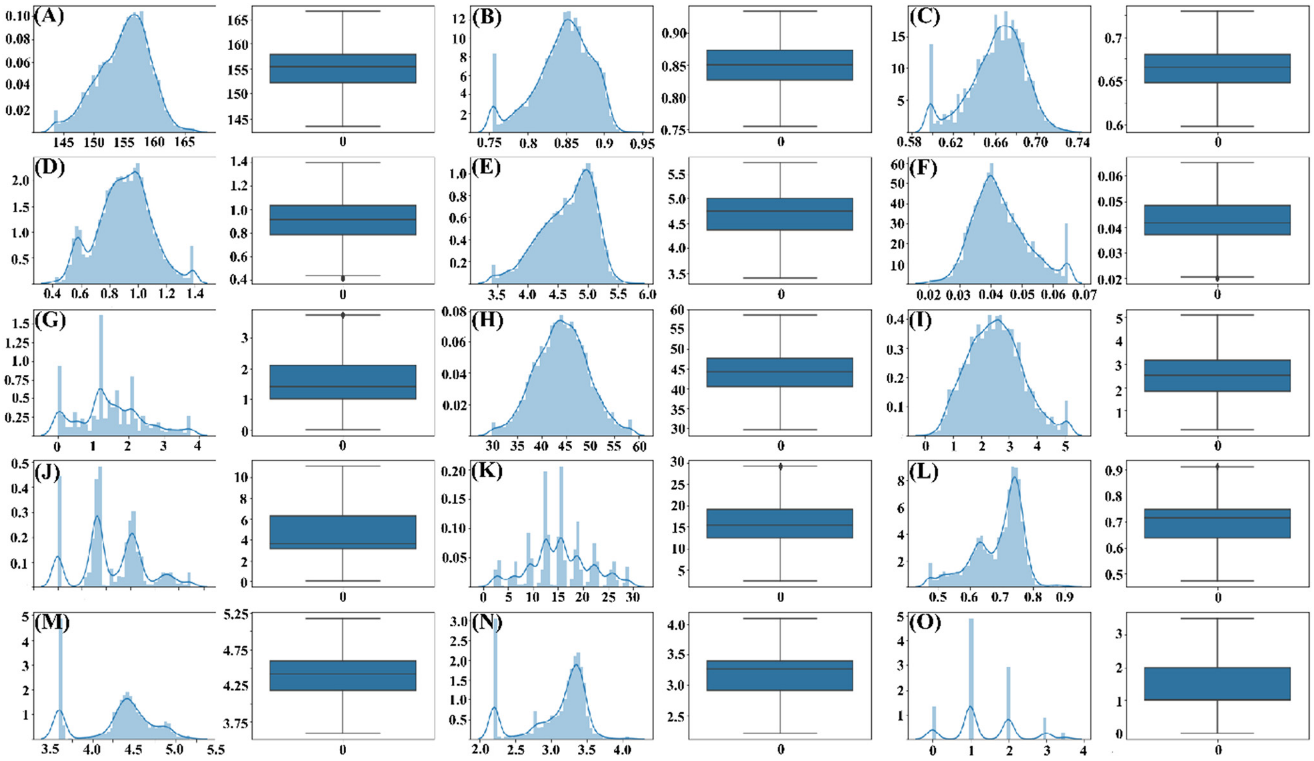

Fig. 4 Distribution and box plots to show the distribution pattern of the capped data. (A) AATS2i, (B) BIC3, (C) ETA_BetaP_s, (D) GATS6m, (E) IC2, (F) JGI3, (G) MDEN-22, (H) MIC2, (I) MLFER_BO, (J) SssNH, (K) WTPT-5, (L) maxHother, (M) maxaaN, (N) maxssNH, and (O) nT6HeteroRing.

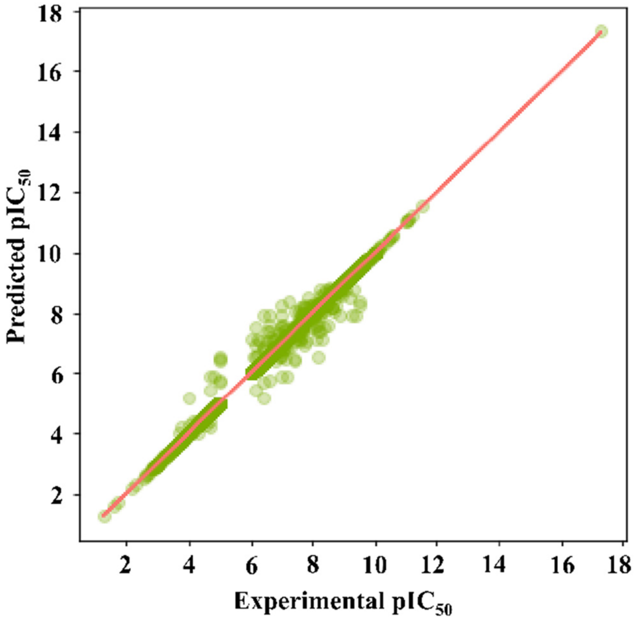

Fig. 5 Prediction plot of the ET algorithm for the training set EGFR inhibitors.

set. The final model gave an R^2 score of 0.99, an RMSE value of 0.07, an MAE of 0.009 and a Pearson correlation coefficient of 0.99 between the predicted and experimental pIC_{50} for the training data (Fig. 5). These values suggested that the developed model was robust and the predictions for the training set were accurate.

These findings were then subjected to validation using the external test set, and we observed an R^2 value of 0.67, an RMSE value of 0.89, an MAE of 0.61 and the Pearson correlation coefficient between the test pIC_{50} and predicted pIC_{50} was 0.82. These numbers suggested that the developed model predicted well for the training set and could be used to predict the biological activities of untested compounds against EGFR.

The model was then tested for overfitting using *k*-fold cross-validation. Cross-validation refers to evaluating the learning models to see if there is any overfitting. The *k*-fold cross-validation is one of the most widely used cross-validation techniques. It involves partitioning the total data into equal *k* chunks or folds. This is followed by *k* iterations of model building, such that in every run, *k*-1 folds are treated as the training set and the remaining fold as the test set.[63] Here, the 10-fold cross-validation and the statistical features were inferred. The cross-validation R^2 score and the RMSE were observed to be 0.63 and 0.94, respectively. The Pearson correlation coefficient between the experimental pIC_{50} and predicted (cross-validation) pIC_{50} was reported to be 0.79 (Fig. 6). These figures suggest that the developed model performed well in the cross-validation run and was not affected much by the training and test set variations. This indicates that the model was not over-fitted and could be used for predictions.

We also performed a comparison between the already reported ML-based tools and $EGFR^{AP}$ to understand how the tool performs. The comparison between the statistical parameters obtained for the $EGFR^{AP}$ tool and other reported machine learning-based tools for EGFR inhibitory activity prediction is provided in Table 1.

The tool presented in this work held well during the rigorous development and validation protocols. It provides an advantage for the medicinal chemist in rational drug design targeting EGFR. It can predict the biological activities of small molecules based on their 2D structural properties. The tool calculates the pIC_{50} values for a subset of well-known EGFR drugs. $EGFR^{AP}$ efficiently predicted the activities of these EGFR drugs, and a correlation of 0.86 was observed between the experimental PIC_{50} values and the PIC_{50} values predicted by the $EGFR^{AP}$ tool (Table 2).

These observations further highlight the accuracy of the $EGFR^{AP}$ tool predictions. After the satisfactory performance of the developed model in the validation stage, we performed a series of structure-based drug design experiments to see if the app's predictions held well in the experiments. Molecular docking studies were conducted first to understand the manner of binding of the inhibitors at the EGFR active.

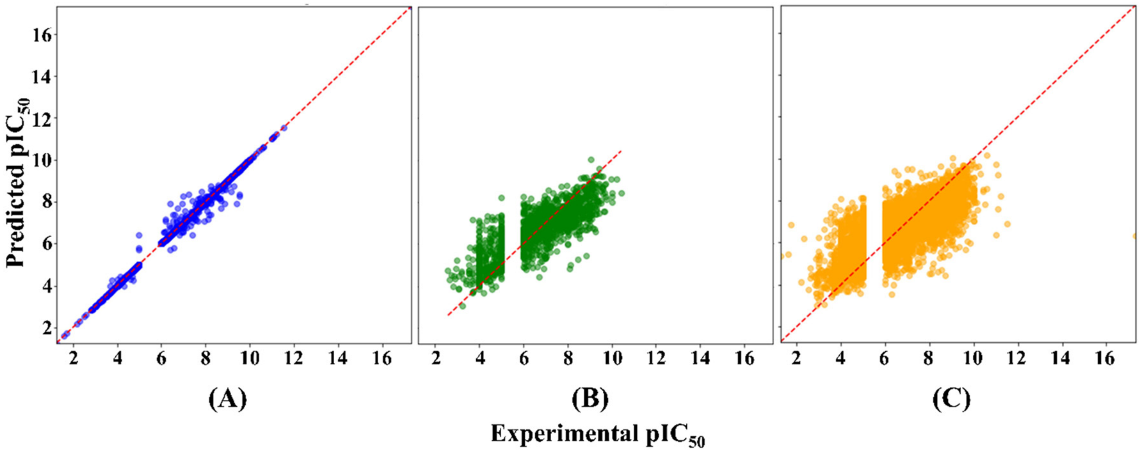

Fig. 6 Prediction plots representing the Pearson correlation coefficients between the experimental and predicted pIC_{50} for (A) the training set, (B) the test set and (C) the cross-validation run.

Table 1 Comparison between the predictive capabilities of the EGFR^{AP} tool and other previously reported tools

S. no.	Source/publication	Training R^2	Training RMSE/RMSD/MSE	Validation R^2	Algorithm used	Test RMSE/RMSD/MSE
1.	EGFR^{AP}	0.99	0.07	0.67	Extra tree regressor	0.39
2.	*ACS Omega*, 2023, **8**(35), 31784–31800	0.959		0.717	Random forest	—
3.	*New J. Chem.*, 2023, **47**, 21513	0.93	0.24		SVM-random forest	0.24
4.	*ACS Omega*, 2024, **9**(2), 2314–2324	0.853	0.147	0.745	Support vector regression	0.255
5.	*J. Chem. Inf. Model.*, 2020, **60**(10), 4640–4652	0.93		0.75	SVM	0.55
6.	*J. Bio. Str. Dyn.*, **41**(22), 12445–12463	0.83	0.54	0.69	Random forest	0.46

Molecular docking revealed a similar binding pattern for the standard EGFR drugs and the considered inhibitors. However, the interactions formed by these inhibitors varied. Almonertinib formed hydrogen bond interactions with Lys745 and Met793, and a salt bridge was observed with Glu762. Meanwhile, Leu718, Ala743, Thr790, Asp800, Tyr801 and Leu844 showed hydrophobic interactions with Almonertinib (Fig. 7(a)). In the case of Gefitinib, no electrostatic interactions were seen; however, a hydrogen bond with Met793 and hydrophobic interactions with Leu718, Ala743, Lys745 and Thr790 were observed (Fig. 7(c)).

Comp1, on the other hand, showed the maximum hydrogen bond interactions at the EGFR active site. It showed hydrogen bond interactions with Lys745, Glu762, Thr854 and Asp855 and a strong electrostatic interaction with Lys745 was also observed (Fig. 7(b)). Similarly, Comp2 was also involved in forming a hydrogen bond with Met793 and a couple of hydrophobic interactions with Leu718, Val726 and Thr790 (Fig. 7(d)). An interaction propensity graph was also prepared, and it was observed that Comp1 showed the maximum number of hydrogen bond interactions, followed by almonertinib, and the other EGFR inhibitors, gefitinib and Comp 2, showed a single hydrogen bond each. The number of hydrophobic interactions was higher in almonertinib, followed by gefitinib Comp2 and Comp1, which showed the least hydrophobic interactions (Fig. 7(e)).

These findings suggest that the EGFR drugs and the molecules reported as active EGFR inhibitors by the EGFR^{AP} tool showed almost similar patterns of intermolecular interactions at the EGFR active site, highlighting effective binding. The amino acids reported here are shown to be crucial for effective inhibitor recognition and binding at the EGFR active site in existing scientific literature. In one such article, 39 amino acids were reported near the ATP binding site in EGFR; Leu718, Val726, Ala743, Met793 and Leu844 were reported as crucial for inhibitor binding.[64] Thr790, on the other hand, is referred to as the gatekeeper of the EGFR active site, and its mutations have resulted in the loss of action of traditional EGFR inhibitors by increasing the ATP affinity of EGFR.[65] These findings suggest that the EGFR^{AP}-predicted molecules effectively bind and mimic the interaction pattern crucial for EGFR inhibition.

Table 2 The EGFR^{AP} predictions for a set of well-known EGFR drugs

EGFR drugs	EGFR^{AP} pIC$_{50}$	Experimental pIC$_{50}$
Abivertinib	6.81	6.19
Almonertinib	7.21	8.47
Brigatinib	7.03	7.17
Erlotinib	6.80	5.99
Icotinib	7.19	8.30
Lapatinib	7.17	7.97
Neratinib	6.79	7.04
Osimertinib	7.03	6.03
Vandetanib	6.67	6.30
Gefitinib	7.85	9.29
	Correlation	0.86

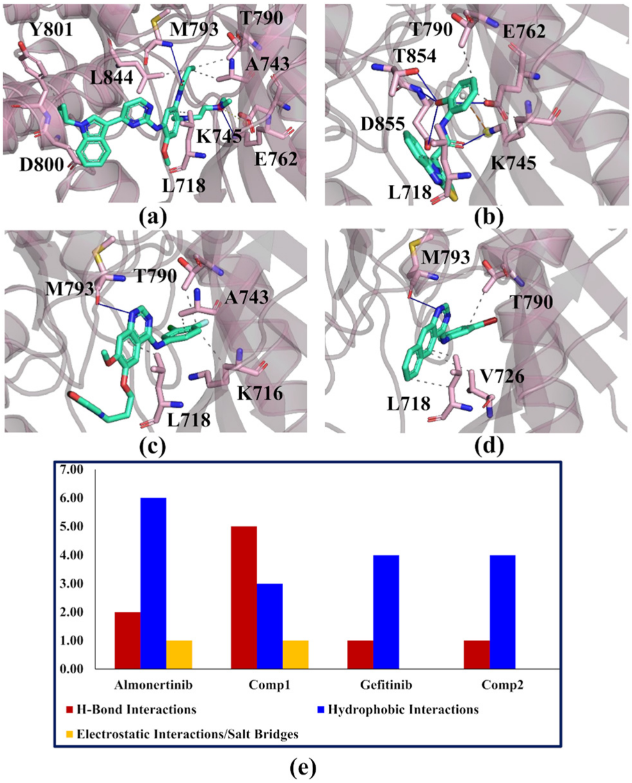

Fig. 7 Binding poses and intermolecular interactions observed at the EGFR active site post molecular docking for (a) Almonertinib, (b) Comp1, (c) Gefitinib, (d) Comp2. The H-bond interactions are represented as solid blue lines, salt bridges with red dotted lines, and hydrophobic interactions as dotted light green lines. Here, (e) represents the propensity of the interactions formed by the inhibitors considered in the study.

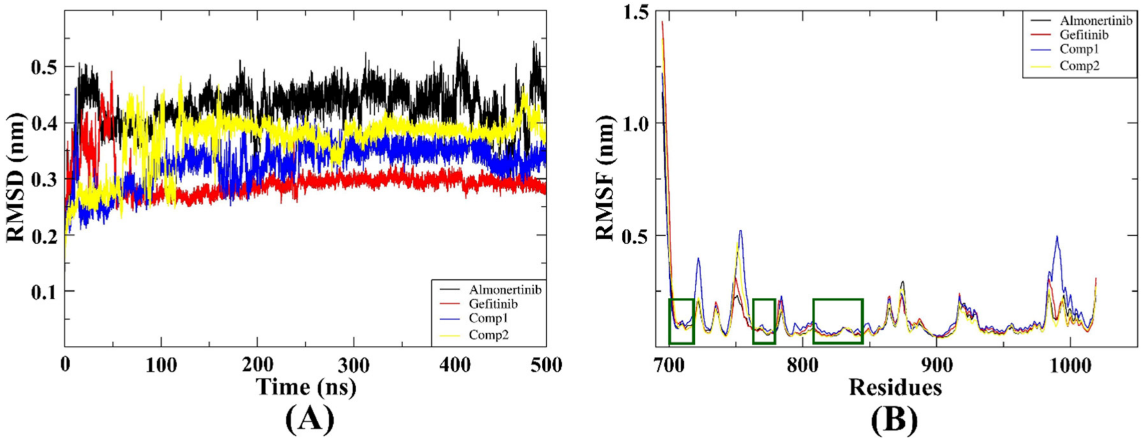

Fig. 8 (A) The root-mean-square deviation (RMSD) plots generated from the trajectories obtained after the MD simulations. (B) The root mean square fluctuations (RMSF) plots for the protein–ligand complexes. The green boxes represent the amino acid regions involved in inhibitor recognition/binding or structural maintenance of the protein structures. Almonertinib is represented by black; gefitinib is red; Comp1 is represented by blue; and Comp2 is represented by yellow.

These protein–ligand complexes were then prepared for a 500 ns simulation production run, and the resulting trajectories were analyzed. Firstly, the simulated complexes were analyzed for structural stability using the root mean square deviation (RMSD) and root means square fluctuations (RMSF) analysis. The RMSD of the complexes is calculated by measuring the overall deviation of the atoms from the initial structure throughout the simulation run time. It provides crucial insights regarding the structural stability and equilibration of the protein–ligand systems. Fig. 8(A) shows the RMSD graphs of the complexes considered here. It is evident from the graphs that the trajectories stabilized after around 200 ns of the simulation run, suggesting that the protein–ligand systems reached equilibration around this time. The EGFR^{AP} predicted molecules showed a slightly lower RMSD profile than the standard EGFR drugs,

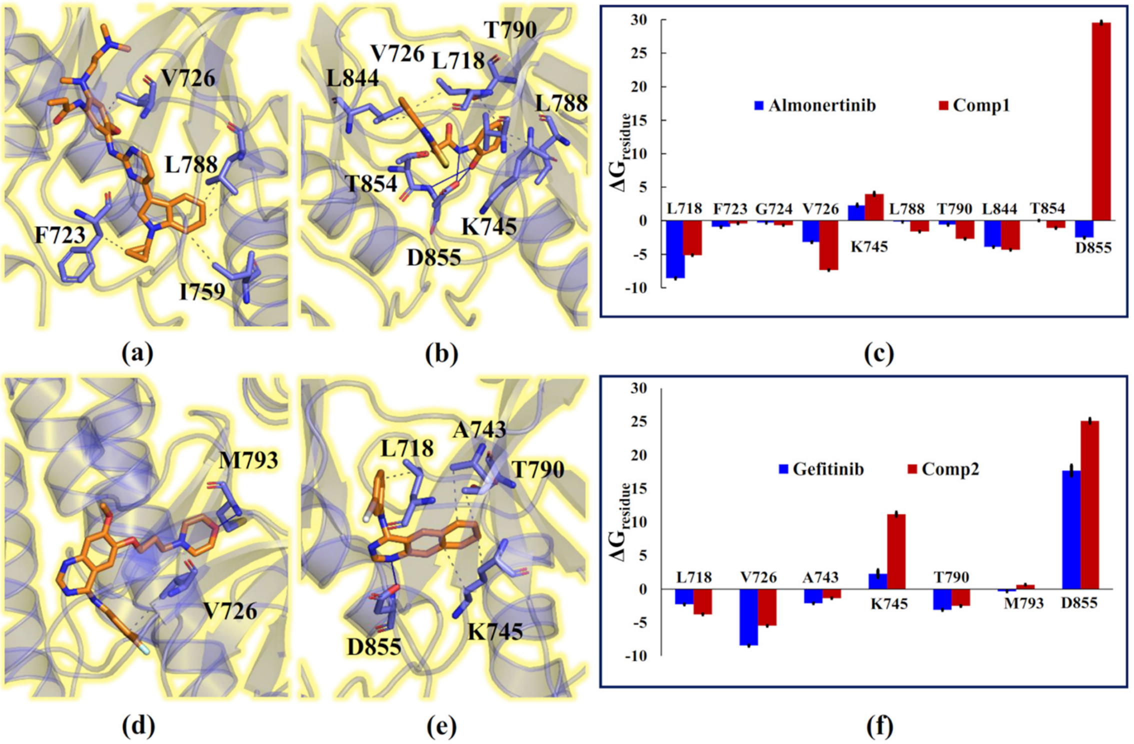

Fig. 9 Binding poses and intermolecular interactions observed at the EGFR active site post-MD simulations for (a) almonertinib, (b) Comp1, (d) gefitinib, (e) Comp2. The solid blue lines represent H-bond interactions, and red dotted lines show salt bridges with and hydrophobic interactions are represented as dotted light green lines. Here, graphs (c) and (f) represent the individual energetic contributions of the interacting amino acids in the form of per-residue decomposition energies (PRDE) in kJ mol^{-1}.

suggesting comparable structural stability. Similarly, RMSF was calculated to evaluate the flexibility of the protein by calculating the amino acid fluctuations during the simulation. It helps identify the flexible and rigid regions in a protein structure. The areas showing lesser fluctuations represent the amino acids involved in the protein's inhibitor binding or structural maintenance. From the RMSF graphs, the amino acids involved in inhibitor binding (post molecular docking) do not show any significant fluctuation in the simulated protein–ligand complexes throughout the simulation. These areas are represented in green boxes in the RMSF graphs (Fig. 8(B)); this suggests that these amino acids were involved in various intermolecular interactions with the ligands throughout the simulations. These patterns were similar for the predicted molecules and the standard EGFR drugs.

These complexes were again explored for post-MD simulation interaction analysis to see how the interaction patterns changed after the simulations. It was observed that although the structural stability was comparable in all the protein–ligand complexes, the interacting amino acids varied. Almonertinib, which showed few hydrogen bonds pre-MD simulations, now interacted only with hydrophobic interactions. It showed hydrophobic interactions with Phe723, Val726, Ile759, and Leu788 (Fig. 9(a)). In the case of Gefitinib, the hydrogen bond interaction with Met793 was retained; however, the hydrophobic interactions dropped drastically and were observed only with Val726 (Fig. 9(d)). In the case of the predicted molecules, Comp1 showed three hydrogen bond interactions with Asp855 and several hydrophobic interactions with Leu718, Val726, Lys745, Leu788, thr790, Leu844 and Thr854 (Fig. 9(b)). Comp2 also exhibited hydrogen bond interaction post-MD simulations with Asp855 and hydrophobic interactions with Leu718, Ala743, Lys745, and Thr790 (Fig. 9(e)). These amino acids are crucial for efficient EGFR inhibition, as per the earlier reports mentioned in the manuscript. These results suggest that the predicted molecules could bind effectively at the EGFR active site without losing their structural integrity and might be explored to identify novel EGFR inhibitors. The individual energetic contribution of these interacting amino acids was then computed using the per-residue decomposition method. The residues showing negative contribution suggest favourable interactions and *vice versa*. As evident from Fig. 9(c) and (f), it is clear that Leu718, F723, Gly724, Val726, Ala743, Leu788, Thr790, Met793, Leu844 and Thr854 showed negative energetic contributions for all the complexes suggesting their role in effective EGFR inhibition. Lys745 and Asp855 show many crucial interactions, but their positive, energetic contribution suggests that they might form unfavourable interactions, possibly due to steric clashes due to structural constraints. All these structural experiments suggest that the molecules predicted as active EGFR inhibitors by the EGFR^{AP} app have comparable *in silico* profiles as the standard EGFR rugs considered in the study. They further reinforce the findings of the developed tool, suggesting its applicability in rational drug design.

This tool can be utilized to develop novel medications to manage a variety of cancers. While the EGFR^{AP} tool yields encouraging results, it is vital to highlight a few limitations. First, the tool's performance is based on the quality and diversity of the training dataset, which may cause biases or limit the predictions for novel compounds outside the chemical space in the publically available datasets. Although the present tool is highly computationally efficient, its applicability to researchers with fewer computational capabilities is limited. In future versions of EGFR^{AP}, a multi-target approach will be adopted to predict the activities of related receptor tyrosine kinases. Additional datasets from other small molecule databases will be integrated to ensure broader coverage of diverse chemical scaffolds.

4. Conclusions

This work presents the EGFR^{AP} tool to predict the pIC$_{50}$ values for ligands against EGFR. The tool was developed using various machine-learning modules in Python. It is based on a learning model developed using an efficient protocol. EGFR^{AP} gives the user an advantage in computing more than 1400 2D molecular descriptors for the input molecules and predicting their biological activities against EGFR. The model performed well in the learning and validation stages; thus, its findings are reliable. It showed a Pearson correlation coefficient (r) of 0.82 between the experimental and predicted pIC$_{50}$ values of the external test set, highlighting its efficiency. The structure-based drug design experiments further reinforced the accuracy of the EGFR^{AP} predictions, where the application's leads showed a comparable binding profile as the standard EGFR drugs considered in the study. EGFR^{AP} is light and can be run on most machines with the minimum hardware and software requirements. These findings suggest that the EGFR^{AP} tool can identify novel lead compounds or aid in modulating the activity of the existing EGFR drugs. The findings presented here can be of significant interest to the medicinal chemists developing novel cancer medications.

Consent for publication

All the authors have read and approved the manuscript for publication in all respects.

Data availability

The relevant code and files required to run the tool can be accessed using the following URL: **https://github.com/amarinderthind/EGFR-ap**. The supplementary files S1–S12 and the Supplementary Figure file can be accessed using the link: **https://github.com/amarinderthind/EGFR-ap/tree/Supplementary-Data**.

Author contributions

Ashish Gupta: planned the protocol, performed data curation and model building/validation, and prepared the MS. Amarinder Singh Thind: performed feature selection and provided expertise in the application of various machine learning protocols and proofread the MS. Rituraj Purohit: conceptualized the entire study and supervised the project.

Conflicts of interest

The authors declare no competing interests.

Acknowledgements

The authors gratefully acknowledge the Director, CSIR-Institute of Himalayan Bioresource Technology (IHBT), Palampur, for providing the facilities to carry out this study and the Indian Council of Medical Research (ICMR), New Delhi, India for providing financial support to Dr. Ashish (ICMR-RA-2021-14168). The CSIR support in the form of project MLP0201 for bioinformatics studies is highly acknowledged. This manuscript represents CSIR-IHBT communication no. 5796.

References

1 M. J. Wieduwilt and M. M. Moasser, *Cell. Mol. Life Sci.*, 2008, **65**, 1566–1584.

2 D. A. Sabbah, R. Hajjo and K. Sweidan, *Curr. Top. Med. Chem.*, 2020, **20**, 815–834.

3 P. Wee and Z. Wang, *Cancers*, 2017, **9**(5), 52–96.

4 S. Talukdar, L. Emdad, S. K. Das and P. B. Fisher, in *Receptor Tyrosine Kinases*, ed. R. Kumar and P. B. B. T.-A. in C. R. Fisher, Academic Press, 2020, vol. 147, pp. 161–188.

5 B. Sharma, V. J. Singh and P. A. Chawla, *Bioorg. Chem.*, 2021, **116**, 105393.

6 M. E. Bahar, H. J. Kim and D. R. Kim, *Signal Transduction Targeted Ther.*, 2023, **8**, 455.

7 C.-H. Yang, H.-C. Chou, Y.-N. Fu, C.-L. Yeh, H.-W. Cheng, I.-C. Chang, K.-J. Liu, G.-C. Chang, T.-F. Tsai, S.-F. Tsai, H.-P. Liu, Y.-C. Wu, Y.-T. Chen, S.-F. Huang and Y.-R. Chen, *Biochim. Biophys. Acta, Mol. Basis Dis.*, 2015, **1852**, 1540–1549.

8 Y.-P. Liu, C.-C. Zheng, Y.-N. Huang, M.-L. He, W. W. Xu and B. Li, *MedComm*, 2021, **2**, 315–340.

9 K. C. Cuneo, M. K. Nyati, D. Ray and T. S. Lawrence, *Pharmacol. Ther.*, 2015, **154**, 67–77.

10 A. Ayati, S. Moghimi, S. Salarinejad, M. Safavi, B. Pouramiri and A. Foroumadi, *Bioorg. Chem.*, 2020, **99**, 103811.

11 Q. Pan, Y. Lu, L. Xie, D. Wu, R. Liu, W. Gao, K. Luo, B. He and Y. Pu, *Mol. Pharmaceutics*, 2023, **20**, 829–852.

12 J. García-Foncillas, Y. Sunakawa, D. Aderka, Z. Wainberg, P. Ronga, P. Witzler and S. Stintzing, *Front. Oncol.*, 2019, **9**, 849.

13 S. Ketzer, K. Schimmel, M. Koopman and H.-J. Guchelaar, *Clin. Pharmacokinet.*, 2018, **57**, 455–473.

14 R. Mariam Raju, J. Joy A, R. Nulgumnalli Manjunathaiah, A. Justin and B. R. Prashantha Kumar, *Results Chem.*, 2024, **7**, 101490.

15 R. Shah and J. F. Lester, *Clin. Lung Cancer*, 2020, **21**, e216–e228.

16 F. Morgillo, C. M. Della Corte, M. Fasano and F. Ciardiello, *ESMO Open*, 2016, **1**, e000060.

17 M. Ge, Y. Zhu, M. Wei, H. Piao and M. He, *Biochim. Biophys. Acta, Rev. Cancer*, 2023, **1878**, 188996.

18 E. da S. Santos, K. A. B. Nogueira, L. C. C. Fernandes, J. R. P. Martins, A. V. F. Reis, J. de B. V. Neto, I. J. da S. Júnior, C. Pessoa, R. Petrilli and J. O. Eloy, *Int. J. Pharm.*, 2021, **592**, 120082.

19 K. S. Alharbi, M. A. Javed Shaikh, O. Afzal, A. S. Alfawaz Altamimi, W. H. Almalki, S. I. Alzarea, I. Kazmi, F. A. Al-Abbasi, S. K. Singh, K. Dua and G. Gupta, *Chem.-Biol. Interact.*, 2022, **366**, 110108.

20 L. Hu, M. Fan, S. Shi, X. Song, F. Wang, H. He and B. Qi, *Eur. J. Med. Chem.*, 2022, **227**, 113963.

21 A. Gupta, N. Chaudhary and P. Aparoy, *Int. J. Biol. Macromol.*, 2018, **119**, 352–359.

22 A. Gupta, N. Chaudhary, K. R. Kakularam, R. Pallu and A. Polamarasetty, *PLoS One*, 2015, **10**, e0134472.

23 H. Chang, Z. Zhang, J. Tian, T. Bai, Z. Xiao, D. Wang, R. Qiao and C. Li, *ACS Omega*, 2024, **9**, 2314–2324.

24 H. Nada, A. R. Gul, A. Elkamhawy, S. Kim, M. Kim, Y. Choi, T. J. Park and K. Lee, *ACS Omega*, 2023, **8**, 31784–31800.

25 D. Huo, S. Wang, Y. Kong, Z. Qin and A. Yan, *J. Chem. Inf. Model.*, 2022, **62**, 5149–5164.

26 I. H. Eissa, R. G. Yousef, E. B. Elkaeed, A. A. Alsfouk, D. Z. Husein, I. M. Ibrahim, M. S. Alesawy, H. Elkady and A. M. Metwaly, *PLoS One*, 2023, **18**, e0282586.

27 https://github.com/amarinderthind/EGFR-ap/tree/Supplementary-Data.

28 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 and A. R. Leach, *Nucleic Acids Res.*, 2024, **52**, D1180–D1192.

29 M. Davies, M. Nowotka, G. Papadatos, N. Dedman, A. Gaulton, F. Atkinson, L. Bellis and J. P. Overington, *Nucleic Acids Res.*, 2015, **43**, W612–W620.

30 C. W. Yap, *J. Comput. Chem.*, 2011, **32**, 1466–1474.

31 S. Jamal, A. Grover and S. Grover, *Front. Pharmacol.*, 2019, **10**, 780–792.

32 S. Kansız, M. Azam, T. Basılı, S. Meral, F. A. Aktaş, S. Yeşilbağ, K. Min, A. A. Ağar and N. Dege, *J. Mol. Struct.*, 2022, **1265**, 133477.

33 C. E. Whitehead, E. K. Ziemke, C. L. Frankowski-McGregor, R. A. Mumby, J. Chung, J. Li, N. Osher, O. Coker, V. Baladandayuthapani, S. Kopetz and J. S. Sebolt-Leopold, *Nat. Cancer*, 2024, **5**, 1250–1266.

34 F. C. Bernstein, T. F. Koetzle, G. J. Williams, E. F. J. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi and M. Tasumi, *J. Mol. Biol.*, 1977, **112**, 535–542.

35 B. Gogoi, P. Chowdhury, N. Goswami, N. Gogoi, T. Naiya, P. Chetia, S. Mahanta, D. Chetia, B. Tanti, P. Borah and P. J. Handique, *Mol. Diversity*, 2021, **25**, 1963–1977.

36 *BIOVIA Dassault Systèmes 2016*, Dassault Systèmes, 2016.

37 M. Kurbanova, M. Ashfaq, M. N. Tahir, A. Maharramov, N. Dege, N. Ramazanzade and E. B. Cinar, *J. Struct. Chem.*, 2023, **64**, 437–449.

38 E. Alaman, A. A. Ağar, M. N. Tahir, M. Ashfaq, E. B. Poyraz and N. Dege, *J. Struct. Chem.*, 2023, **64**, 1314–1328.

39 S. Kansiz, N. Dege, S. Ozturk, N. Akdemir, E. Tarcan, A. Arslanhan and E. Saif, *Acta Crystallogr., Sect. E: Crystallogr. Commun.*, 2021, **77**, 138–141.

40 S. Daoui, C. Baydere, F. Akman, F. El Kalai, L. Mahi, N. Dege, Y. Topcu, K. Karrouchi and N. Benchat, *J. Mol. Struct.*, 2021, **1225**, 129180.

41 N. Dege, Ö. Özge, D. Avcı, A. Başoğlu, F. Sönmez, M. Yaman, Ö. Tamer, Y. Atalay and B. Zengin Kurt, *Spectrochim. Acta, Part A*, 2021, **262**, 120072.

42 N. Arumugam, A. I. Almansour, R. S. Kumar, A. J. Mohammad Ali Al-Aizari, S. I. Alaqeel, S. Kansız, V. S. Krishna, D. Sriram and N. Dege, *RSC Adv.*, 2020, **10**, 23522–23531.

43 B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan and M. Karplus, *J. Comput. Chem.*, 1983, **4**, 187–217.

44 M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess and E. Lindah, *SoftwareX*, 2015, **1–2**, 19–25.

45 B. Hess, C. Kutzner, D. Van Der Spoel and E. Lindahl, *J. Chem. Theory Comput.*, 2008, **4**, 435–447.

46 D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark and H. J. C. Berendsen, *J. Comput. Chem.*, 2005, **26**, 1701–1718.

47 J. Huang, S. Rauscher, G. Nawrocki, T. Ran, M. Feig, B. L. de Groot, H. Grubmüller and A. D. MacKerell, *Nat. Methods*, 2017, **14**, 71–73.

48 U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee and L. G. Pedersen, *J. Chem. Phys.*, 1995, **103**, 8577–8593.

49 B. Hess, H. Bekker, H. J. C. Berendsen and J. G. E. M. Fraaije, *J. Comput. Chem.*, 1997, **18**, 1463–1472.

50 N. Chaudhary and P. Aparoy, *Heliyon*, 2020, **6**, e04944.

51 M. Tian, H. Li, X. Yan, J. Gu, P. Zheng, S. Luo, D. Zhangsun, Q. Chen and Q. Ouyang, *Front. Mol. Biosci.*, 2022, **9**, 848353.

52 M. W. Heymans and J. W. R. Twisk, *J Clin. Epidemiol.*, 2022, **151**, 185–188.

53 H. Kang, *Korean J. Anesthesiol.*, 2013, **64**, 402–406.

54 H. Mai, T. C. Le, D. Chen, D. A. Winkler and R. A. Caruso, *Chem. Rev.*, 2022, **122**, 13478–13515.

55 S. A. Atimbire, J. K. Appati and E. Owusu, *Sci. Rep.*, 2024, **14**, 4530.

56 S. N. Kigo, E. O. Omondi and B. O. Omolo, *Sci. Rep.*, 2023, **13**, 17315.

57 S. Kumar, R. Bhowmik, J. M. Oh, M. A. Abdelgawad, M. M. Ghoneim, R. H. Al-Serwi, H. Kim and B. Mathew, *Sci. Rep.*, 2024, **14**, 4868.

58 D. Bonchev and N. Trinajstić, *SAR QSAR Environ. Res.*, 2001, **12**, 213–236.

59 V. Consonni and R. Todeschini, *Recent Advances in QSAR Studies: Methods and Applications*, ed. T. Puzyn, J. Leszczynski and M. T. Cronin, Springer Netherlands, Dordrecht, 2010, pp. 29–102.

60 Z. Wang, L. Mu, H. Miao, Y. Shang, H. Yin and M. Dong, *Energy*, 2023, **275**, 127438.

61 R. Gupta, A. K. Yadav, S. K. Jha and P. K. Pathak, *Int. J. Green Energy*, 2024, **21**, 1853–1873.

62 M. Ghazwani and M. Y. Begum, *Sci. Rep.*, 2023, **13**, 10046.

63 P. Refaeilzadeh, L. Tang and H. Liu, Cross-validation, in *Encyclopedia of Database Systems*, ed. L. Liu and M. T. Özsu, Springer US, Boston, MA, 2009, pp. 532–538.

64 Z. Zhao, L. Xie and P. E. Bourne, *J. Chem. Inf. Model.*, 2019, **59**, 453–462.

65 C.-H. Yun, K. E. Mengwasser, A. V. Toms, M. S. Woo, H. Greulich, K.-K. Wong, M. Meyerson and M. J. Eck, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**, 2070–2075.