

Structural identification of novel pyrimidine derivatives as epidermal growth factor receptor inhibitors using 3D QSAR, molecular docking, and MMGBSA analysis: a rational approach in anticancer drug design

Pradip Jana ¹, Shivangi Agarwal ¹, Varsha Kashaw ², RatneshDas ³, Anshuman Dixit ⁴, and Sushil K. Kashaw ^{1,*}

¹ Department of Pharmaceutical Sciences, Dr. HarisinghGour University (A Central University), Sagar (MP), India

² Sagar Institute of Pharmaceutical Sciences, Sagar-M.P.

³ Department of Chemistry, Dr. HarisinghGour University (A Central University), Sagar (MP), India

⁴ Institute of Life Sciences, Nalco Square, Bhubaneswar, 751023, Odisha, India

Abstract: Non-small cell lung cancer (NSCLC) has evolved into the deadliest in the present scenario. The progression of NSCLC is mainly due to the dysregulation of the tyrosine kinase family's epidermal growth factor receptor (EGFR). Thus, EGFR has been widely studied as a major target in the treatment of NSCLC, but the lack of selectivity and drug resistance limit the use of existing therapeutic agents. Considering the urgent necessity for the advanced development of EGFR inhibitors, we have implemented a three-dimensional structure-activity relationship (3D QSAR), molecular docking, and MMGBSA studies on a series of pyrimidine derivatives. In the 3D QSAR, the comparative molecular field analysis model (CoMFA) was obtained with a correlation coefficient (r^2) = 0.698, cross-validated correlation coefficient (q^2) = 0.541, and predictive r^2 (r^2_{pred}) = 0.509. The comparative molecular similarity indices analysis (CoMSIA) model also displayed similar results with r^2 = 0.72, q^2 = 0.586, and r^2_{pred} = 0.495. The statistical parameters fulfill the acceptability criteria of the models. Docking studies revealed the binding interactions of the pyrimidine derivatives with double mutant EGFR^{L858R/T790M}. Docking scores of the top two selected compounds 29 and 34 were 92.99 and 92.13, respectively. Analyzing 3D QSAR contour plots and docking results reviewed some important structural attributes of EGFR^{L858R/T790M} selective inhibitors, which directed the designing of some new molecules. The designed compounds showed good predictive activity and exhibited higher binding interactions with EGFR^{L858R/T790M} than the reference ligand gefitinib. Moreover, to evaluate the binding of selected top hits from docking and designed compounds, MMGBSA (Molecular Mechanics-Generalized Born Surface Area) analysis was performed, which revealed that the designed compound (N7) showed a good binding affinity with EGFR^{L858R/T790M} (dG = -68.59 kcal/mol) as compared to other compounds. Further, in silico ADME predictions revealed the drug-likeness of the designed compounds. Thus, this work will guide researchers in future developments of pyrimidine derivatives as EGFR inhibitors.

Keywords: EGFR; pyrimidine; 3D QSAR; CoMFA; CoMSIA; Molecular docking; MMGBSA.

Abbreviations:

3D QSAR- Three-Dimensional Quantitative Structure-Activity Relationship

ADME- Absorption Distribution Metabolism Excretion

BBB- Blood Brain Barrier

CoMFA- Comparative Molecular Field Analysis

CoMSIA- Comparative Molecular Similarity Indices Analysis

EGFR- Epidermal Growth Factor Receptor

EGFRwt- Epidermal Growth Factor Receptor wild type

GA- Genetic Algorithm

GOLD- Genetic Optimization for Ligand Docking

LOO- Leave One Out

*Corresponding author: Sushil K. Kashaw

Email address: sushilkashaw@gmail.com

DOI: <http://dx.doi.org/10.13171/mjc02304141691kashaw>

Received February 10, 2023

Accepted March 9, 2023

Published April 14, 2023

MMGBSA-Molecular Mechanics-Generalized Born Surface Area

MSA- Multiple Sequence Alignment

NSCLC- Non-Small Cell Lung Cancer

ONC- Optimal Number of Components

PDB- Protein Data Bank

PLS- Partial Least Square

SEE- Standard Error of Estimate

TKIs- Tyrosine Kinase Inhibitors

1. Introduction

In today's world, non-small cell lung cancer (NSCLC) has become the deadliest form of lung cancer, with an alarming number of cases and death counts¹. In almost 50% of NSCLC patients, epidermal growth factor receptor (EGFR) overexpression and dysregulation of related downstream signaling are reported². The EGFR (a receptor tyrosine kinase family member) controls vital cellular activities such as cell proliferation, migration, differentiation, and apoptosis through signaling pathways³. The signaling pathway gets triggered when a ligand molecule binds at the specific binding site, stabilizes the dimer structure of the receptor, and auto-phosphorylation of tyrosine residue within the intracellular kinase domain takes place. Therefore, inhibition of the EGFR tyrosine kinase receptor has evolved as a promising approach to treating NSCLC^{4,5}. But the incidence of drug resistance due to several mutations of EGFR receptor limits the use of available drugs. The first-generation EGFR tyrosine kinase inhibitors (EGFR-TKIs) gefitinib and erlotinib, approved by US Food and Drug Administration in 2002 and 2004⁶, have shown effective tumor regression against NSCLC, especially in patients with EGFR-sensitive mutants (e.g. EGFR^{L858R})⁷⁻⁹. Despite clinical significance, the efficacy of first-generation inhibitors was lost because of acquired T790M point mutation (Threonine⁷⁹⁰→Methionine⁷⁹⁰) in EGFR, constituting almost 50% of clinically developed resistance cases⁶. Though this acquired drug resistance was overcome with second-generation irreversible EGFR-TKIs such as afatinib and neratinib, dose-limiting toxicities and side effects like diarrhea and skin rashes were also observed simultaneously due to loss of selectivity on EGFR wild-type (EGFRwt)^{10,11}. Further efforts have been made to overcome drug resistance related to T790M mutation by developing third-generation EGFR-TKIs (e.g. osimertinib). The third-generation EGFR-TKIs have shown covalent binding with Cys797 residue and exhibited good selectivity over EGFRwt, reducing the chances of side effects^{12,13}.

Even though promising results were reported, third-generation inhibitors certainly bring newly acquired resistances like C797S mutation in EGFR⁶.

Therefore, it is of great value to develop novel EGFR inhibitors with better selectivity towards EGFR. Several works have reported substituted pyrimidine derivatives as EGFR inhibitors exhibiting good selectivity against mutant EGFR. Chang and co-workers have synthesized a series of pyrimido-pyrimidine derivatives and showed potential inhibitory activity against EGFR mutants⁶. Ji and the group synthesized a novel series of 6-alkenylamides substituted 4-anilinothieno pyrimidines and evaluated them as irreversible inhibitors. Many of these compounds exhibited good potency against EGFR wild type and EGFR^{T790M/L858R} mutant type¹⁴. Recently, Zhang et al. have designed and developed a novel series of pyrido-pyrimidine derivatives to overcome acquired drug resistance, showing exciting results that can be accounted for further¹⁵.

To better understand the structural requirements of pyrimidine derivatives to acquire more selective and potent EGFR inhibitors, we have utilized a combined molecular modeling strategy, including 3D QSAR, molecular docking, and MMGBSA analysis, with *in silico* ADME predictions. CoMFA and CoMSIA-based 3D QSAR studies were performed to develop statistically significant models, which will further help identify important structural attributes. Molecular docking studies were conducted to understand the most likely binding interactions between the compounds and the EGFR. To compute the binding free energy of docked compounds, MMGBSA analysis was performed. While *in silico*, ADME studies assisted in finding out the drug-likeness of the molecules. The valuable structural information of pyrimidine derivatives identified from our research is believed to help design more selective and potent EGFR-TKIs.

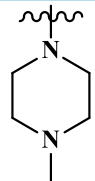
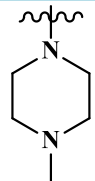
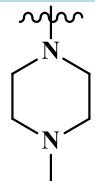
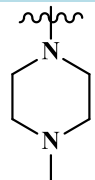
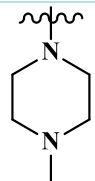
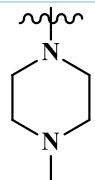
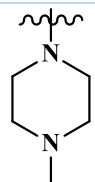
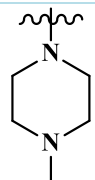
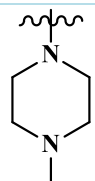
2. Materials and methods

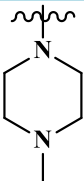
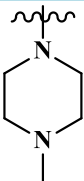
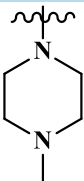
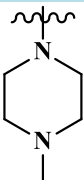
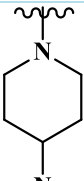
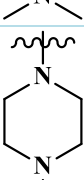
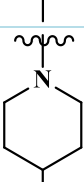
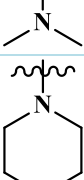
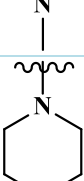
2.1. Dataset

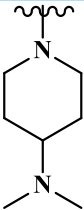
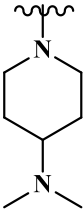
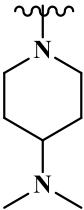
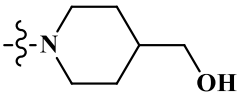
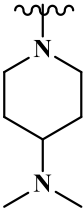
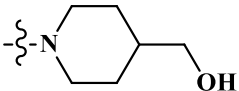
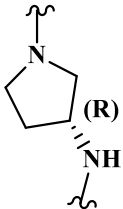
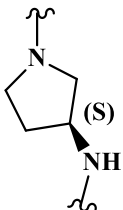
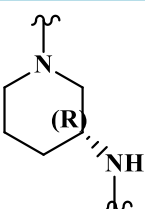
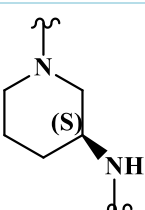
A total of 38 pyrimidine derivatives, along with their inhibitory activities, were taken from the literature^{6,15}. All the biological activity values (IC₅₀) were represented as pIC₅₀ (-logIC₅₀). This is presented in Table 1.

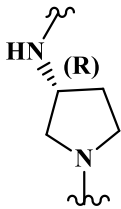
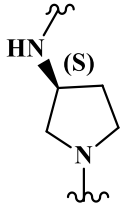
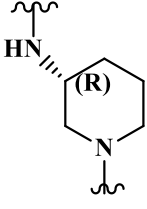
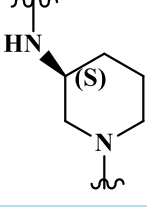
Table 1. EGFR^{L858R/T790M} inhibitors with their actual and predicted pIC_{50} values.

| No ^a | <i>R</i> ₁ | <i>R</i> ₂ | <i>R</i> ₃ | Actual <i>pIC</i> ₅₀ | Pred- <i>pIC</i> ₅₀ | |
|-----------------|--|-----------------------|-----------------------|---------------------------------|--------------------------------|--------|
| | | | | | CoMFA | CoMSIA |
| | | | | | | |
| | <p style="text-align: center;">Template A (1-20) Template B (21-30) Template C (31-38)</p> | | | | | |
| 01* | | 2'-MeO | methyl | 8.5031 | 8.0995 | 8.1266 |
| 02 | | 2'-MeO | methyl | 7.9393 | 8.4423 | 8.4229 |
| 03* | | 2'-MeO | methyl | 7.4385 | 8.1506 | 8.1021 |
| 04* | | 2'-MeO | methyl | 7.5399 | 8.031 | 8.1359 |
| 05 | | 2'-MeO | methyl | 8.2865 | 8.4095 | 8.4492 |
| 06 | | 2'-MeO | methyl | 8.284 | 8.1704 | 8.0999 |
| 07 | H | H | methyl | 7.5482 | 7.794 | 7.8832 |
| 08 | | H | methyl | 8.9245 | 8.4697 | 8.5152 |

| | | | | | | |
|-----|---|-----------|--------------|--------|--------|--------|
| 09 |  | 3'-MeO | methyl | 8.6904 | 8.439 | 8.491 |
| 10 |  | 2'-EtO | methyl | 8.1487 | 8.0226 | 8.0669 |
| 11 |  | 2'-Pr(i)O | methyl | 7.4401 | 7.9706 | 8.019 |
| 12 |  | 2'-Me | methyl | 8.3788 | 8.4792 | 8.5134 |
| 13 |  | 2'-MeO | i-propyl | 8.3546 | 8.3575 | 8.2498 |
| 14 |  | 2'-MeO | cyclo-propyl | 8.2204 | 8.5873 | 8.5618 |
| 15* |  | 2'-MeO | phenyl (Ph) | 9.0088 | 8.1239 | 8.1099 |
| 16 |  | 2'-MeO | 2-naphthyl | 8.6126 | 8.0517 | 8.0914 |
| 17 |  | 2'-MeO | benzyl | 9.0315 | 8.1615 | 8.1105 |

| | | | | | | |
|-----|---|-----------------|--------------------|--------|--------|--------|
| 18* |  | 2'-MeO | 4-biphenyl | 8.3335 | 8.1855 | 8.2439 |
| 19 |  | 2'-MeO | 4-phenoxyphenyl | 8.4473 | 8.3802 | 8.51 |
| 20* |  | 2'-MeO | 4-benzoyloxyphenyl | 8.2 | 8.4825 | 8.5331 |
| 21 |  | pyrrolidin-1-yl | -NHPh | 7.0706 | 7.5579 | 7.5389 |
| 22 |  | pyrrolidin-1-yl | -NHPh | 7.1831 | 7.1292 | 7.1828 |
| 23 |  | piperidin-1-yl | -NHPh | 7.3468 | 7.1093 | 7.1575 |
| 24 |  | piperidin-1-yl | -NHPh | 7.6345 | 7.1093 | 7.1575 |
| 25* |  | morpholine-4-yl | -NHPh | 7.2741 | 7.6586 | 7.6124 |
| 26 |  | morpholine-4-yl | -NHPh | 7.2388 | 7.1207 | 7.1756 |

| | | | | | | |
|-----------------------|---|---|---------------|--------------------------------|------------------------------|---------------|
| 27 |  | morpholine-4-yl | -NHPH 4-F | 7.5406 | 7.1348 | 7.198 |
| 28 |  | morpholine-4-yl | -NHPH-2,4-diF | 7.5229 | 7.7083 | 7.6296 |
| 29 |  |  | -NHPH | 7.3556 | 7.1293 | 7.1165 |
| 30 |  |  | -NHPH-4-F | 7.4685 | 7.4085 | 7.3881 |
| <i>No^a</i> | <i>Linker</i> | | | <i>Actual pIC₅₀</i> | <i>Pred-pIC₅₀</i> | |
| | | | | | <i>CoMFA</i> | <i>CoMSIA</i> |
| 31 |  | | | 7.4365 | 7.7211 | 7.6179 |
| 32 |  | | | 7.5768 | 7.7211 | 7.6179 |
| 33 |  | | | 6.786 | 7.1137 | 7.0886 |
| 34* |  | | | 6.4424 | 7.1137 | 7.0886 |

| | | | | |
|-----|--|--------|--------|--------|
| 35* |  | 6.9731 | 7.1313 | 7.0547 |
| 36 |  | 6.6232 | 7.1313 | 7.0547 |
| 37 |  | 6.8854 | 7.1462 | 7.0613 |
| 38* |  | 7.6321 | 7.1462 | 7.0613 |

*Compound number, *test set compounds.

2.2. QSAR study

2.2.1. Dataset division

The collected compounds were divided into two sets, i.e., training and testing. The training set comprises 28 compounds (75%) for model building. Whereas 10 compounds (25%) were kept in the test set to predict the build models¹⁶.

2.2.2. Energy minimization and alignment

In order to obtain the best-desired conformer of each molecule, energy minimization was performed using SYBYL X-2.1.1 software. During energy minimization, the force field was set to 'Tripos', 'Powell gradient' was chosen with a convergence of 0.005 kcal/(mol*Å), and a maximum iteration count of 1000 was applied. The Gasteiger-Huckel method was selected for the calculation of partial atomic charge. The lowest energy conformation of each

molecule was used for QSAR studies¹⁷. In the next step, molecular alignment was done. Molecular alignment is one of the most crucial steps while building 3D QSAR models. The most popular approaches for molecular alignment are maximum standard structure alignment and distill rigid alignment. In top typical structure alignment, other molecules in the dataset are aligned based on the available common structures concerning the template. While in distill rigid alignment, molecules are aligned according to their steric and electrostatic field on the template molecule¹⁸.

In our study distill rigid alignment type was applied to the training set for molecular alignment. Compound 17, the most potent molecule in the dataset, was used as the template molecule. The structural alignment of the compounds is displayed in Figure 1.

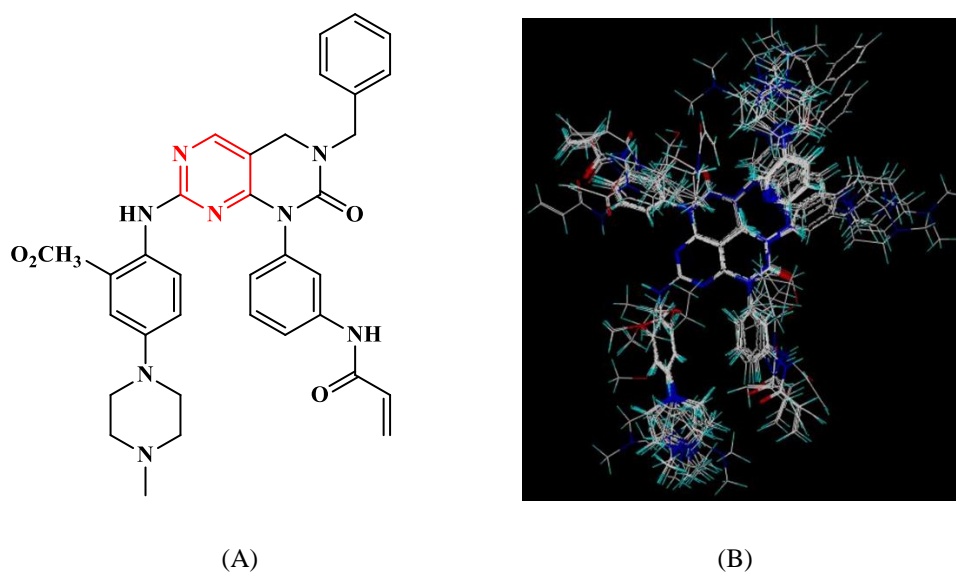


Figure 1. Molecular alignment. (A) template for alignment: compound 17, the red colored part of the structure is the core for alignment; (B) distill rigid alignment of training set compounds.

2.2.3. Comparative molecular field analysis (CoMFA)

In CoMFA, steric fields were calculated utilizing Lennard-Jones potentials, whereas Coulombic potentials were used to determine electrostatic fields. Aligned compounds were enclosed in a 3D cubic lattice with a grid spacing of 2.0 Å. The automatically generated grid points use an sp^3 carbon atom probe to calculate fields.

2.2.4. Comparative molecular similarity indices analysis (CoMSIA)

The CoMSIA method computed similarity indices at different points in a regularly spaced grid for the aligned training set compounds. It has multiple benefits over the CoMFA method, like robustness, no application of arbitrary cutoffs, and effortlessly interpretable contour maps. This method comprises the calculation of five fields (such as steric, electrostatic, hydrophobic, hydrogen bond donor, and hydrogen bond acceptor). Similar to CoMFA, grid spacing was maintained at 2 Å.

2.2.5. Partial least square (PLS) analysis and internal validation

The Partial Least Square method was applied to training set molecules to determine the correlation between the QSAR models and biological activities. The reliability of the developed models was assessed by implementing the leave-one-out (LOO) internal validation method. The LOO method measures the cross-validated correlation coefficient (q^2) and the optimal number of components (ONC). While evaluating a model, other parameters considered are the square of non-cross-validation coefficient or squared correlation coefficient (r^2) and standard error of estimate (SEE)¹⁹.

2.2.6. External validation

The predictive ability of developed CoMFA and CoMSIA models was assessed with the help of test set compounds. The predictive factor r^2 (r_{pred}^2) was calculated considering the predicted activity of the test set compounds utilizing the following formula:

$$r_{pred}^2 = \frac{x - y}{x}$$

Where x = Total of squared deviations between the activities of test set compounds and the mean activity of the training set compounds; y = Total of squared deviations between actual and predicted activities of test assigned compounds²⁰.

2.3. Molecular docking

Molecular docking was executed to explore appropriate binding interactions and poses employing GOLD (Genetic Optimization for Ligand Docking) 5.2 software. The target protein EGFR L858R/T790M crystal structure was retrieved from Protein Data Bank (PDB ID: 4I22)²¹. In the protein preparation phase, polar hydrogens and gasteiger charges were added. The co-crystallized ligand was removed from the ligand binding site, all bound water molecules were deleted, and the binding site was specified using the 'GOLD setup'. Previously saved multi-MOL2 files of energy-minimized pyrimidine derivatives were used as ligands for the docking studies. To validate the docking protocol, a co-crystallized ligand (gefitinib) was extracted and re-docked in the same binding site of the EGFR protein. Then, the pyrimidine-based EGFR inhibitors were docked into the ATP binding cavity of the kinase domain of the protein. The Genetic Algorithm (GA) run was kept as 10, and other parameters were kept as default in GOLD settings. Two fitness functions were applied, 'ChemPLP' as the scoring function and 'Chemscore

as rescoring function. Docking results were visualized in GOLD as well as in PyMOL viewer. 15 newly designed compounds were also docked at the same binding pocket using previously optimized parameters.

Docking was performed with GOLD 5.2 software. GOLD uses default scoring functions such as ChemPLP and Chemscore, which are obtained as positive values. In ChemPLP, the PLP function (fPLP) is used to model steric complementarity between protein and ligand. In addition to distance and angle dependence, hydrogen bonding and metal terms are also taken into account, which is represented as follows:

$$\text{fitnessPLP} = -(\text{wPLP} \cdot \text{fPLP} + \text{Wlig-clash} \cdot \text{f lig-clash} + \text{wlig-tors} \cdot \text{flig-tors} + \text{fchem-cov} + \text{wprot} \cdot \text{fchem-prot} + \text{wcons} \cdot \text{fcons})$$

$$\text{fitnessChemPLP} = \text{fitnessPLP} - (\text{fchem-hb} + \text{fchem-cho} + \text{fchem-met})$$

The overall score is the negative value of the sum of the component energy terms. The highest fitness scores are, therefore, the best.

Thus, GOLD docking scores are represented as a positive value but are considered the negative value of the sum of all components.

2.4. Multiple sequence alignment

Multiple sequence alignment (MSA) is an alignment of three or more biological sequences, usually protein, DNA, or RNA. MSA is often used to assess the sequence conservation of protein domains, tertiary and secondary structures, and even individual amino acids or nucleotides. Our study of amino acid sequences of ten different species was taken from uniprot (<http://www.uniprot.org/>). The organisms whose amino acid sequences of EGFR protein were used include *Homo sapiens* (Human), *Drosophila melanogaster* (Fruit fly), *Mus musculus* (Mouse), *Macaca mulatta* (Rhesus macaque), *Rattus norvegicus* (Rat), *Mesocricetus auratus* (Golden hamster), *Xenopus tropicalis* (Frog), *Danio rerio* (Zebrafish), *Pelodiscus sinensis* (Turtle), and *Meleagris gallopavo* (Wild turkey). The sequences were obtained in “.fasta” format and aligned in the PRALINE web server (<http://www.ibi.vu.nl/programs/pralinewww/>) in search of conserved residues of EGFR protein among different species. PRALINE (PRofileALignment) performs multiple alignments using a progressive approach. The BLOSUM62 matrix did scoring. The conservation is presented by color code, where the scale denotes 0 for the least conserved residue and 10 for the most conserved residue. The conserved

residues are considered to be functionally essential residues ²².

2.5. Binding free energy analysis

The binding free energy calculations were performed for top selected hits from docking and the designed compounds using the MMGBSA (Molecular Mechanics-Generalized Born Surface Area) method with the help of the Prime module of Schrodinger Maestro 9.3. The calculations were performed on docked protein-ligand complexes of top-ranked ligands. The binding energy is calculated as per equation 1.

$$\text{DG bind} = \text{E_complex (minimized)} - \text{E_ligand (minimized)} - \text{E_receptor (minimized)} \quad (1)$$

Prime calculations combine the OPLS_2005 force field, VSGB solvation model for polar solvation, and a nonpolar solvation term. Thus, the calculation uses various energy components (Eq. 2)

$$\text{DG bind} = \Delta\text{EMM} + \Delta\text{Gsolv} + \Delta\text{GSA} \quad (2)$$

Where,

E_{MM} = molecular mechanical energy

G_{solv} = polar contribution towards solvation energy

ΔGSA = non-polar solvation term

The prime outputs MMGBSA free energy of binding (Prime MMGBSA DG bind) of the selected complexes.

2.6. In silico ADME and drug-likeness screening

In silico predictions can be significant in screening early hits and can be applied before the laboratory synthesis of the designed compounds ²³. ADME properties, which constitute the pharmacokinetic profile of a drug molecule, are essential in evaluating its pharmacodynamic activities. In silico predictions of newly designed compounds were performed with the SwissADMEonline prediction tool (<http://www.swissadme.ch/>) ²⁴.

Compounds were converted into their canonical ‘SMILE’ format and put in the ‘SwissADME’ server. The software predicts the physicochemical properties, lipophilicity, water solubility, pharmacokinetics, drug-likeness, and synthetic accessibility of the compounds.

3. Results and Discussion

3.1. Results of QSAR study (CoMFA and CoMSIA)

The statistical parameters obtained in the developed CoMFA and CoMSIA models are presented in Table 2.

Table 2. Statistical parameters of the comparative molecular field analysis (CoMFA) and molecular similarity indices analysis (CoMSIA) models.

| Model | q ² | r ² | ONC | SEE | r ² _{pred} | Relative contribution | | | | |
|---------------|----------------|----------------|-----|-------|--------------------------------|-----------------------|-------|-------|-------|-------|
| | | | | | | S | E | D | H | A |
| <i>CoMFA</i> | | | | | | | | | | |
| S | 0.519 | 0.688 | 1 | 0.378 | 0.551 | 1.000 | - | - | - | - |
| E | 0.526 | 0.683 | 1 | 0.376 | 0.429 | - | 1.000 | - | - | - |
| SE | 0.541 | 0.698 | 1 | 0.360 | 0.509 | 0.498 | 0.502 | - | - | - |
| <i>CoMSIA</i> | | | | | | | | | | |
| SEH | 0.605 | 0.728 | 1 | 0.348 | 0.289 | 0.268 | 0.403 | - | 0.329 | - |
| SEHD | 0.602 | 0.729 | 1 | 0.348 | 0.585 | 0.175 | 0.335 | 0.227 | 0.264 | - |
| SEHA | 0.587 | 0.719 | 1 | 0.354 | 0.016 | 0.203 | 0.303 | - | 0.245 | 0.250 |
| SEDA | 0.603 | 0.815 | 2 | 0.293 | 0.320 | 0.197 | 0.317 | 0.237 | - | 0.240 |
| EHDA | 0.593 | 0.821 | 2 | 0.288 | 0.293 | - | 0.309 | 0.222 | 0.242 | 0.226 |
| SEHDA | 0.586 | 0.720 | 1 | 0.353 | 0.495 | 0.159 | 0.244 | 0.191 | 0.193 | 0.430 |

q²: cross-validated correlation coefficient; r²: non-cross-validated correlation coefficient; ONC: optimal number of components; SEE: standard error of estimate; r²_{pred}: predictive correlation coefficient; S: steric fields; E: electrostatic fields; D: hydrogen-bond donor fields; H: hydrophobic fields; A: hydrogen-bond acceptor fields.

As shown in Table 2, two descriptor fields of CoMFA, steric (S) and electrostatic (E) were used in all three possible combinations (S, E, and SE) to build the models. While in CoMSIA, 'S', 'E', hydrophobic (H), hydrogen bond donor (D), and hydrogen bond acceptor (A) fields were used in combination. The statistical parameters of most of the models in Table 2 meet the desired internal validation criteria indicating the developed models' acceptance. Further, the generated models' predictive correlation coefficient (r²_{pred}) was also calculated with the predicted activities of the respective test sets. The CoMFA model generated with 'S' field and the CoMSIA model generated with a combination of 'SEHD' fields gave maximum external predictive ability (r²_{pred} 0.551 and 0.585, respectively). While CoMFA model developed with 'SE' fields combination and CoMSIA model with the 'SEHDA' fields combination also showed good predictive ability (r²_{pred} 0.509 and 0.495 respectively). In the latter two CoMFA and CoMSIA models, the best

possible combination of fields was applied and thus considered further. In Table 2, we can see these two models exhibited good q² (0.541 and 0.586), r² (0.698 and 0.72) along with a considerable SEE value (0.36 and 0.353). In the selected CoMFA model, the relative contributions of 'S' and 'E' fields were 0.498 and 0.502, respectively. It indicates a greater contribution of 'E' field. Contributions of 'S', 'E', 'H', 'D', and A fields in the selected CoMSIA model were 0.159, 0.244, 0.191, 0.193, and 0.213, respectively. It was found that the E field has the highest contribution in the CoMSIA model, similar to CoMFA model. The scatter plots of actual versus predicted pIC₅₀ values of training and test sets for the selected QSAR models are presented in Figure 2. Both models fit nicely along the diagonal line visible in the plots. The predictive power of the constructed models was also found to be satisfactory. The above results signify that the constructed models are powerful enough for further prediction of activities.

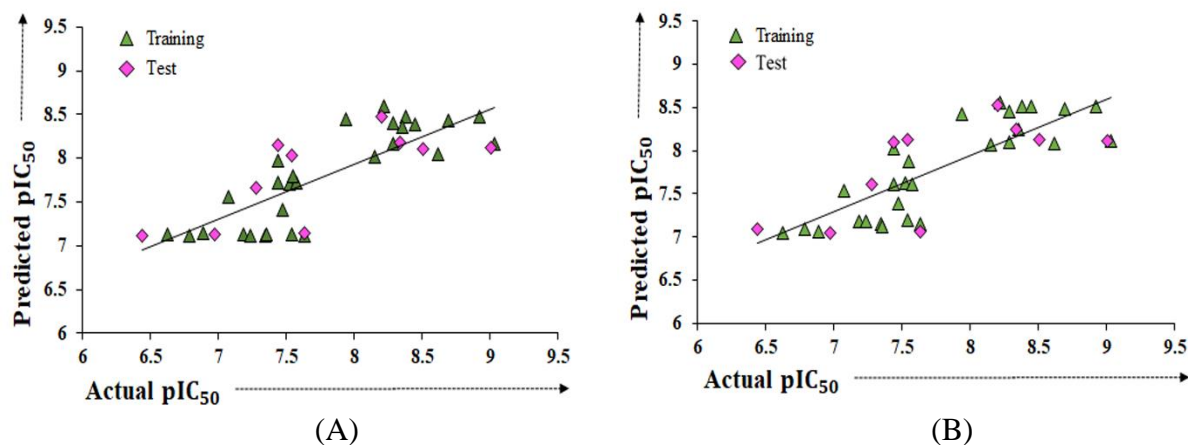


Figure 2. Scatter plot of actual versus predicted pIC₅₀ values of the training set and test set compounds, for (A) CoMFA and (B) CoMSIA model.

3.1.1. Analysis of CoMFA Contours

The 'S' and 'E' contour maps of the CoMFA model are represented in Figure 3. Compound 17, the template molecule (the most potent molecule in the dataset), was selected to represent contour maps. In the 'S' contour map, green contours indicate regions favorable for steric contributions, while the yellow contours represent unfavorable regions. A large yellow steric contour is seen around the benzyl substituent in CoMFA (Figure 3A), which suggests the unsuitability of steric substitution in this region for

EGFR inhibitory activity. This can explain the more excellent biological activity of compound 15 with phenyl substitution compared to compound 20 with benzyloxyphenyl substituent. The green contour around the piperazine ring and above the pyrimido-pyrimidine indicate the presence of steric bulk in this region is necessary. This may explain why the piperazine ring as R₁ substituent in compound 8 is more potent than compound 7, which contains hydrogen as a substituent.

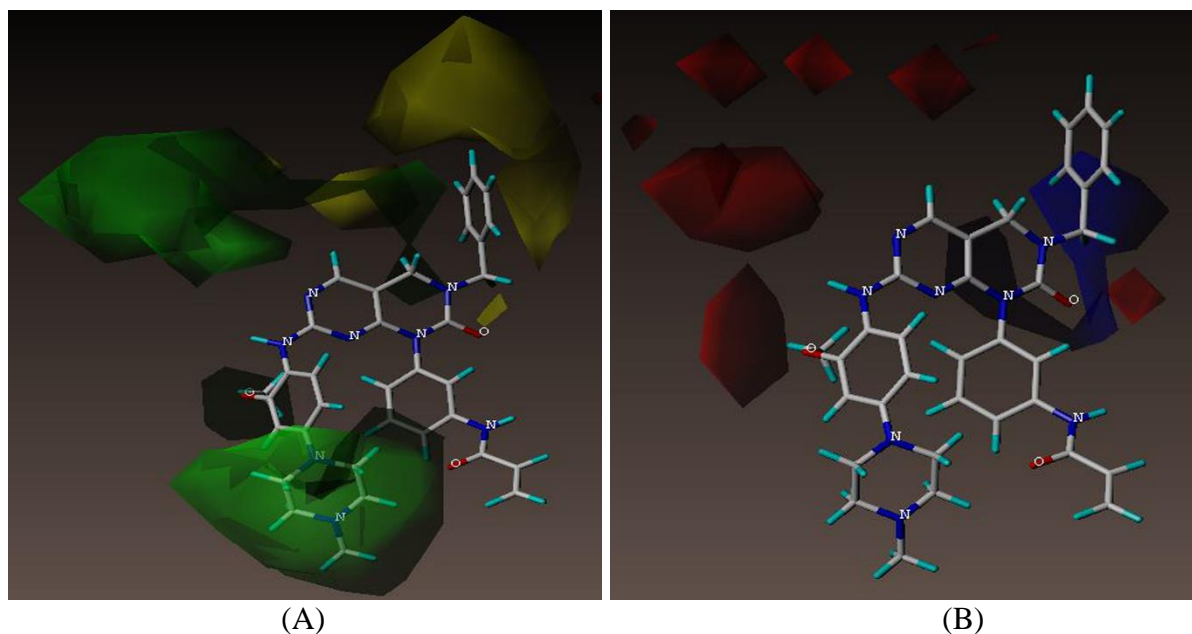


Figure 3. CoMFA contour maps presented with template compound 17. (A) CoMFA steric field contour map (green contour is favored; yellow contour is disfavored); (B) CoMFA electrostatic field contour map (blue contour is preferred for electropositive groups; red contour is favored for electronegative groups).

A green contour near the $-\text{OCH}_3$ group also suggests steric bulk favorability in this position. The blue and red 'E' contours specify regions for suitability of electropositive and electronegative groups, respectively. Blue contours near the benzyl substituent in CoMFA electrostatic map (Figure 3B) suggest the suitability of electropositive groups in this area. Two red contours near the $-\text{OCH}_3$ group indicate electronegative groups are favorable in this region.

3.1.2. Analysis of CoMSIA Contours

The 'S' and 'E' contour maps obtained in the CoMSIA model were the same as the CoMFA contours discussed earlier (Figure 3(A, B)). This section discusses the remaining fields of CoMSIA 'D', 'A', and 'H'. The obtained CoMSIA contour maps are presented in Figure 4.

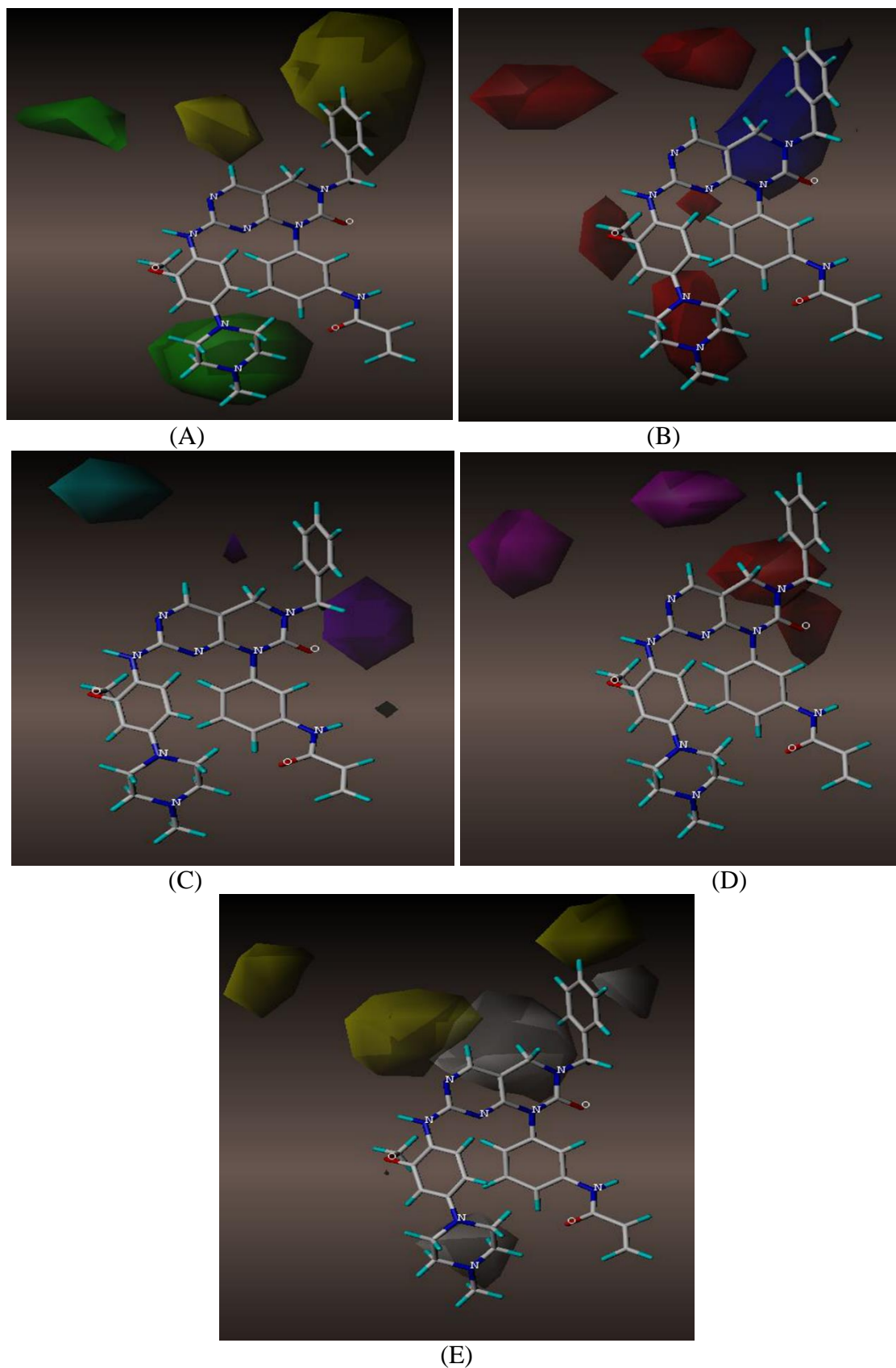


Figure 4. CoMSIA steric field contour map (green contour is favored; the yellow outline is unfavored); (B) CoMSIA electrostatic field contour map (blue contour is preferred for electropositive groups; red contour is chosen for electronegative groups) ; (C) CoMSIA hydrogen donor field contour map (cyan contour is selected; purple contour is unfavored); (D) CoMSIA hydrogen acceptor field contour map (magenta contour is favored; red contour is unfavored); (E) CoMSIA hydrophobic field contour map (yellow contour is favored; white contour is unfavored)

In CoMSIA 'D' field (Figure 4C), cyan contours denote the hydrogen bond donor group suitable area, whereas purple contours represent the hydrogen bond donor unsuitable area. For example, a large cyan contour above the 6th N of the pyrimido-pyrimidine ring suggests that the hydrogen bond donor group is desirable in this region. In contrast, a sizeable purple contour near the benzyl substituent indicates that the hydrogen bond donor in this area is unsuitable for activity.

In 'A' field maps (Figure 4D), the magenta contour is for the favorable hydrogen bond acceptor region, while the red contour denotes regions where hydrogen bond acceptors are unfavorable. Magenta contours seen above 2-methoxy aniline substituent and above 5th position of the pyrimido-pyrimidine ring suggests that the addition of hydrogen bond acceptor group in these regions could increase EGFR inhibitory activity. Conversely, one red contour near 2nd and another near 3rd and 4th positions of the pyrimido-pyrimidine ring indicate that substitutions with hydrogen bond acceptors could decrease activity.

The 'H' contour map is shown in Figure 4E, where yellow contours indicate favorable hydrophobic areas, while white contours suggest unfavorable hydrophobic areas. For example, two large yellow

contours near the -para position of the benzyl substituent and above 6th position of the pyrimido-pyrimidine ring suggest the favorability of hydrophobic groups in these regions. Conversely, a large white contour near 4th and 6th position of the pyrimido-pyrimidine ring indicates that hydrophobic substituents in this position are undesirable. Further white contour, the near-meta position of benzyl substituent, and around piperazine substituent are also disfavored for biological activity.

3.2. Molecular Docking results

Docking studies were performed to understand the probable binding interactions between the pyrimidine derivatives and the EGFR L858R/T790M tyrosine kinase. Before docking dataset compounds, the docking protocol was validated by re-docking the co-crystallized ligand gefitinib into the binding cavity of EGFR. After re-docking, the co-crystallized ligand attained a similar conformation as its original conformation in crystallized protein structure, thereby successfully validating our docking method (Supplementary Figure 1, p. 144). Then, all 38 molecules were docked in the ATP binding pocket of the kinase domain of the EGFR. The docking scores of the dataset molecules and co-crystallized gefitinib are presented in Table 3.

Table 3. Docking scores of the dataset compounds.

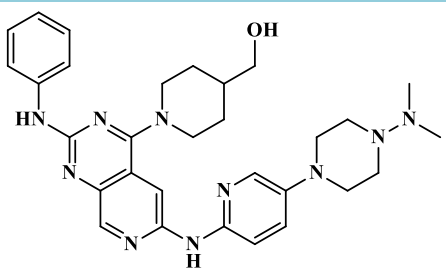
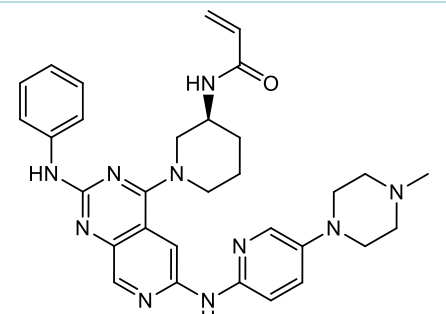
| Compound No. | Docking Score (ChemPLP. Fitness) | Docking Rescore (Chemscore. Fitness) |
|--------------|----------------------------------|--------------------------------------|
| 1 | 78.24 | 22.64 |
| 2 | 73.39 | 21.98 |
| 3 | 79.20 | 27.73 |
| 4 | 76.45 | 20.62 |
| 5 | 67.14 | 24.18 |
| 6 | 77.23 | 26.89 |
| 7 | 81.77 | 26.58 |
| 8 | 75.24 | 27.46 |
| 9 | 79.36 | 23.51 |
| 10 | 73.45 | 22.01 |
| 11 | 76.60 | 25.01 |
| 12 | 74.48 | 25.50 |
| 13 | 83.92 | 21.99 |
| 14 | 84.56 | 25.87 |
| 15 | 87.56 | 29.67 |
| 16 | 82.69 | 26.44 |
| 17 | 88.26 | 26.08 |
| 18 | 88.22 | 26.95 |
| 19 | 90.51 | 26.47 |
| 20 | 87.34 | 21.74 |

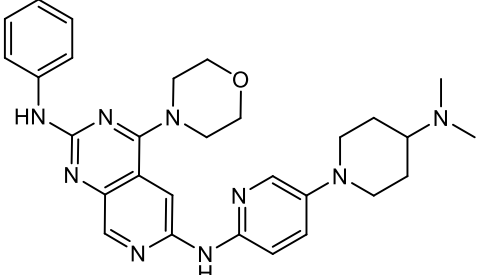
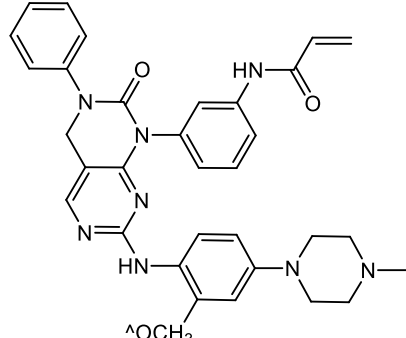
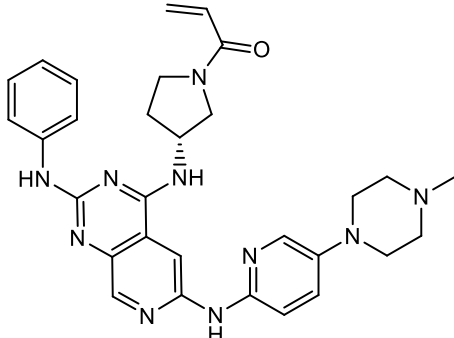
| | | |
|-----------|-------|-------|
| 21 | 78.01 | 29.89 |
| 22 | 85.25 | 31.69 |
| 23 | 86.76 | 32.33 |
| 24 | 86.60 | 32.77 |
| 25 | 79.84 | 31.13 |
| 26 | 89.08 | 30.36 |
| 27 | 79.15 | 25.76 |
| 28 | 73.96 | 23.87 |
| 29 | 92.99 | 31.95 |
| 30 | 91.33 | 27.86 |
| 31 | 89.63 | 31.26 |
| 32 | 85.44 | 25.95 |
| 33 | 89.52 | 25.51 |
| 34 | 92.13 | 31.70 |
| 35 | 87.45 | 30.23 |
| 36 | 86.26 | 30.20 |
| 37 | 84.86 | 28.14 |
| 38 | 88.12 | 31.05 |
| Gefitinib | 79.19 | 29.02 |

Based on docking scores, five top-ranked compounds 29, 34, 26, 15, and 35 were selected (Table 4). Compound 29 possesses the highest docking score

(92.99), and all the selected compounds have a more excellent docking score than gefitinib (79.19) (Table 3).

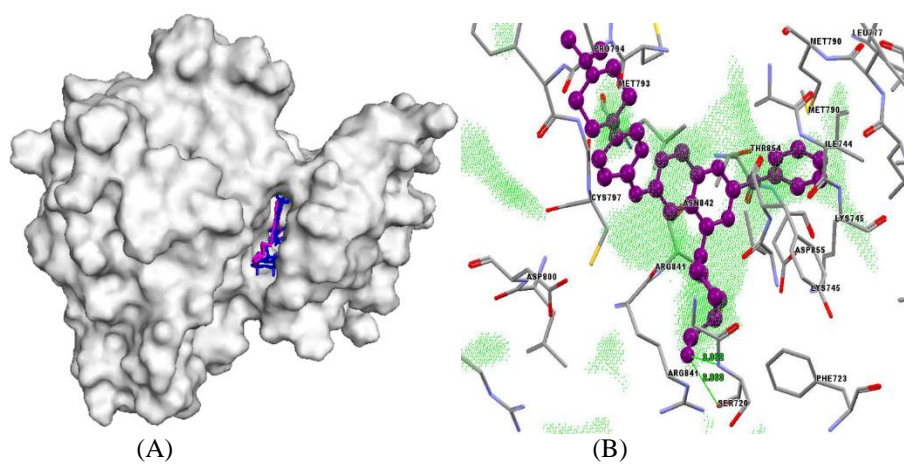
Table 4. Chemical structures and IUPAC names of top five selected dataset compounds.

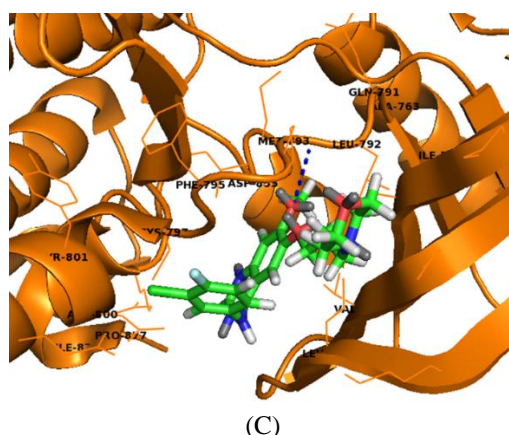
| S. No. | Compound | Chemical Structure & IUPAC Name |
|--------|----------|--|
| 1 | 29 |  <p>(1-(6-(5-(4-(dimethylamino) piperazin-1-yl)pyridin-2-ylamino)-2-(phenylamino)pyrido[3,4-d]pyrimidin-4-yl)piperidin-4-yl)methanol</p> |
| 2 | 34 |  <p>(S)-N-(1-(6-(5-(4-methylpiperazin-1-yl)pyridin-2-ylamino)-2-(phenylamino)pyrido[3,4-d]pyrimidin-4-yl)piperidin-3-yl)acrylamide</p> |

| | | |
|---|----|--|
| 3 | 26 |  <p>N6-(5-(4-(dimethylamino)piperidin-1-yl)pyridin-2-yl)-4-morpholino-N2-phenylpyrido[3,4-d]pyrimidine-2,6-diamine</p> |
| 4 | 15 |  <p>N-(3-(7-(2-methoxy-4-(4-methylpiperazin-1-yl)phenylamino)-2-oxo-3-phenyl-3,4-dihydropyrimido[4,5-d]pyrimidin-1(2H)-yl)phenyl)acrylamide</p> |
| 5 | 35 |  <p>(R)-1-(3-(6-(5-(4-methylpiperazin-1-yl)pyridin-2-ylamino)-2-(phenylamino)pyrido[3,4-d]pyrimidin-4-ylamino)pyrrolidin-1-yl)prop-2-en-1-one</p> |

The molecular surface of the mutant EGFR protein with gefitinib (pink) and compound 29 (blue) is shown in [Figure 5A](#). It is seen that compound 29 well

occupied the binding cavity similar to gefitinib. Compound 29 well occupied the fit points generated by GOLD software ([Figure 5B](#)).





(C)

Figure 5. Docking results. (A) Molecular surface of the protein presented with co-crystallized ligand (pink) and compound 29 (blue) at the binding cavity of the protein; (B) Well occupancy of the appropriate points (green dots) by compound 29 (violet); (C) Binding interactions in between co-crystallized ligand gefitinib and mutant EGFR^{L858R/T790M}.

The hydrogen bond (H-bond) interactions of the selected compounds 29, 34, 26, 15, and 35 and the co-crystallized ligand gefitinib are presented in Table 5. The oxygen atom of co-crystallized ligand gefitinib showed H-bond interaction with the nitrogen atom of

MET 793 amino acid residue of EGFR^{L858R/T790M} at a distance of 3.4 Å (Figure 5C). Compound 29, having the highest docking score (92.99) (Table 5), formed two H-bonds with SER 720.

Table 5. Interaction analysis of selected five top ranked dataset compounds and co-crystallized ligand gefitinib.

| S. No | Compound | Docking Score (ChemPLP) | Rescore (Chemscore) | Interacting residues | Type of interaction | Bond distance (Å) |
|-------|-----------|-------------------------|---------------------|-------------------------------|----------------------------|-------------------------|
| 1 | Gefitinib | 79.19 | 29.02 | MET 793 | H-Bond | 3.4 |
| 2 | 29 | 92.99 | 31.95 | SER 720 | 2H-Bond | 2.863 & 3.062 |
| 3 | 34 | 92.13 | 31.70 | MET 793 | H-Bond | 2.679 |
| 4 | 26 | 89.08 | 30.36 | ASP 855 MET 790 | H-Bond Short contact | 2.931 2.87 |
| 5 | 15 | 87.56 | 29.67 | ARG 841 ASP 855 LYS 745 | H-Bond H-Bond H-Bond | 2.825 2.581 3.047 |
| 6 | 35 | 87.45 | 30.23 | THR 854 SER 720 | H-Bond H-Bond | 3.013 3.001 |

The first H-bond was between –NH group of SER 720 and the hydroxyl group of the ligand at a distance of 2.863 Å, and the second one was between the carbonyl group of SER 720 and the hydroxyl group of the ligand at a distance of 3.062 Å (Figure 6A). Compound 34, having the second highest docking score (92.13) (Table 5), showed H-bond interaction with MET 793 amino acid similar to the co-crystallized ligand gefitinib. The –NH group of compound 34 interacted with the carbonyl group of the hinge region MET 793 residue at a distance of 2.679 Å (Figure 6B). In compound 26, –NH group

formed an H-bond with a carbonyl group of ASP 855 at a distance of 2.931 Å (Figure 6C). Further, Compound 15 also showed H-bond interaction with ASP 855 residue and formed two new H-bonds with ARG 841 and LYS 745 (Figure 6D). While compound 35 formed H-bond with SER 720 similar to compound 29 (Figure 6E). Besides MET 793 residue, new H-bond interactions with residues like ASP 855, SER 720 are expected to reflect in the improvement of the binding specificity of the ligands into the binding pocket of double mutant EGFR^{L858R/T790M}.

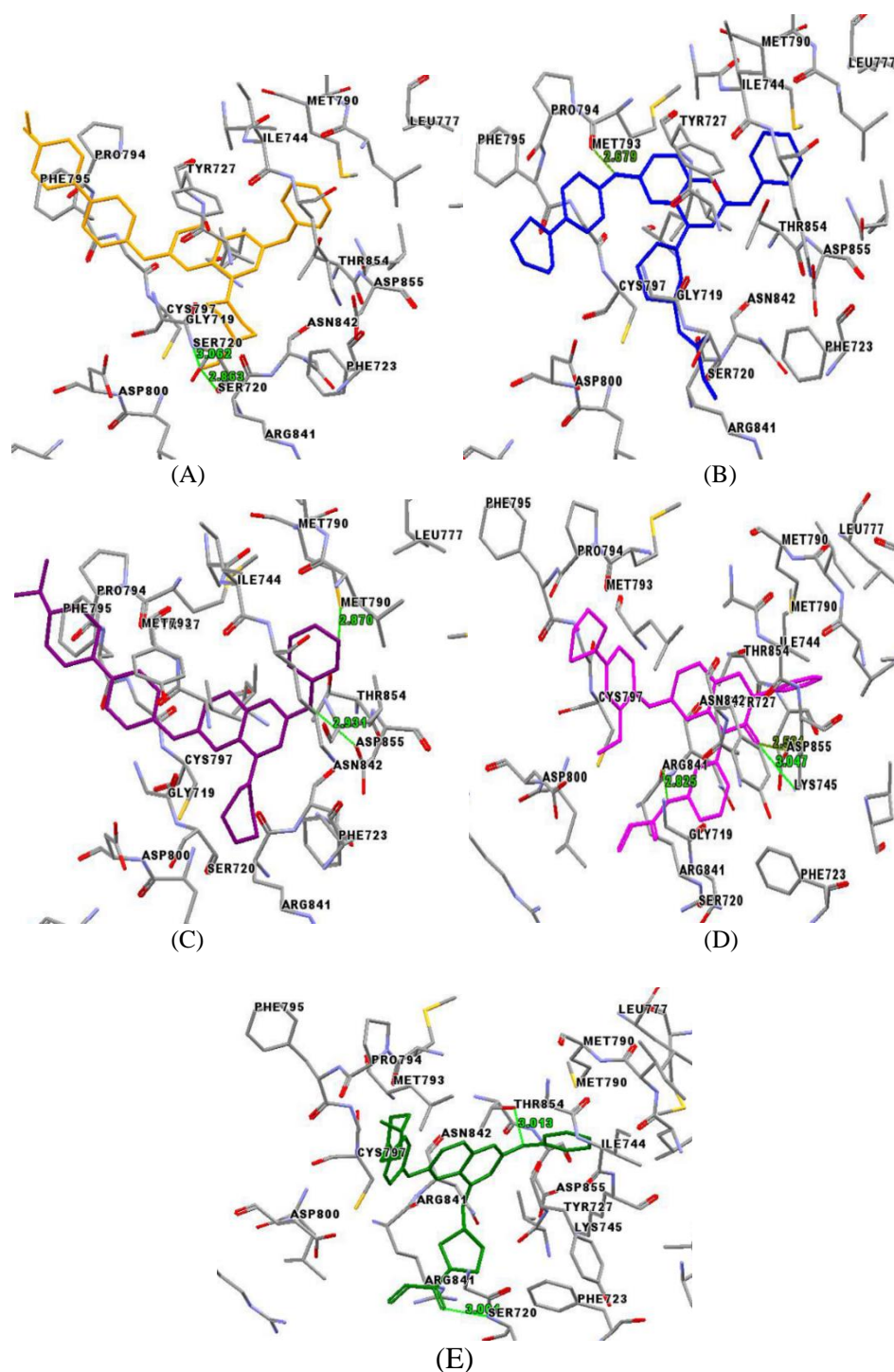
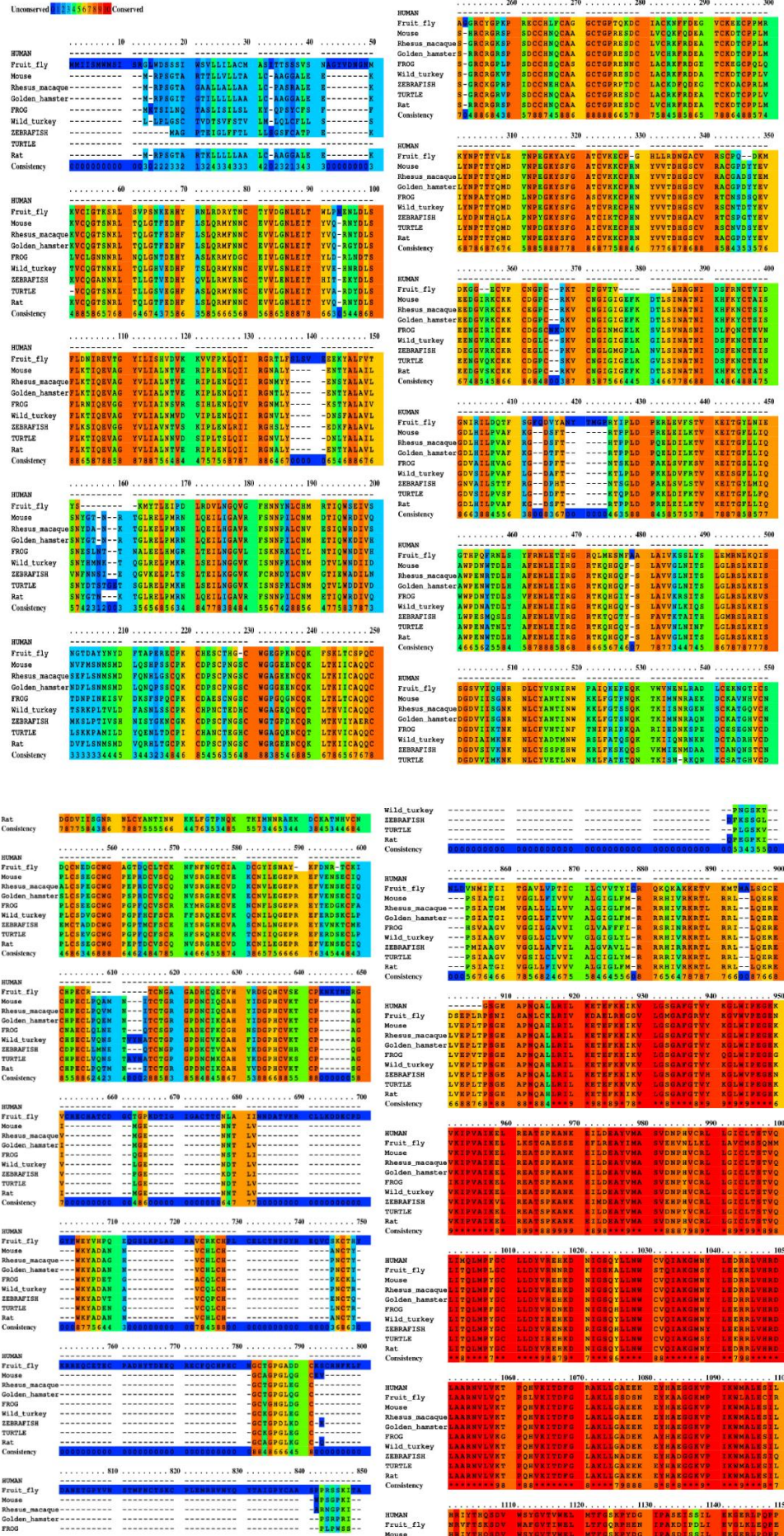


Figure 6. Binding interactions. (A) in between compound 29 (yellow) and amino acid residues (labeled in black); (B) in between compound 34 (blue) and amino acid residues; (C) in between compound 26 (violet) and amino acid residues; (D) in between compound 15 (pink) and amino acid residues; (E) in between compound 35 (green) and amino acid residues.

3.3. Multiple Sequence Alignment results

Multiple sequence alignment was performed to check whether the amino acids of EGFR protein interacted with the molecules and whether the other amino acids present in the binding site are conserved amino acid residues. The alignment of amino acid residues of different organisms is shown in Figure 7. The color index from blue to red indicates the increase in the consistency of conservation (unconserved to conserved). Figure 8 presents the character of

preserving crucial amino acid residues in a simplified form. Here it can be observed that interacting amino acids MET 793, ASP 855, ASP 800, ARG 841, THR 854, and MET 790 in the docking study of EGFR protein are highly conserved among all species. Conservation confirms that the functionally significant amino acid residues among different species have interacted in the docking study. The alignment scores and results are summarized in Table 6.



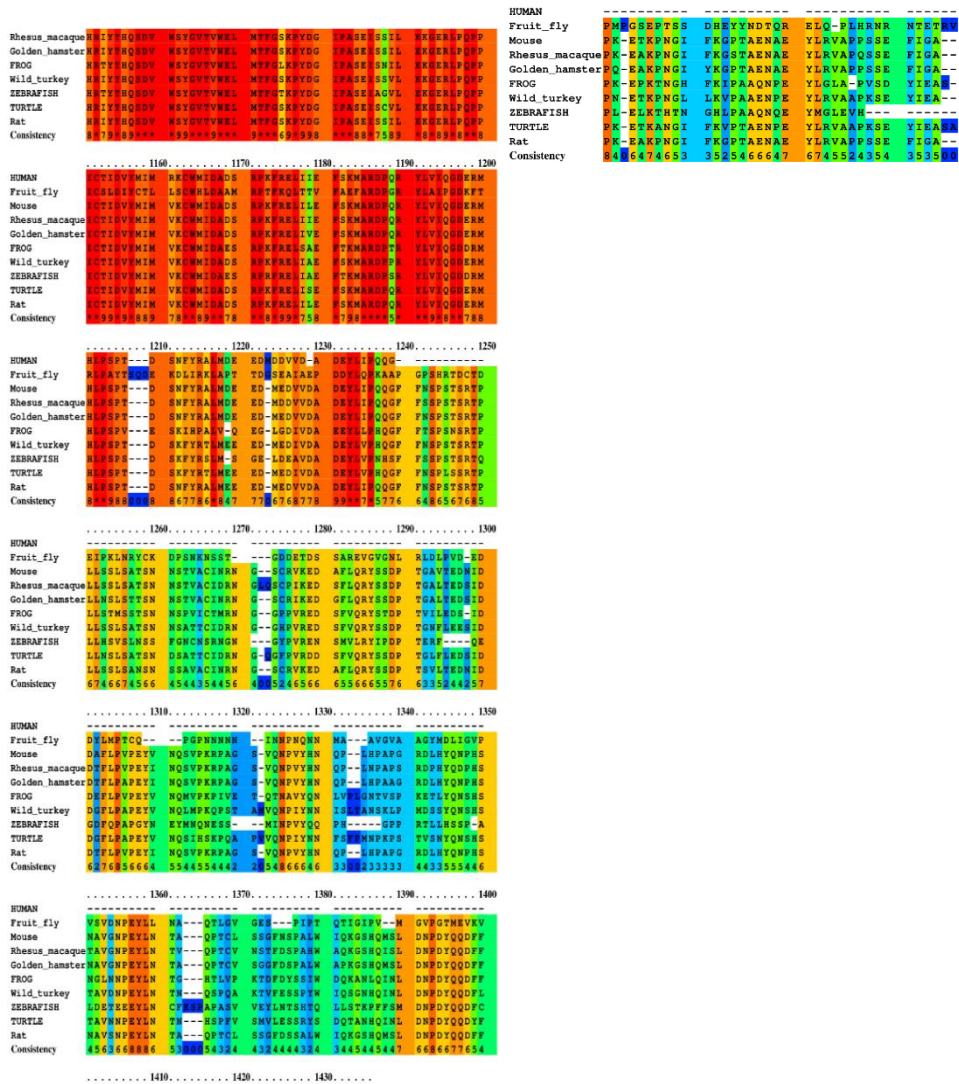


Figure 7. Alignment of amino acid sequences of EGFR obtained from ten different species as *Homo sapiens* (Human), *Drosophila melanogaster* (Fruit fly), *Mus musculus* (Mouse), *Macaca mulatta* (Rhesus macaque), *Rattus norvegicus* (Rat), *Mesocricetus auratus* (Golden hamster), *Xenopus tropicalis* (Frog), *Danio rerio* (Zebrafish), *Pelodiscus sinensis* (Turtle), and *Meleagris gallopavo* (Wild turkey). The conservation index colors reserved residues.

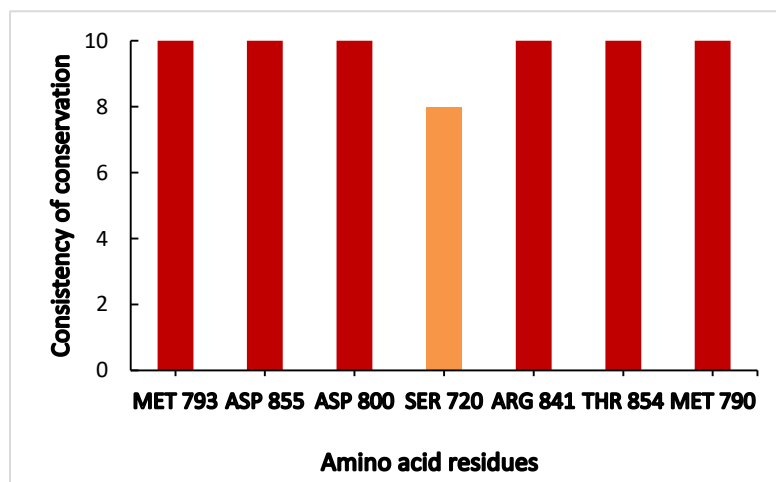


Figure 8. Simplified presentation of conservation consistency of important amino acid residues.

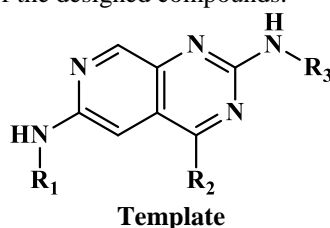
Table 6. Results of multiple sequence alignment of EGFR by PRALINE server.

| Alignment results for amino acids of EGFR |
|--|
| Alignment score = 704656.00 |
| Alignment score per aligned residue pair = 15.43 |
| Sequence identities = 31475 |
| Percent sequence identity = 0.69 |
| Number of sequences = 10 |
| Alignment length = 1436 |
| Number of residues = 11347 |
| Number of gaps = 3013 |

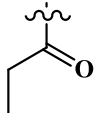
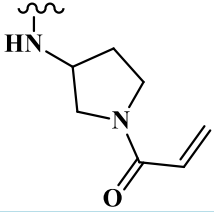
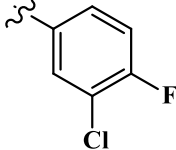
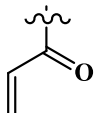
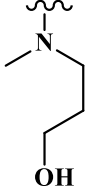
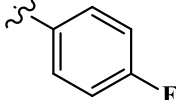
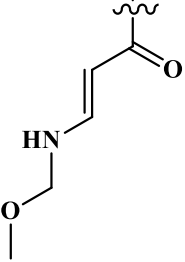
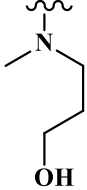
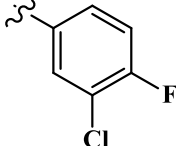
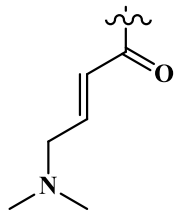
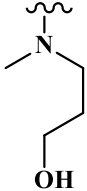
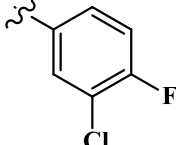
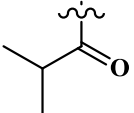
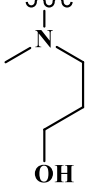
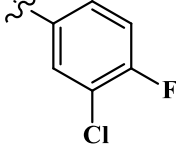
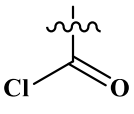
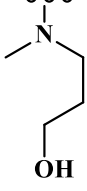
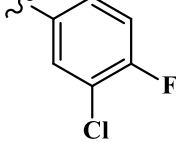
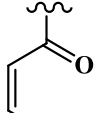
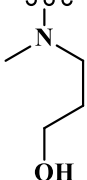
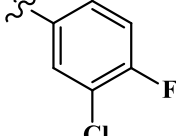
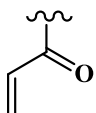
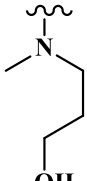
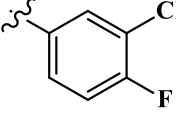
3.4. Designing of New Compounds, QSAR, and docking predictions

Based on structural attributes obtained from CoMFA and CoMSIA contour mapsin3D QSAR and analysis of docking results, several new pyrimidine-based tyrosine kinase inhibitors could be designed, which may help to overcome the drug resistance developed

over the double mutant EGFR^{T790M/L858R} to obtain improved biological activity against non-small cell lung cancer. The chemical structures of the newly designed compounds and their inhibitory activities against EGFR^{L858R/T790M} predicted by the CoMFA and CoMSIA models are enlisted in Table 7.

Table 7. Predicted biological activities of the designed compounds.

| No. | R ₁ | R ₂ | R ₃ | CoMFA Predicted pIC ₅₀ | CoMSIA Predicted pIC ₅₀ |
|-----|----------------|----------------|----------------|-----------------------------------|------------------------------------|
| N1 | | | | 7.7845 | 7.8712 |
| N2 | | | | 7.7653 | 7.8706 |
| N3 | | | | 7.7898 | 7.8694 |
| N4 | | | | 7.8046 | 7.8725 |

| | | | | | |
|-----|---|---|--|--------|--------|
| N5 |  |  |  | 7.7762 | 7.8621 |
| N6 |  |  |  | 7.7946 | 7.8865 |
| N7 |  |  |  | 7.8663 | 7.8726 |
| N8 |  |  |  | 7.8162 | 7.8837 |
| N9 |  |  |  | 7.8407 | 7.8755 |
| N10 |  |  |  | 7.8021 | 7.8729 |
| N11 |  |  |  | 7.8339 | 7.8695 |
| N12 |  |  |  | 7.8897 | 7.8652 |

| | | | | | |
|-----|--|--|--|--------|--------|
| N13 | | | | 7.8855 | 7.8663 |
| N14 | | | | 7.8759 | 7.8698 |
| N15 | | | | 7.8392 | 7.8567 |

The predictions for the newly designed pyrimidines are quite significant. Further, these compounds were docked in EGFR^{L858R/T790M} protein (PDB ID- 4I22) to investigate the binding interactions. The docking scores with binding interactions of a few selected, designed compounds N7, N4, and N1 are presented in Table 8. Among them, the docking score of compound N7 (79.59) is higher than the reference ligand gefitinib (79.19). Compound N7 formed H-bond with MET 793 similar to the co-crystallized gefitinib and dataset compound 34 at a distance of 2.665 Å and with

ASP 800 at a distance of 2.894 Å (Figure 9). Compounds N4 and N1 also showed H-bond interaction with MET 793. Further, H-bonding with SER 720 residue is observed in N4, similar to compound 29 of the dataset (Supplementary Figure 2, p. 144) On the basis of information collected from the contour plots in CoMFA and CoMSIA models and interactions identified from docking studies, some important structural requirements are highlighted in Figure 10.

Table 8. Docking scores and interactions of newly designed compounds.

| S. No | Compound Name | Docking Score (ChemPLP) | Rescore (Chemscore) | H-bond interactions with residues & distances (Å) |
|-------|---------------|-------------------------|---------------------|---|
| 1 | N7 | 79.5964 | 22.1072 | MET 793 (2.665 Å), ASP 800 (2.894 Å) |
| 2 | N4 | 77.3696 | 21.4103 | MET 793 (2.704 Å), SER 720 (3.068 Å) |
| 3 | N1 | 75.3703 | 22.7289 | MET 793 (2.729 Å), ASP 800 (2.94 Å) |

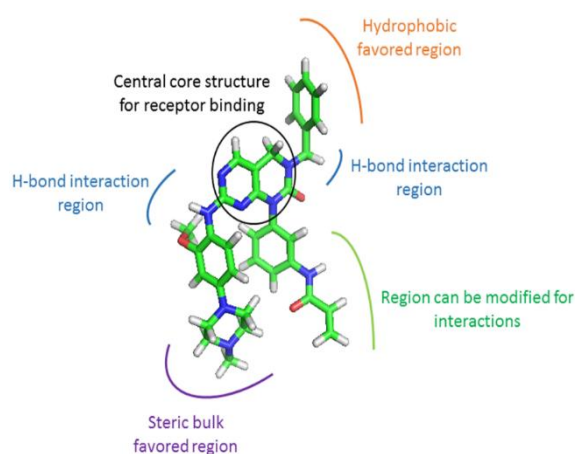


Figure 10. Identified structural requirements of novel pyrimidine derivatives as EGFR^{T790M/L858R} inhibitors summarized on template compound 17.

Among them, the docking score of compound N7 (79.59) is higher than the reference ligand gefitinib (79.19). Moreover, compound N7 formed H-bond

with MET 793 similar to the co-crystallized gefitinib and dataset compound 34 at a distance of 2.665 Å and with ASP 800 at a distance of 2.894 Å (Figure 9).

Compounds N4 and N1 also showed H-bond interaction with MET 793. Further, H-bonding with SER 720 residue is observed in N4, similar to

compound 29 of the dataset (Supplementary Figure 2).

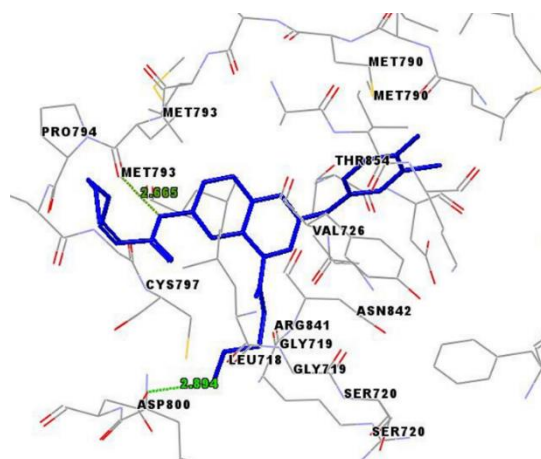


Figure 9. Binding interactions in between compound N7 and EGFR^{L858R/T790M} protein.

3.5. Binding Free energy analysis

The docked complexes of the top-scored ligands (34 and 29) and newly designed compound (N7) were further proceeded for binding free energy calculations using the MMGBSA method. Interestingly, MMGBSA analysis computed that the binding free energy of N7 ($dG = -68.59$ kcal/mol) was found to be greater than compound 29 ($dG = -62.42$ kcal/mol) and compound 34 ($dG = -58.37$ kcal/mol), which revealed stronger binding of our designed compound (N7) as compared to dataset ligands. This is as per our expectations, as compound N7 was designed by

optimizing the different parameters based on the interpretation of COMFA and COMSIA results.

3.6. ADME and drug-likeness prediction of designed compounds

Before proceeding with laboratory synthesis, in silico ADME predictions can be beneficial to collect necessary information about the prepared compounds' pharmacokinetics and druggability. The data on pharmacokinetic and drug-likeness properties of the designed compounds obtained from the ADME screening is presented in Table 9.

Table 9. In silico pharmacokinetics and drug-likeness predictions of newly designed compounds.

| No. | Pharmacokinetics | | | | Drug likeness | | | |
|-----|---------------------------|---------------|--------------|---|---------------------------|------------------------|------------------------|-------------------------|
| | Log S (ESOL) ^a | GI absorption | BBB permeant | Log P _{o/w} (MLOGP) ^b | Lipinski rule; Violations | Ghose rule; Violations | Bio-availability score | Synthetic accessibility |
| N1 | -4.39 | High | No | 2.4 | Yes; 0 | Yes; 0 | 0.55 | 3.19 |
| N2 | -4.39 | High | No | 2.4 | Yes; 0 | Yes; 0 | 0.55 | 3.17 |
| N3 | -4.39 | High | No | 2.4 | Yes; 0 | Yes; 0 | 0.55 | 3.19 |
| N4 | -4.63 | High | No | 2.62 | Yes; 0 | Yes; 0 | 0.55 | 3.28 |
| N5 | -4.9 | High | No | 2.86 | Yes; 0 | No; 2 | 0.55 | 3.88 |
| N6 | -3.8 | High | No | 2.18 | Yes; 0 | Yes; 0 | 0.55 | 3.12 |
| N7 | -4.6 | High | No | 1.96 | Yes; 0 | No; 1 | 0.55 | 3.76 |
| N8 | -4.38 | High | No | 2.53 | Yes; 0 | No; 2 | 0.55 | 3.71 |
| N9 | -4.77 | High | No | 2.96 | Yes; 0 | Yes; 0 | 0.55 | 3.29 |
| N10 | -4.82 | High | No | 2.93 | Yes; 0 | Yes; 0 | 0.55 | 3.02 |
| N11 | -4.62 | High | No | 2.89 | Yes; 0 | Yes; 0 | 0.55 | 3.38 |
| N12 | -4.62 | High | No | 2.89 | Yes; 0 | Yes; 0 | 0.55 | 3.38 |
| N13 | -4.02 | High | No | 2.4 | Yes; 0 | Yes; 0 | 0.55 | 3.32 |
| N14 | -4.62 | High | No | 2.89 | Yes; 0 | Yes; 0 | 0.55 | 3.38 |
| N15 | -4.26 | High | No | 2.62 | Yes; 0 | Yes; 0 | 0.55 | 3.41 |

^aESOL: Topological method implemented from (Delaney, 2005). Solubility class: Log S scale. Insoluble < -10 < Poorly < -6 < Moderately < -4 < Soluble < -2 < Very < 0 < Highly. ^bMLOGP: Topological method implemented by (Moriguchi et al., 1992).

According to the pharmacokinetic properties obtained, all compounds showed moderate solubility except compound N6 which is soluble in the Log S (ESOL) scale implemented from ²⁵. All the compounds possess high gastrointestinal (GI) absorption, and no one has the blood-brain barrier (BBB) permeability. The lipophilicity (MLOGP) of the compounds is well under the desired value ($MLOGP \leq 5$) ²⁷. The Lipinski ²⁷ and Ghose ²⁸ filters were applied to check the drug-likeness parameter. The Lipinski (Pfizer) filter is the most popular rule-of-five for drug candidate selection, states that absorption of a compound is more likely to take place when the molecular weight is under 500 g/mol, Log P value is below 5, and the compound is having a maximum of 5 hydrogen bond donor and 10 hydrogen bond acceptor atoms. Interestingly each of the designed compounds passed the Lipinski rule with zero violation. Ghose rule screens drug-like candidates based on the following limits of parameters: molecular weight between 160 and 480 g/mol, Log P value ranging between -0.4 to 5.6, molar refractivity between 40 and 130, and the total number of atoms between 20 and 70. Maximum designed compounds met the criteria of the Ghose rule with no violation except compounds N5 and N7 with two and N8 with one violation. Martin developed a score determining the probability of a compound having bioavailability ($F > 10\%$) in a rat model. If this score for a compound is 0.55, it means that the compound passes the rule-of-five (Lipinski rule) and has a 55% probability of giving $F > 10\%$ in the rat. While the score of 0.17 represents failing of the Lipinski rule and still has a 17% probability of giving $F > 10\%$ ²⁹. In our study, all the designed compounds have a bioavailability score = 0.55, as they pass the Lipinski rule and thus have a 55% probability of $F > 10\%$. The synthetic accessibility of the designed compounds was also determined. Here the score for synthetic accessibility is categorized in the range of 1 (very easy to synthesize) to 10 (very difficult to synthesize) ²⁴. The synthetic accessibility scores for all 15 designed compounds range from 3.02 to 3.88. It suggests that our designed compounds can be synthesized further.

4. Conclusion

In the present work, 3D QSAR and molecular docking studies were performed in search of necessary structural attributes and potential binding interactions to improve the potency of the pyrimidine derivatives as EGFR-TKIs in non-small cell lung cancer. Considering derived structural information, some new pyrimidine-based tyrosine kinase inhibitors were designed, which showed good predictions using developed 3D QSAR models. The designed compounds showed H-bond interaction with MET 793; newer interactions were also seen with residues like ASP 800 and SER 720. Dataset compounds 29, 34, 26, 15, 35 and designed compound N7 showed higher docking scores with promising binding interactions than the reference ligand gefitinib, which

could be an encouraging finding for further optimizations. Interestingly, the MMGBSA binding energy calculations revealed the stronger binding affinity of the designed compound N7 compared to other ligands. In silico ADME predictions of the designed compounds were also carried out to understand drug-likeness better. Finally, we summarize that our findings from the above studies will be helpful for the future development of novel pyrimidine derivatives as EGFR tyrosine kinase inhibitors which could overcome the developed drug resistances.

Acknowledgment

Pradip Jana is thankful to the All India Council for Technical Education (AICTE), New Delhi, for awarding the fellowship. In addition, the authors acknowledge the Department of Pharmaceutical Sciences, Dr. Harisingh Gour University (A Central University), India, for providing research facilities.

Declarations

Ethics approval and consent to participate: No animal experiments were performed

Availability of data and materials: All relevant data are within the manuscript.

Competing interests: The authors declare that they have no competing interests

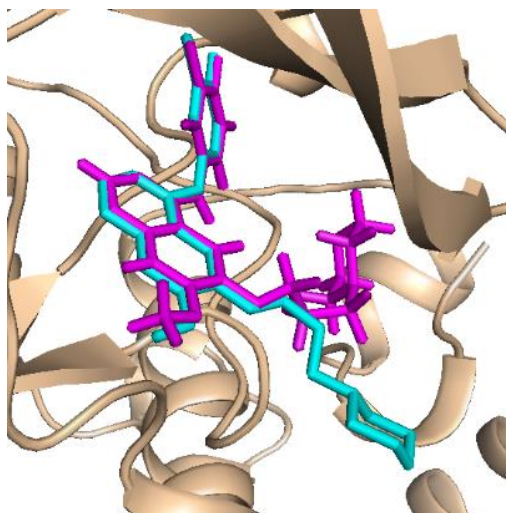
Funding: All India Council for Technical Education (AICTE), New Delhi, with grant number (ID:2018-00000472).

References

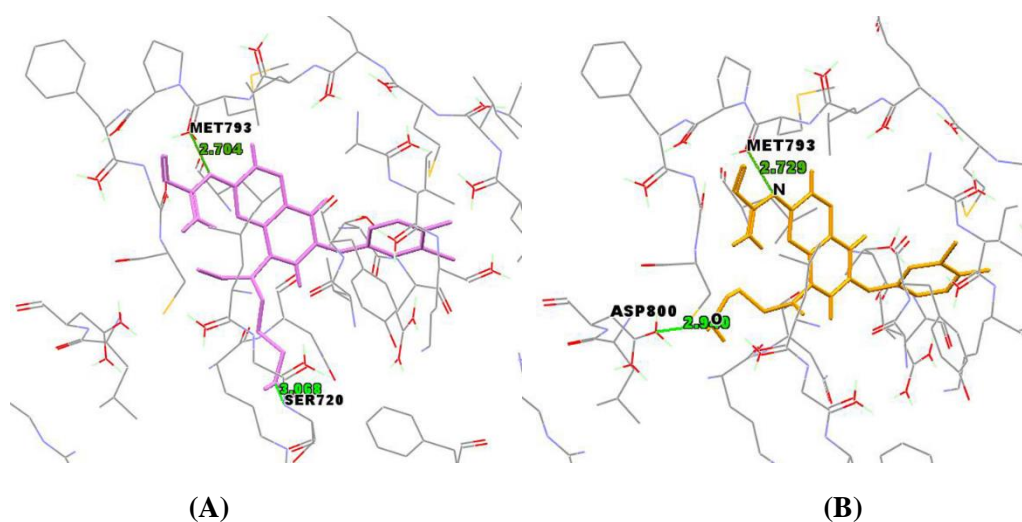
1. R.L. Siegel, K.D. Miller, A. Jemal, Cancer statistics, 2016. CA: a cancer journal for clinicians, **2016**, 66(1), 7-30.
2. S.V. Sharma, D.W. Bell, J. Settleman, D.A. Haber, Epidermal growth factor receptor mutations in lung cancer, Nature Reviews Cancer, **2007**, 7(3), 169-181.
3. M.A. Olayioye, The ErbB signaling network: receptor heterodimerization in development and cancer, The EMBO journal, 2000, 19(13), 3159-3167.
4. M.M. Moasser, Targeting the function of the HER2 oncogene in human cancer therapeutics. Oncogene, **2007**, 26(46), 6577-6592.
5. C.E. Geyer, Lapatinib plus capecitabine for HER2-positive advanced breast cancer, New England Journal of Medicine, **2006**, 355(26), 2733-2743.
6. S.Chang, Design, synthesis, and biological evaluation of novel conformationally constrained inhibitors targeting epidermal growth factor receptor threonine790→methionine790 mutant, Journal of medicinal chemistry, **2012**, 55(6), 2711-2723.
7. T.S. Mok, Gefitinib or carboplatin–paclitaxel in pulmonary adenocarcinoma. New England Journal of Medicine, **2009**, 361(10), 947-957.

8. F.A. Shepherd, Erlotinib in previously treated non-small-cell lung cancer, *New England Journal of Medicine*, **2005**, 353(2), 123-132.
9. K.D. Carey, Kinetic analysis of epidermal growth factor receptor somatic mutant proteins shows increased sensitivity to the epidermal growth factor receptor tyrosine kinase inhibitor, erlotinib. *Cancer Research*, **2006**, 66(16), 8163-8171.
10. M.L. Sos, Chemogenomic profiling provides insights into the limited activity of irreversible EGFR Inhibitors in tumor cells expressing the T790M EGFR resistance mutation, *Cancer Research*, **2010**, 70(3), 868-874.
11. Kim, Y., et al., The EGFR T790M mutation in acquired resistance to an irreversible second-generation EGFR inhibitor, *Molecular cancer therapeutics*, **2012**, 11(3), 784-791.
12. Z. Song, Challenges and perspectives on the development of small-molecule EGFR inhibitors against T790M-mediated resistance in non-small-cell lung cancer: miniperspective, *Journal of medicinal chemistry*, **2016**, 59(14), 6580-6594.
13. H. Patel, Recent updates on third-generation EGFR inhibitors and emergence of fourth generation EGFR inhibitors to combat C797S resistance. *European Journal of Medicinal Chemistry*, **2017**, 142, 32-47.
14. X. Ji, Design, synthesis and biological evaluation of novel 6-alkenylamides substituted of 4-anilinothieno [2, 3-d] pyrimidines as irreversible epidermal growth factor receptor inhibitors, *Bioorganic & medicinal chemistry*, **2014**, 22(7), 2366-2378.
15. H. Zhang, Discovery of 2, 4, 6-trisubstituted pyrido [3, 4-d] pyrimidine derivatives as new EGFR-TKIs, *European Journal of Medicinal Chemistry*, **2018**, 148, 221-237.
16. B. Bhardwaj, Insight into structural features of phenyltetrazole derivatives as ABCG2 inhibitors for the treatment of multidrug resistance in cancer, SAR and QSAR in *Environmental Research*, **2019**, 30(7), 457-475.
17. Y. Jian, Molecular modeling study for the design of novel peroxisome proliferator-activated receptor gamma agonists using 3D-QSAR and molecular docking, *International Journal of Molecular Sciences*, **2018**, 19(2), 630.
18. A. Dixit, Development of CoMFA, advance CoMFA and CoMSIA models in pyrroloquinazolines as thrombin receptor antagonist, *Bioorganic & medicinal chemistry*, **2004**, 12(13), 3591-3598.
19. S. Bhattacharya, Comparative Molecular Field Analysis on 4 (3H) Quinazolinone derivatives for the Development of Potential Antimicrobial Agents, *Journal of Drug Delivery and Therapeutics*, **2019**, 9(2), 603-611.
20. A. Golbraikh, A. Tropsha, Beware of q2! *Journal of molecular graphics and modelling*, **2002**, 20(4), 269-276.
21. K.S. Gajiwala, Insights into the aberrant activity of mutant EGFR kinase domain and drug recognition, *Structure*, **2013**, 21(2), 209-219.
22. S. Agarwal, A. Dixit, S.K. Kashaw, Ligand and structure-based virtual screening of chemical databases to explore potent small molecule inhibitors against breast invasive carcinoma using recent computational technologies, *Journal of Molecular Graphics and Modelling*, **2020**, 107591.
23. M. Mishra, Integrated computational investigation to develop molecular design of quinazoline scaffold as promising inhibitors of plasmodium lactate dehydrogenase, *Journal of Molecular Structure*, **2020**, 1207, 127808.
24. A. Daina, O. Michielin, V. Zoete, SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules, *Scientific reports*, **2017**, 7, 42717.
25. J.S. Delaney, Predicting aqueous solubility from structure, *Drug Discovery Today*, **2005**, 10(4), 289-295.
26. I. Moriguchi, Simple method of calculating octanol/water partition coefficient, *Chemical and pharmaceutical bulletin*, **1992**, 40(1), 127-130.
27. C.A. Lipinski, Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced drug delivery reviews*, **1997**, 23(1-3), 3-25.
28. A.K. Ghose, V.N. Viswanathan, J.J. Wendoloski, A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases, *Journal of combinatorial chemistry*, **1999**, 1(1), 55-68.
29. Y.C. Martin, A bioavailability score, *Journal of medicinal chemistry*, **2005**, 48(9), 3164-3170.

Structural identification of novel pyrimidine derivatives as epidermal growth factor receptor inhibitors using 3D QSAR, molecular docking and MMGBSA analysis: a rational approach in anticancer drug design



Supplementary Figure 1. Superimposition of co-crystallized gefitinib (sky) and re-docked co-crystallized gefitinib (purple) in the binding site of EGFR^{L858R/T790M} protein (PDB ID: 4I22).



Supplementary Figure 2. Binding interactions in between EGFR^{L858R/T790M} protein and newly designed compounds. (A) Compound N4; (B) Compound N1.