Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Descriptor-Free QSAR: Effectiveness and Screening for Putative Inhibitors of FGFR1

Lateef Sulaimon (adlat4best@yahoo.com)
University of Lagos https://orcid.org/0000-0002-5530-6205
Ireoluwa Joel
University of Ilorin
Temidayo Adigun
University of Ilorin
Rahmat Adisa
University of Lagos
Titilola Samuel
University of Lagos
Taiwo Ademoye
University of Lagos
Moyosore Ogunleye
University of Lagos

Article

Keywords: Descriptor-free QSAR, FGFR1, QM-MM optimization, LSTM, Induced-fit docking

Posted Date: February 8th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-154245/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Descriptor-Free QSAR: Effectiveness and Screening for Putative Inhibitors of FGFR1

Sulaimon, L.A.[1,*], Adigun, T.O.[2,*], Joel, I.Y.[2], Adisa, R.A.[1], Samuel, T.A.[1], Ademoye, T.A.[1] and Ogunleye, M.O.[1]

[1]*Department of Biochemistry, Faculty of Basic Medical Sciences, College of Medicine of University of Lagos, Idi-araba, Lagos, Nigeria.*

[2]*Department of Biochemistry, University of Ilorin, Ilorin, Kwara State, Nigeria*

**Corresponding authors: [1]adlat4best@yahoo.com; [2]prothesis4life@gmail.com*

Abstract

The effectiveness of descriptors-utilizing quantitative structure-activity relationship models in drug design remains limited by the quality of descriptors used in training, this then raises the question: can QSAR models be directly trained on compound SMILES? Long short-term memory (LSTM) algorithm has been employed to answer this question however, direct application remain scarce. The effectiveness of a descriptor-free QSAR (LSTM-SM) in modeling FGFR1 inhibitors dataset while comparing with two conventional QSAR using descriptors (126bits Morgan fingerprint and 2D descriptors respectively) was investigated in this study. The validated descriptor-free QSAR model was thereafter used to screen for active FGFR1 inhibitors in ChemDiv database and subjected to molecular docking, induced-fit docking, and QM-MM optimization to filter for compounds with high binding affinity and suggest putative mechanism of inhibition and specificity. The LSTM-SM model, when compared with the conventional QSAR models, performed better having accuracy, specificity, and sensitivity of 0.92, model loss of 0.025 and AUC of 0.95.Fifteen thousand compounds were predicted as actives from the ChemDiv database and four compounds finally selected. Of the four, three showed putatively effective binding interactions with key active site residues and were also effective against acquired resistance due to gateway residue mutations. The advent of self-feature extracting machine learning algorithms, therefore, has provided the possibility of better predictive model quality that is not necessarily limited by compound descriptors thus we apply this approach in discovering putatively active FGFR1 inhibitors and elucidated putative mechanism of inhibition and specificity for the obtained compounds.

Keywords: Descriptor-free QSAR; FGFR1; QM-MM optimization; LSTM; Induced-fit docking

Introduction

The effectiveness of "conventional" quantitative structural activity relationship (QSAR) models in computer-aided drug discovery are well documented [1–3]—we define conventional QSAR as any model utilizing descriptors in the course of training. However, there are challenges i.e. QSAR models are only as good as their descriptors ("garbage in garbage out"). Different types of descriptors have been developed [4] to have robust, and reliable models; but these are not without problems which include but not limited to: descriptor interpretations, bias, need for third party software for descriptor calculations, effective descriptors selection algorithm, inter-correlations, etc. [5–7].

This raises the question; can QSAR models be trained directly on the compound SMILES while eliminating descriptors? Recent works have employed long short term memory recurrent neural networks (LSTM-RNN) algorithm to build a descriptor-free QSAR model on large and diverse datasets as proof of concepts [8], but there is paucity of data on direct application of this method for drug design and discovery.

Squamous-cell carcinomas account for 20–30% of non-small cell lung cancer (NSCLC) [9]. Squamous-NSCLC (Sq-NSCLC), unlike lung adenocarcinomas, lacks commonly targetable oncogenic aberrations such as EGFR mutations, ROS-1, or ALK rearrangements[10,11] but recent discoveries have revealed the basic fibroblast growth factor receptor 1 (FGFR1) as a crucial druggable target in squamous non-small cell lung cancer. FGFR1 is a transmembrane receptor tyrosine kinase having an extracellular domain for binding of ligand and a catalysis-mediated intracellular domain being responsible for the receptor kinase activity [12]. It plays a physiological role in the basic hallmarks of cell development including cell proliferation, growth, differentiation, angiogenesis, migration, and survival [13,14] but dysregulated in Sq-NSCLC condition through mechanisms including over-amplification of chromosome 8p12 and/or aberrant transcriptional regulation[9-11].

Several small-molecule FGFR1 kinase inhibitors have been developed with a substantial amount (including AZD4547; BGJ398; JNJ-42756493; LY2874455; BAY1163877, etc.) undergoing clinical evaluation[14,16] but there exists the challenge of specificity, toxicity, acquired resistance (via mutation of "gateway" residues), etc.

We therefore aim to investigate the effectiveness of descriptor-free QSAR (comparably with conventional QSAR) in modeling FGFR1 inhibitors dataset in Chembl repository; employ the

model to screen for potential active FGFR1 inhibitors in Chemdiv database; determine their putative mechanism of inhibition and specificity as well as examine their ability to selectively overcome acquired FGFR1 drug resistance.

Materials and Methods

Hardware and Software

A Google collaboratory notebook runtime using a 12GB RAM, single-core 2GB GPU, and a Linux Ubuntu 18.04 distro system running on a 12 GB RAM, core i5, 4 Cores, 2.5GHz was used for the analyses. Python packages: Tensorflow v2.3.0 [17]; Scikit-learn v0.22.2[18]; Feature selector; Numpy v1.18.1,Pandas v1.0.3, Matplotlib v3.2.1, and TALOS v0.11.1 were used for model training, evaluation, feature extractions and preparations, data waggling, data visualizations, and hyperparameter tuning respectively. All python packages run on Python v3.6 using Jupyter Lab v1.2.6.

Data Extraction, Descriptor Calculation, and Preparation

All inhibitors of FGFR1 (4123) were downloaded from the ChEMBL [19] database and imported into a standalone MySQL database that was created for analysis. All inhibitors with no IC_{50} values were removed, while inhibitor smiles and corresponding IC_{50} values were extracted into a CSV sheet.

The SMILES were one-hot encoded using Molvecgen module for descriptor-free QSAR modelling whereas the Morgan fingerprint (126bits) was calculated using RDKit AllChem. Get Morgan Fingerprint function and MOE 2D descriptors were calculated. The Feature Selector module was used to pre-process 2D descriptors; it removed descriptors that are inter-correlated (correlation threshold was set at 0.75) and descriptors with little or no contribution to 0.95 cumulative importance (Feature Selector uses XGBoost algorithm [20] to estimate the feature importance of the descriptors).

The IC_{50} values (nM) were converted to pIC_{50} values ($pIC_{50} = 9\text{-}\log_{10}(IC_{50})$; which were in turn converted to categorical values of active (1) and non-active (0). Activity threshold for conversion was set at pIC_{50} 7.

The data were split into three sets: training-set (70%), test-set (20%), and validation-set (10%) using the RDKit Max-Min Picker module. The Max-Min algorithm calculates the fingerprints for the whole dataset, evaluates the Tanimoto distance between fingerprints (MACCS) and diverse subsets selected [21].

Model Training and Evaluation

Three models were built in this study:

- long short-term memory (LSTM) with canonical SMILES. i.e., no descriptors (LSTM-SM);
- neural network model with molecular fingerprints as descriptors (NN-FP);
- random forest model with MOE 2D descriptors (RF-2D).

LSTM-SM Model

i) LSTM Principle

Long short-term memory (LSTM) was introduced to solve the long term dependence problem of Recursive Neural Network (RNN) [22]; it utilizes cell states—serving as a form of "memory"—connected network-wide. To update cell states during training, "Gates" are introduced: a forget layer gate (f_t)that determines part of the cell state to be discarded (Eq1); an input layer gate (i_t) which determines part of the cell state that has to be updated (Eq2); and a tanh layer gate ($\check{C}_t$) that creates new candidate values that would be added to the cell state (Eq3).

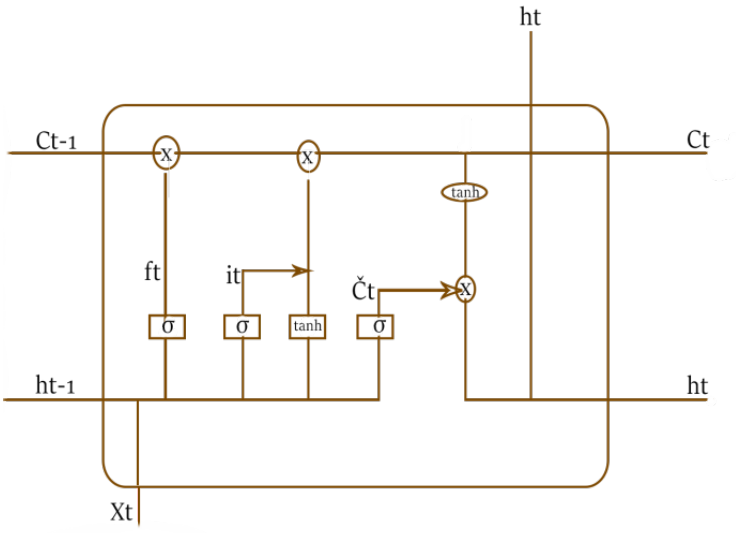

Figure 1: A cell Unit in a LSTM Network(σ: *Sigmoid activation Function;* C_{t-1}*: previous cell state;* h_{t-1}*: previous hidden state;* f_t*: forget gate;* i_t*: input layer gate;* $\check{C}_t$*: tanh layer gate; tanh: tanh activation function;* h_t*: new hidden state;* C_t*: new cell state)*

$$f_t = \sigma\,(W_f \,.\, [h_{t-1},\ x_t] + b_f) \qquad \textit{Eq1}$$

$$i_t = \sigma\,(W_i \,.\, [h_{t-1},\ x_t] + b_i) \qquad \textit{Eq2}$$

$$\check{C}_t = \tanh\,(W_{\check{C}} \,.\, [h_{t-1}, xt] + b_{\check{C}}) \qquad \textit{Eq3}$$

- σ = Sigmoid activation Function
- W_i = Input gate weight
- h_{t-1} = Previous hidden state
- X_t = Inputted vector
- b_i= Input gate bias

To create a new cell state (C_t) the old cell sate (C_{t-1}) is updated (Eq 4).

$$C_t = f_t * C_{t-1} + i_t * \check{C}_t \qquad \textit{Eq (4)}$$

Finally, an output gate layer (O_t) is created to determine which aspect of the new cell state (C_t) to be outputted as a hidden state (h_t) to the next cell in the network (Eq5, Eq6).

$$O_t = \sigma\,(W_o\,.\,[h_{t-1},\ x_o] + b_o) \qquad Eq(5)$$

$$h_t = O_t * tanh(C_t) \qquad Eq(6)$$

ii) Model Training

Talos module was used for hyperparameter tuning: a python dictionary specifying different hyperparameters and value ranges was provided and the module randomly selects from this parameter dictionary and constructs various LSTM models to return the model performance of each selection.

Table 1: LSTM-SM hyperparameters

Model	Number of layers	Loss function	Optimizer	Units (Neurons)	Dropout rate	Batchsize	Epoch
LSTM-SM	3	Logcosh	Adam	256	0.3	128	100

iii) Evaluation Metrics

The performance of the models was evaluated [23]using:

- accuracy = (Tp + Tn) / (Tp + Fp + Tn + Fn)
- sensitivity = Tp / (Tp + Fn)
- specificity = Tn/ (Tn +Fp)
- the area under the curve (AUC) calculated from the receiver operator curve (ROC) plots

iv) Model Validation

Due to the stochastic nature of neural networks, the LSTM-SM model was validated using the following protocols (model performance was validated at different random seeds):

- 10-Fold Cross-validation: the LSTM-SM model was subjected to a 10% split 10-fold cross-validation and the average performance of the model was computed
- Y-randomization: the training label was randomized and trained,this new model was used to predict the test set and validation sets. This protocol ascertaining that the observed performancewas not due to chance. It is expected that the new model would perform significantly lower than the original model (unrandomized). This process is iterated 10 times and average performance wascomputed.
- Validation-set: Validation data set were also used to evaluate model performance at every stage of training and testing.

Baseline Models

Two baseline models were built to serve as representative of conventional QSAR:

- A fully connected neural network (NN-SM) trained on RDKit Morgan fingerprint (126 bits)
- A Random Forest model (RF-2D) trained on MOE 2D descriptors

The hyperparameters used for NN-SM and RF-2D model is stated in Table 2 and 3

Table 2: NN-FP Hyperparameters

Model	Number of layers	Loss function	Optimizer	Units (Neurons)	Dropout rate	Batchsize	Epoch
NN-FP	3	Log cosh	Adam	256	0.2	5	15

Table 3: RF-2D hyperparameters

Model	Estimators	Criterion	Minimum Samples Split	Minimum Samples Leaf
RF-2D	100	Gini	2	1

Applicability Domain

Two protocols were applied to define the applicability domain of the model:

- Enalos Similarity KNIME node [24] was used to flag compounds that were not similar to the training dataset. The fingerprints of the compounds were computed and subjected to the Enalos similarity node.
- Compounds with functional groups not found in the training set were considered outside the applicability domain of the model

Database Screening

ChemDiv database representative compounds (300,000) were downloaded for screening: in screening, active class prediction probability was restricted to 0.75 and above.Also, due to the stochastic nature of neural networks,the model predictions were repeated four times, and only compounds consistent in at least 3 predictions were selected.

Molecular Docking

The crystal structure of FGFR1 in complex with AZD4547(PDB ID: 4V05) was downloaded from the RCSB protein database and prepared using the Schrödinger protein preparation wizard [25]; missing side chains and loops were filled with prime [26], water beyond 5Å from the het group was deleted and het states were generated using Epik [27](pH 7.0 +/- 2.0) while all other parameters were left at default values. The predicted active compounds were prepared using the Schrödinger LigPrep module in which force field minimization using OPLS2005 [28] and Het states were generated using Epik [27] (pH 7.0 +/- 2.0) while the active site coordinates of the FGFR1 was extracted using the receptor grid generation module of Schrödinger.

The predicted active compounds were thereafter docked using Schrödinger virtual screening workflow (consisting of a filtering stage: based on drug-likeness criteria, docking, and binding affinity calculation). The docking stage was a three-step process utilizing the three Schrodinger glide docking algorithms: high throughput virtual screening (HTVS), Standard Precision (SP), and Extra precision (XP) sequentially—each with an increasing level of accuracy [29]. We initially docked the compounds in the first step in which 10% of the top scored compounds were returned as input for step two which involved the glide SP docking of the compounds for returns

of 10% of the top best scoring compounds. Finally, the resultant compounds were docked using glide XP to retain the top best 100 compounds prior to their binding affinity calculations using the MMGBSA protocol.

Molecular Mechanics Generalized Born Surface Area

Compounds binding affinity was calculated using the prime molecular mechanics-generalized Born surface area (MM-GBSA) [30]. MM-GBSA aids in optimizing the binding free energies calculation after minimization of the docked protein-ligand complex under VSGB 2.0 implicit solvation model and OPLS-2005 force field. The compound binding free energy in this study was calculated according to Equation 7.

$$\Delta G_{bind} = G_{complex} - G_{protein} - G_{ligand} \qquad Eq\ (7)$$

where $G_{complex}$, $G_{protein}$ and G_{ligand} represent the binding free energies of the protein-ligand complex, protein, and ligand respectively.

Molecular Docking Protocol Validation

The docking protocol was validated by redocking the co-crystalized ligand and superimposing the redocked pose with the crystalized pose. The RSMD value of pose differences was calculated. An enrichment study was done: 20 FGFR1 inhibitors reported in the literature were mixed with decoys and docked (this investigated how well the docking protocol was able to select active compounds ahead of decoys); ROC curve was plotted and AUC calculated.

Induced-fit Docking

The top 12 compounds from the molecular docking study were subjected to induced-fit docking to predict the binding pose of the compounds and calculate corresponding binding affinities (using MMGBSA).

Maestro induced-fit docking module [32, 36] was used; briefly, the compounds were docked into the active site (with the active site residues held rigidly), prime module refined the active sites residue backbone, and finally redocked the compounds into the refined protein conformation.

Residue Mutation Analysis

Schrodinger maestro mutates utility was used to mutate FGFR1 Val561 to Met561; the mutated protein was subjected to induced-fit docking with selected compounds to investigate possible bonding interactions with the mutated residue.

QM-MM Optimization

Optimization studies were further carried out on the compound poses obtained from the induced-fit docking experiment using the Schrodinger Q-site module [31]. This was to validate bonding interactions observed in induced-fit docking poses [32]. The ligand and active site residues (side-chain and backbone) involved in the interactions are treated as the quantum mechanics (QM) region while the protein complex (excluding active site residue and ligand) was treated as the molecular mechanics (MM) region. The QM calculation was done using density functional theory (DFT) with Becke's three-parameter exchange potential, Lee-Yang-Parr correlation function (B3LYP) and basis set 631G** level, while the MM region was treated using OPLS2005; minimization was done using Truncated Newton, 1000 maximum cycle, with the convergence criterion set to energy gradient while all other parameters were set at default.

Results

Model Building and Evaluation

Three models were trained and evaluated: LSTM-SM (long short-term memory model built using canonical smiles only), NN-FP (Neural Network model using fingerprints), and RF-2D (Random forest model using 2D descriptors). The LSTM-SM model accuracy ranged from 0.88 to 0.95 over different datasets splits, including training-set, test-set, and validation-set, with 10-fold cross-validation accuracy of 0.92 and drop in accuracy to 0.62 when subjected to Y-randomization (Table 4). There was also a progressive increase in accuracy and reduction of model loss over 100 epochs, while the plotted model ROC curve for each data split had an AUC of 0.95 (Figure 2).

Table 4: LSTM-SM model performance on different datasets and validation protocols (Cross-validation and Y-randomization)

Model evaluation	Loss	Sensitivity	Specificity	Accuracy
Training-set	0.0371	0.8898	0.8898	0.8898
Test-set	0.0114	0.9705	0.9705	0.9705
Validation-set	0.0184	0.9521	0.9521	0.9521
Cross-validation	0.0252	0.9282	0.9282	0.9282
Y-Randomization	0.1512	0.6175	0.6175	0.6175

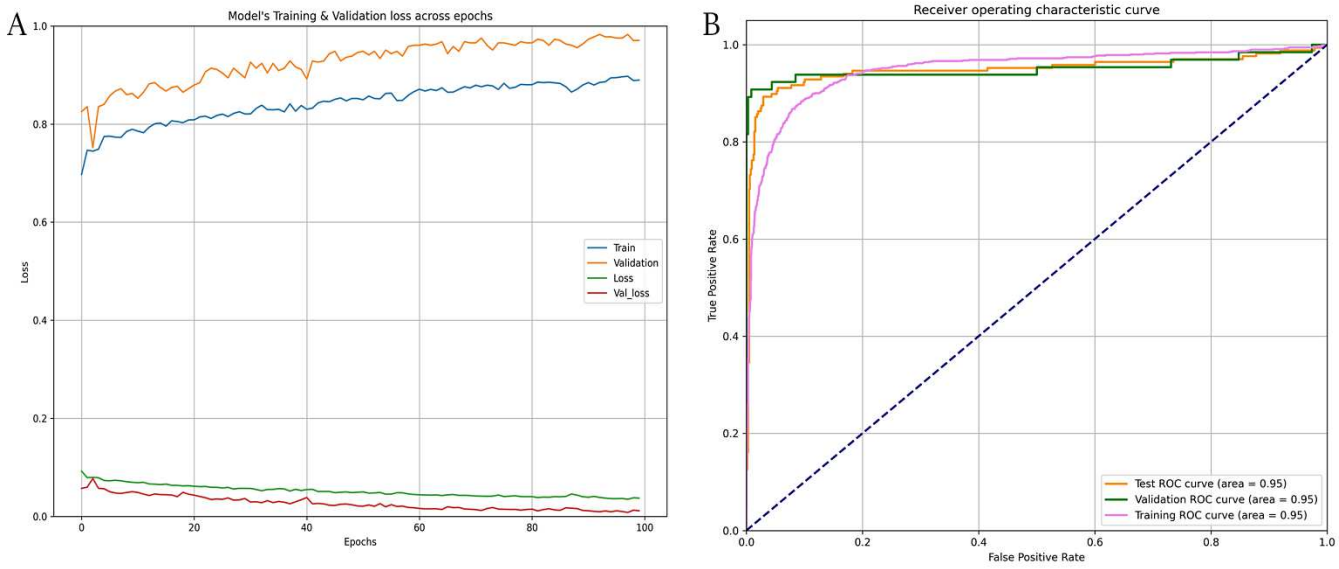

Figure 2: LSTM-SM model evaluation a) accuracy and loss over 100 epoch b) ROC curves*: Training-set: AUC 0.95; Test-set: AUC 0.95; Validation-set: AUC 0.95*

The cross-validation of the NN-FP model showed accuracy, sensitivity, and specificity of 0.91, while there was a significant reduction in the model performance (Table 5) with the training label randomized through Y-randomization. The NN-FP model training history (15 epochs) showed a progressive increase in accuracy and reduction in model loss while the test-set and validation-set had an AUC of 0.99 (Figure 3). The RF-2D model had a sensitivity of 0.47, a specificity of 0.59, and an accuracy of 0.66 over 10-fold cross-validation.

Table 5: NN-FP model performance on different datasets and validation protocols (Cross-validation and Y-randomization)

Model evaluation	Loss	Sensitivity	Specificity	Accuracy
Training-set	0.0398	0.8779	0.8779	0.8779
Test-set	0.0146	0.9607	0.9607	0.9607
Validation-set	0.0146	0.9607	0.9607	0.9607
Cross-validation	0.03	0.9076	0.9076	0.9076
Y-Randomization	0.149	0.5476	0.54699	0.54807

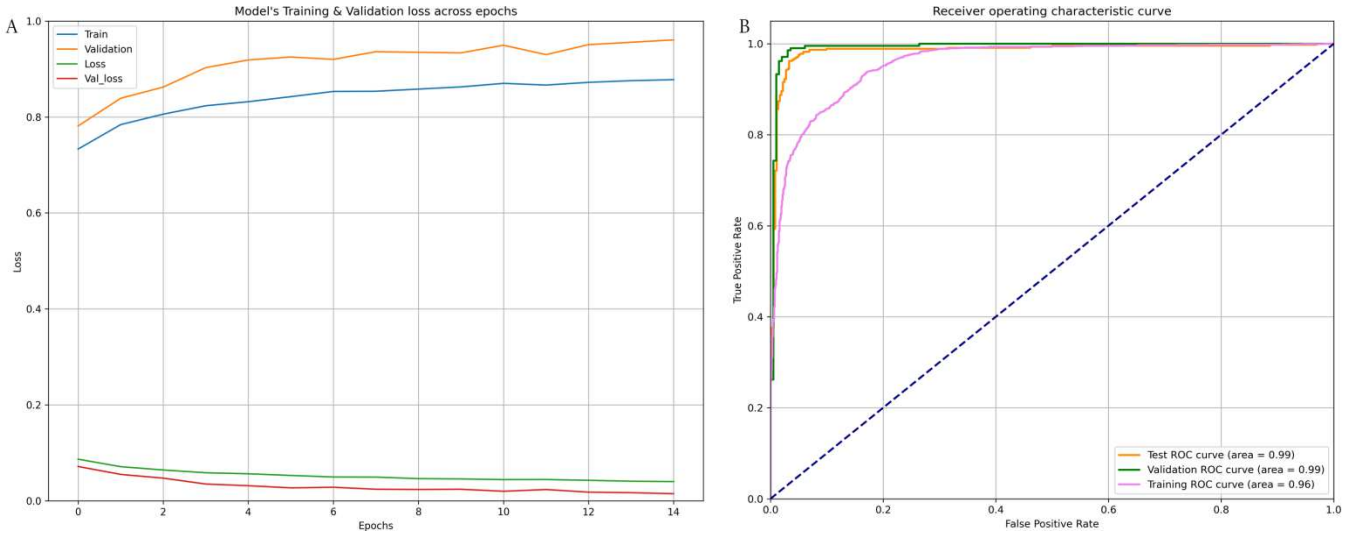

Figure 3: NN-FP model evaluation a) accuracy and loss over 15 epoch b) ROC-CURVEs*: Training-set: AUC 0.95; Test-set: AUC 0.99; Validation-set: AUC 0.99*

The LSTM-SM model was used to screen the ChemDiv database and a total of 15,000 compounds were predicted as actives. These compounds were subjected to molecular docking (to filter out compounds with low binding affinity) and induced-fit docking (to elucidate plausible binding modes indicating putative mechanism of inhibition and specificity).

Molecular Docking and Induced-fit Docking

The docking protocol was validated (see methods); superimposing RSMD was 0.789Å and the AUC of the enrichment ROC curve was 0.99 (Figure 4). The top 12 compounds (resulting from molecular docking) with binding affinities ranging from -103.61 kcal/mol to -90.26 kcal/mol were selected for induced-fit docking (criteria for selection was MMGBSA calculated binding affinity). Induced-fit docking poses had binding affinities ranging from -144.09 kcal/mol to -100.22 kcal/mol; top four compounds were: 2912 (-144.06kcal/mol), 3488 (-132.70kcal/mol), 5277 (-125.6kcal/mol), and 1717 (-124.36kcal/mol) (Table 5). The co-crystallized ligand (AZD4547) was selected as control/standard; its molecular docking pose had a binding affinity of -126.84 kcal/mol and induced-fit docking pose -139.25kcal/mol.

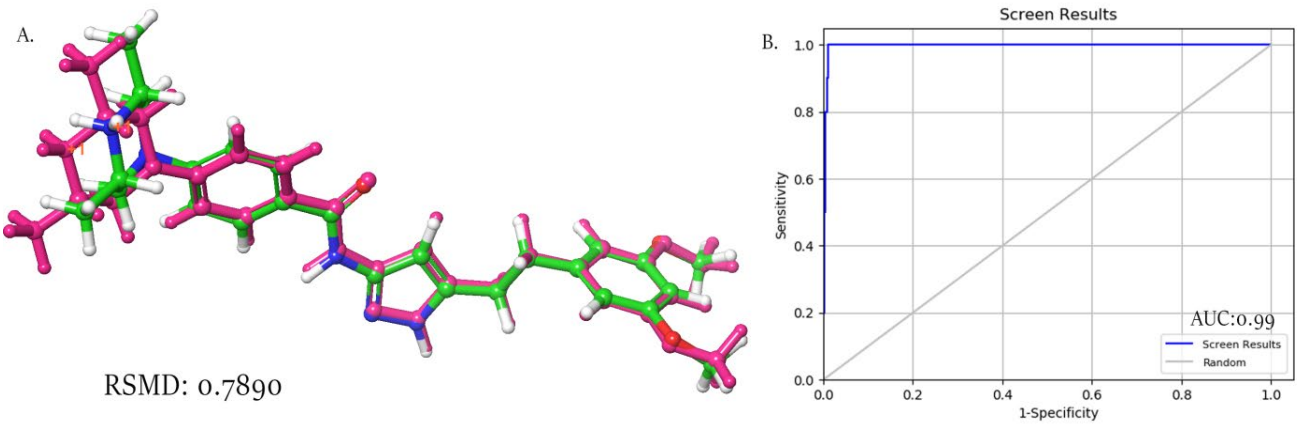

Figure 4: Validation of docking protocol: a) Superimposing of redocked co-crystallized ligand pose (magenta) with the crystalized pose (green) RSMD*: 0.789Å b) ROC curve of the docking protocol enrichment study (AUC:0.99).*

Table 6: Molecular Docking and Induced-fit Docking Binding Affinity

Compound-ID	Molecular docking (kcal/mol)	Induced-fit docking (kcal/mol)
AZD4547	-126.84	-139.25
2912	-103.61	-144.06
3488	-100.59	-132.7
1717	-96.64	-124.36
7110	-93.88	-111.03
875	-93.66	-116.63
3634	-92.49	-100.51
6302	-91.94	-111.1
5550	-91.7	-114.19
4191	-91.55	-119.63
1449	-91.36	-107.61
9800	-91.14	-100.22
5227	-90.26	-125.64

Examining the comparative optimal binding poses of each of the top four hit compounds and AZD4547 (with respect to the calculated optimal binding conformation of the target post-induced fit docking) as well as the interactions between the compounds and different key FGFR1 active site regions including the hinge region (Glu562 – Lys566), P-loop (Lys482 – Leu494), α-C-helix (Gly531), gateway residue (Val561), and DFG-motif (Asp641, Phe642, and Gly643) shows varied binding poses and interactions as shown in figures 5, 6, 7.

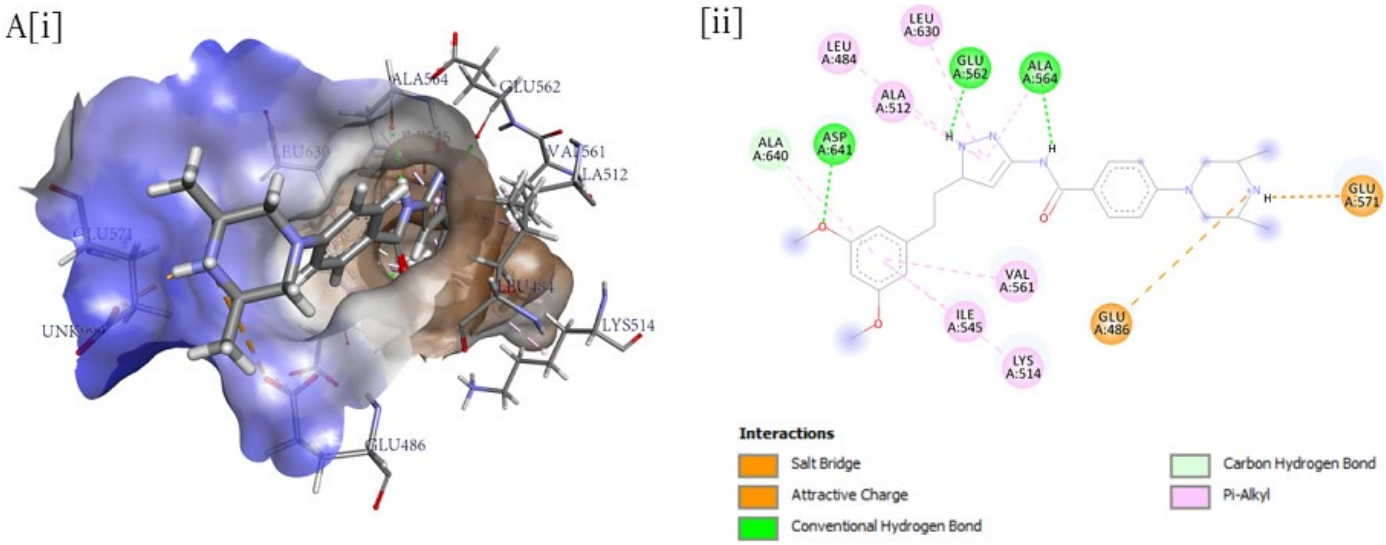

Figure 5: Induced-fit docking of AZD4547: binding affinity -139.25kcal/mol;*[i] 3D interactions B[ii] 2D interactions*

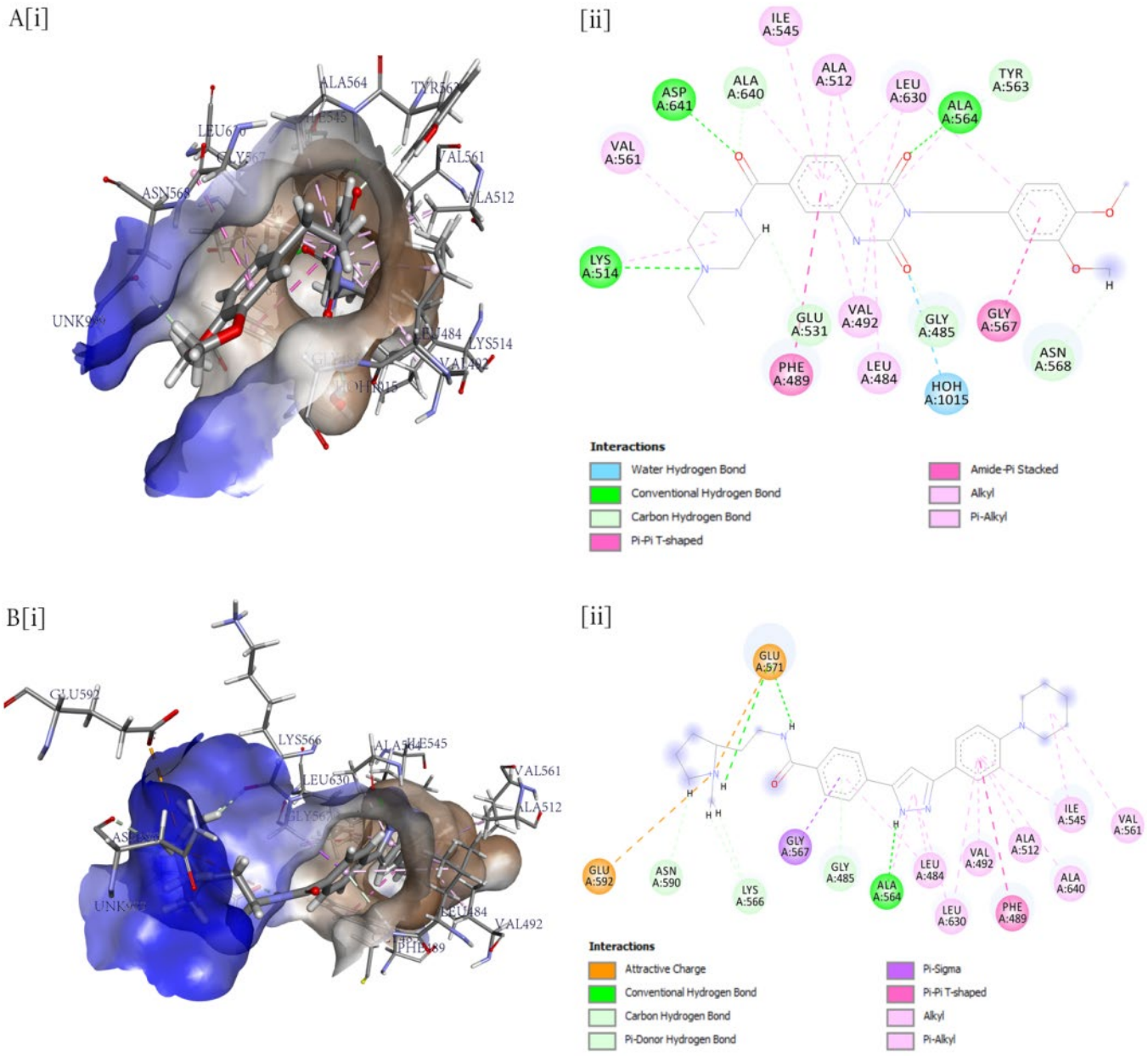

Figure 6: Induced-fit Docking of a)Compound 2912: binding affinity -144.06kcal/mol b) Compound 3448*: binding affinity -132.70 kcal/mol; [i] 3D interactions [ii] 2D interactions*

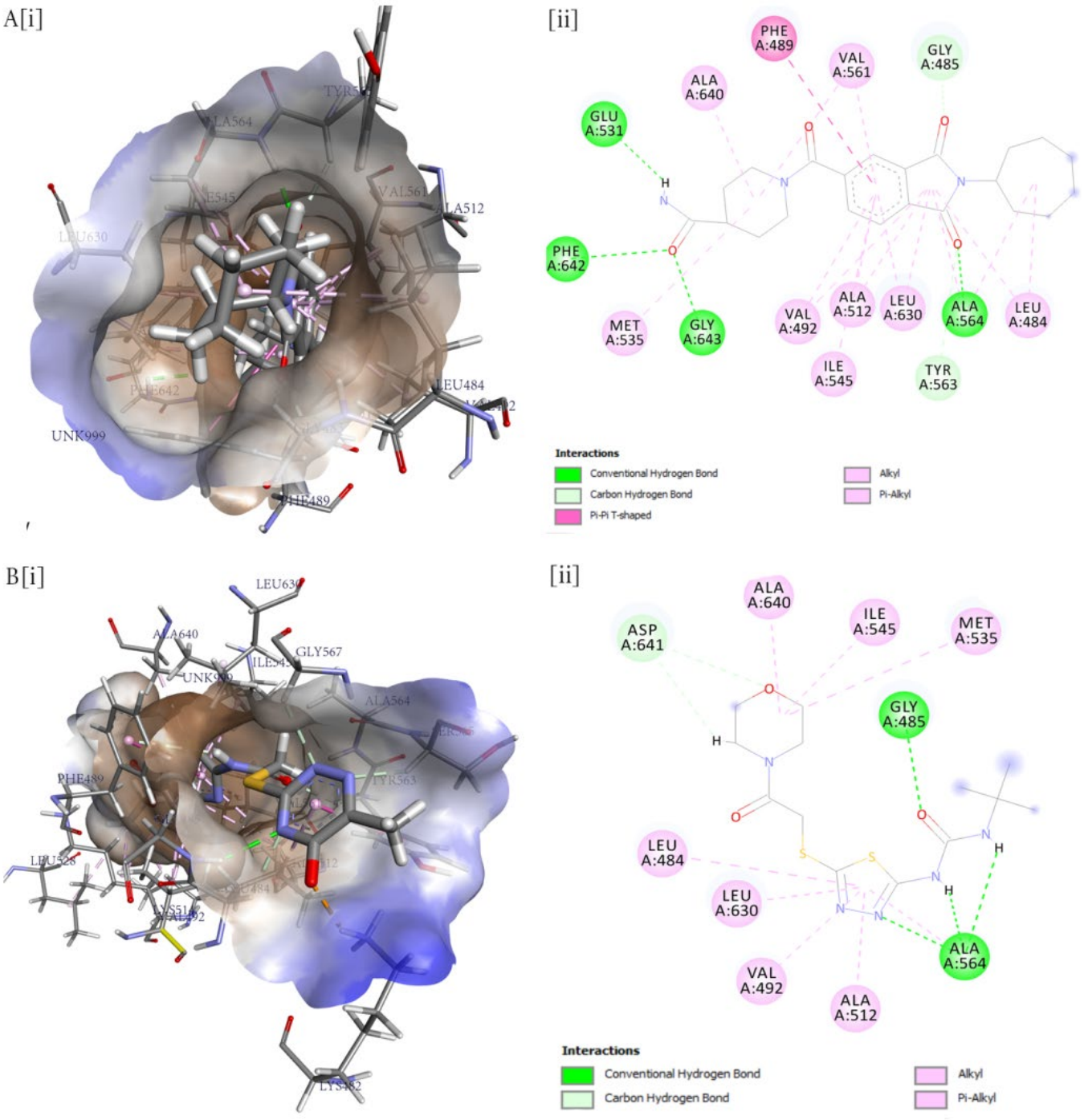

Figure 7: Induced-Fit Docking of a) Compound 5227: binding affinity -125.64kcal/mol b)Compound 1717 binding affinity -124.36 kcal/mol; *[i] 3D interactions [ii] 2D interaction*

QM-MM optimization

The QM-MM calculation was employed to optimize the predicted induced-fit poses. AZD4547 (standard) formed a new hydrophobic bond (π-π stacked) with Phe489; compound 2912 lost its hydrogen bond with the water moiety; compound 3488 lost its bonding interactions with Phe489 and Gly567; compound 5227 lost its interaction with Phe489; compound 1717 gained a hydrophobic interaction with Phe489 and two hydrogen bonds with Glu531, but lost its interaction with Val492 (figures 8, 9). We further observed a reduction in the binding affinity of compound AZD4547, 2912, 5227, 1717, and an increase in compound 3488 binding affinity (Table 7). The QM-MM optimized poses were then selected as final poses and binding affinities.

Table 7: Comparative Binding Affinity of Top Hit Compoundsfollowing Induced-fit Docking and QM-MM Optimization

Compound ID	Induced-fit docking (kcal/mol)	QM-MM optimization (kcal/mol)
AZD4547	-139.25	-136.63
2912	-144.06	-135.85
3488	-132.70	-133.45
5227	-125.64	-123.88
1717	-124.64	-122.08

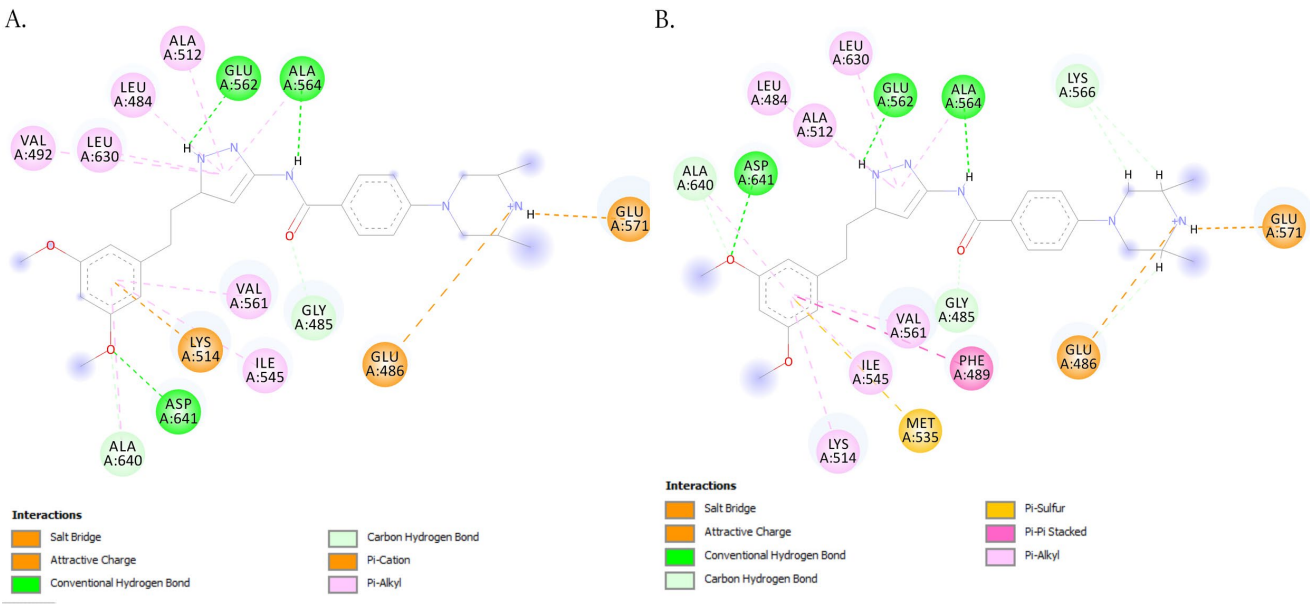

Figure 8: QM-MM Optimization of AZD4547 Induced-fit Poses: *a) Un-optimized Induced-fit pose b) Optimized Induced-fit pose*

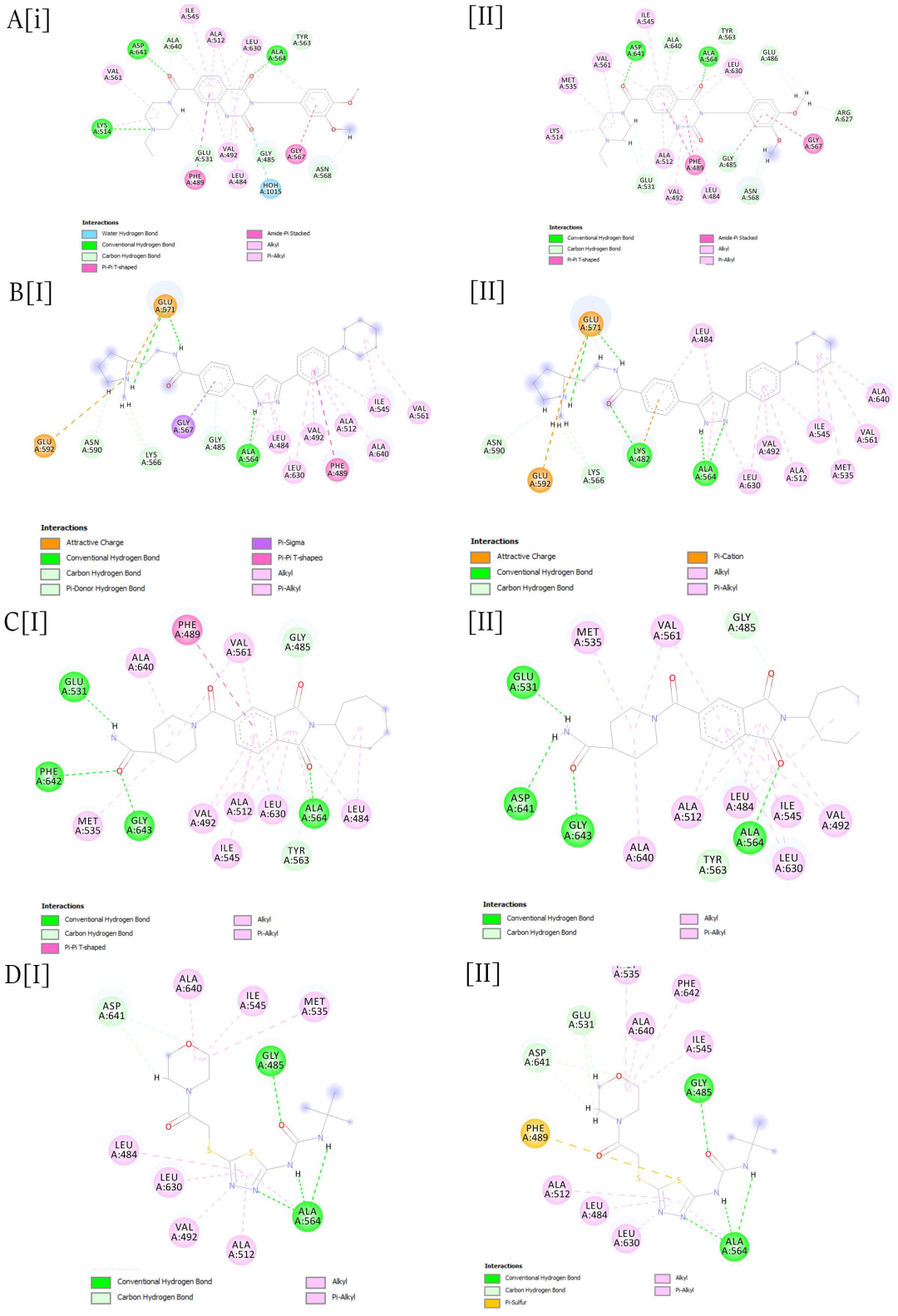

Figure 9: QM-MM Optimization of Induced-fit Poses: *a) Compound 2912 b) 3488 c) 5227 d) 1717 (i: Un-optimized Induced-fit pose ii: Optimized Induced-fit pose)*

Gateway Residue Mutation Analysis

We mutated Val561 to Met561 to investigate possible interactions with this mutated residue; compound 3488 and compound 5227 formed π-alkyl interactions with Met561 and compound 2912 formed π-sulfur interactions with Met561 as shown in figure 10. The compound 1717 was not subjected to this experiment as it did not make any initial interaction with Val561.

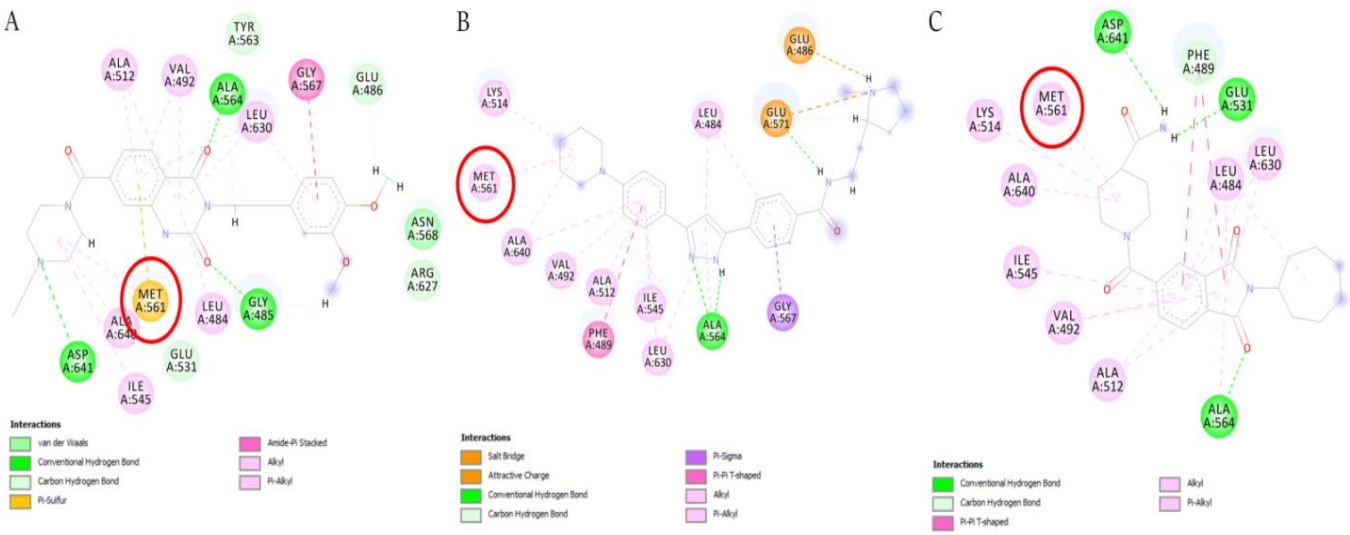

Figure 10: Induced-fit Docking of Mutated FGFR1 Protein (Val531 to Met531): *a) Compound 2912 b) Compound 3488 c) Compound 5227*

Discussion

From the results, we find that the descriptor-free QSAR (LSTM-SM) effectively modeled the FGFR1 inhibitor dataset: the cross-validated model had an accuracy, sensitivity, and specificity of 0.928, and a model loss of 0.025, randomization of the training label (Y-randomization) resulted in a significant reduction in model performance (accuracy, specificity, sensitivity: 0.617) thus eliminating possibility of chance correlation [33,34].

When compared with conventional QSAR models (neural network fingerprint model (NN-FP) and Random Forest 2D descriptors model (RF-2D)), we find that the LSTM-SM model performed slightly better than the NN-FP model (LSTM-SM accuracy, specificity, and sensitivity: 0.928; NN-FP accuracy, specificity, and sensitivity: 0.907) and outperformed the RF-2D model (RF-2D sensitivity: 0.47, specificity: 0.59, and accuracy: 0.66).

LSTM-SM model screened the ChemDiv dataset and 15,000 compounds were predicted as actives. Of the 15,000 compounds predicted as actives, four compounds (2912, 3488, 5227, and 1717) are presented in this study—after being subjected to molecular docking, induced-fit docking, and QM-MM optimization (selection criterion was binding affinity).

We utilized the predicted interactions with key active site residues to suggest putative mechanism of inhibition and specificity. Our suggestions are based on observed interactions of experimentally validated inhibitors. Inhibitors of FGFR1 (non-covalent) developed so far inhibits via two mechanisms: type I and II. Type I (e.g. AZD4547) inhibits FGFR1 in its DFG-in conformation via interaction with Asp641 thus interrupting the coordination of ATP phosphate group [35,36]; Type II (e.g. Ponatinib) inhibits FGFR1 in its DFG-out conformation and forms interactions with the conserved Glu531 of the αC helix region [36,37].

Specificity for FGFR1 occurs via interactions with certain regions of FGFR1 active site. This interactions includes: interactions with Phe489 of the P-loop region which induces its closure (P-loop closure) over the inhibitor thereby creating a better fit; the conserved sites in the FGFR family makes most inhibitors to be active over a wide range of FGFRs (pan-FGFR inhibitors), but the difference in specific residue positions (Tyrosine, Cysteine, or Phenylalanine) in the hinge regions of FGFRs can be exploited for specificity; Tyr563 have been identified in FGFR1 [38]. Val561 confers a natural resistance on FGFR1 via steric hindrance [36] thereby serving as a

gateway residue for FGFR1 the ability to form bonds with this residue also suggest specificity. Finally, the ability to interact with hinge residues (Glu562 – Lys566) also suggests a level of specificity [36].

We, therefore, suggest the following: compound 2912 exhibits both type I and II features (interaction with Asp641 and Glu531 is observed), with its mechanism of specificity via interactions with Phe489 (P-loop), Val561 (gateway residue), Ala564, and Try563 (FGFR1 hinge residues); mechanism of inhibition for Compound 3488 is not clear (absence of bonding interactions with Asp641 or Glu531), however, specificity mechanism could be via interaction with Val561, Ala564, and Try563; Compound 5227 could inhibit via type II mechanism (interaction with Glu531) and specificity via interaction with Glu531, Phe642, Ala564, Try563, and Val561; Compound 1717 also exhibits both type I and II features (interaction with Asp641 and Glu531), with specificity via interactions with Phe489 and Ala564.

The optimized AZD4547 (standard for the study) pose saw interactions which were consistent with experimentally observed interactions, most importantly interactions with Phe489 (P-loop), Asp641 (DFG-motif), Ala564 and Glu562 (Hinge residues) [35]. Furthermore, compound 2912, 3488 and AZD4547 showed similar binding affinity for FGFR1 (-135kcal/mol, -133kcal/mol and -136kcal/mol respectively) suggesting a potential similar biological activity.

Acquired resistance is a challenge when considering long term efficacy of FGFR1 inhibitors, this resistance occurs via mutation of the gateway residue with a bulky amino acid e.g. methionine or isoleucine (resulting in the "gates been closed") [36,37,39]. Simulating this mutation(Val561 to Met561) we predict that the compounds might still be effective regardless of such mutations since the compounds still interacted with Met561: compound 2912 formed π-sulfur interaction, compound 3488, and 5227 from π-alkyl interaction suggesting effectiveness despite this acquired resistance. For compound 1717, interactions with Val561 or Met561 were absent.

With this we submit compounds 2912, 5227, and 3488 as potential specific inhibitors of FGFR1; however, the exact mechanism of inhibitions for the compounds needs to be experimentally verified. Compound 1717 (despite its high binding affinity) is not considered an effective inhibitor due to its in ability to interact with the gateway residue.

We however, recognize that training a descriptor-less QSAR model is computationally intensive, but with advances in computing power this should be a non-issue (it is therefore, a choice between either speed or a QSAR model based purely on SMILES representation); also the stochastic nature of neural networks might result in some compound been missed.

For future perspective, we recommend further tuning of hyperparameters (this could reduce the number of epochs required to train), training with more FGFR1 inhibitors from other databases to make the model more robust, bidirectional LSTM algorithms could be experimented with, improvement or invention of a new textual representation of compounds purely for descriptor-less QSAR modeling to ensure that attention values are easily mapped back to compound structure is also recommended. Finally, the suggested mechanism of inhibition and specificity of the compounds remain predictions and experimental validation is needed.

Conclusion

With the advent of more sophisticated machine learning algorithms capable of self-feature extractions, we have shown that by allowing the model to extract its own descriptor/features, it can perform as good as feeding the model with descriptors if not better; we have also shown its effectiveness in screening for active compounds and elucidated a putative mechanism of inhibition and specificity for selected compounds. Hence, we can affirm that “descriptor calculation, preparation, filtering, and selection steps in the QSAR workflow can be eliminated”.

References

[1] Muratov, E.N., Bajorath, J., Sheridan, R.P., Tetko, I.V., Filimonov, D., Poroikov, V., et al. QSAR without borders. Chem Soc Rev 49, 3525–64, https://doi.org/10.1039/d0cs00098a (2020).

[2] Tandon, H., Chakraborty, T., Suhag, V.A. Concise Review on the Significance of QSAR in Drug Design. Chem Biomol Eng 4, 45, https://doi.org/10.11648/j.cbe.20190404.11 (2019).

[3] Muhammad U, Uzairu A, Ebuka Arthur D. Review on: quantitative structure activity relationship (QSAR) modeling. J Anal Pharm Res 7, 240–2, https://doi.org/10.15406/japlr.2018.07.00232 (2018).

[4] Mauri, A., Consonni, V., Todeschini, R. Molecular Descriptors. Handb Comput Chem 2016, 1–29. https://doi.org/10.1007/978-94-007-6169-8_51-1 (2016).

[5] Idakwo, G., Luttrell, I.V.J., Chen, M., Hong, H., Gong, P., Zhang, C. A Review of Feature Reduction Methods for QSAR-Based Toxicity Prediction. Challenges Adv Comput Chem Phys 30, 119–39, https://doi.org/10.1007/978-3-030-16443-0_7 (2019).

[6] Goodarzi, M., Dejaegher, B., Heyden, Y.V. Feature selection methods in QSAR studies. J AOAC Int 95, 636–51, https://doi.org/10.5740/jaoacint.SGE_Goodarzi (2012).

[7] Khan, P.M., Roy, K. Current approaches for choosing feature selection and learning algorithms in quantitative structure–activity relationships (QSAR). Expert Opin Drug Discov 13, 1075–89, https://doi.org/10.1080/17460441.2018.1542428 (2018).

[8] Chakravarti, S.K., Alla S.R.M. Descriptor Free QSAR Modeling Using Deep Learning With Long Short-Term Memory Neural Networks. Front Artif Intell 2019, 2. https://doi.org/10.3389/frai.2019.00017 (2019).

[9] Travis, W.D. Pathology of Lung Cancer. Clin Chest Med 32, 669–92, https://doi.org/10.1016/j.ccm.2011.08.005 (2011).

[10] Rekhtman, N., Paik, P.K., Arcila, M.E., Tafe, L.J., Oxnard, G.R., Moreira, A.L., et al. Clarifying the spectrum of driver oncogene mutations in biomarker-verified squamous

carcinoma of lung: Lack of EGFR/KRAS and presence of PIK3CA/AKT1 mutations. Clin Cancer Res 18, 1167–76, https://doi.org/10.1158/1078-0432.CCR-11-2109 (2012).

[11] Marchetti, A., Martella, C., Felicioni, L., Barassi, F., Salvatore, S., Chella, A., et al. EGFR mutations in non-small-cell lung cancer: Analysis of a large series of cases and development of a rapid and sensitive method for diagnostic screening with potential implications on pharmacologic treatment. J Clin Oncol 23, 857–65, https://doi.org/10.1200/JCO.2005.08.043 (2005).

[12] Lemmon, M.A., Schlessinger, J. Cell signaling by receptor tyrosine kinases. Cell 141, 1117–34, https://doi.org/10.1016/j.cell.2010.06.011 (2010).

[13] Turner, N., Grose, R. Fibroblast growth factor signalling: From development to cancer. Nat Rev Cancer 10, 116–29, https://doi.org/10.1038/nrc2780 (2010).

[14] Haugsten, E.M., Wiedlocha, A., Olsnes, S., Wesche, J. Roles of fibroblast growth factor receptors in carcinogenesis. Mol Cancer Res 8, 1439–52, https://doi.org/10.1158/1541-7786.MCR-10-0168 (2010).

[15] Weiss, J., Sos, M.L., Seidel, D., Peifer, M., Zander, T., Heuckmann, J.M., et al. Frequent and focal FGFR1 amplification associates with therapeutically tractable FGFR1 dependency in squamous cell lung cancer. Sci Transl Med 2012, 4. https://doi.org/10.1126/scitranslmed.3004128 (2012).

[16] Sleeman, M., Fraser, J., McDonald, M., Yuan, S., White, D., Grandison, P., et al. Identification of a new fibroblast growth factor receptor, FGFR5. Gene 271, 171–82, https://doi.org/10.1016/S0378-1119(01)00518-2 (2001).

[17] Abadi, M., Paul, B.J., Chen, I., Chen, Z., Davis, A., Dean, J., et al. TensorFlow: A System for Large-Scale Machine Machine Learning. 12th USENIX Symp. Oper. Syst. Des. Implement. https://doi.org/10.1016/0076-6879(83)01039-3 (2016).

[18] Varoquaux, G., Buitinck, L., Louppe, G., Grisel, O., Pedregosa, F., Mueller, A. Scikit-learn. GetMobile Mob Comput Commun 19, 29–33, https://doi.org/10.1145/2786984.2786995 (2015).

[19] Gaulton, A., Hersey, A., Nowotka, M.L., Patricia, B.A., Chambers, J., Mendez, D., et al. The ChEMBL database in 2017. Nucleic Acids Res 45, D945–54, https://doi.org/10.1093/nar/gkw1074 (2017).

[20] Chen, T., Guestrin, C. XGBoost: A Scalable Tree Boosting System. Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min 42, 785–94, https://doi.org/doi.org/10.1145/2939672.2939785 (2016).

[21] Ashton, M., Barnard, J., Casset, F., Charlton, M., Downs, G., Gorse, D., et al. Identification of diverse database subsets using property-based and fragment-based molecular descriptions. Quant Struct Relationships 21, 598–604, https://doi.org/10.1002/qsar.200290002 (2002).

[22] Sepp, H., Jurgen, S. Long Short-Term Memory. Neural Comput 9, 1735–80, https://doi.org/10.1162/neco.1997.9.8.1735 (1997).

[23] M H., M.N S. A Review on Evaluation Metrics for Data Classification Evaluations. Int J Data Min Knowl Manag Process 5, 01–11, https://doi.org/10.5121/ijdkp.2015.5201 (2015).

[24] Melagraki, G., Afantitis, A. Enalos KNIME nodes: Exploring corrosion inhibition of steel in acidic medium. Chemom Intell Lab Syst 123, 9–14, https://doi.org/10.1016/j.chemolab.2013.02.003 (2013).

[25] Madhavi, S.G., Adzhigirey, M., Day, T., Annabhimoju, R., Sherman, W. Protein and ligand preparation: Parameters, protocols, and influence on virtual screening enrichments. J Comput Aided Mol Des 27, 221–34, https://doi.org/10.1007/s10822-013-9644-8 (2013).

[26] Jacobson, M.P., Pincus, D.L., Rapp, C.S., Day, T.J.F., Honig, B., Shaw, D.E., et al. A Hierarchical Approach to All-Atom Protein Loop Prediction. Proteins Struct Funct Genet 55, 351–67, https://doi.org/10.1002/prot.10613 (2004).

[27] Shelley, J.C., Cholleti, A., Frye, L.L., Greenwood, J.R., Timlin, M.R., Uchimaya, M. Epik: A software program for pKa prediction and protonation state generation for drug-like molecules. J Comput Aided Mol Des 21, 681–91, https://doi.org/10.1007/s10822-007-

9133-z (2007).

[28] Harder, E., Damm, W., Maple, J., Wu, C., Reboul, M., Xiang, J.Y., et al. OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. J Chem Theory Comput 12, 281–96, https://doi.org/10.1021/acs.jctc.5b00864 (2016).

[29] Halgren, T.A., Murphy, R.B., Friesner, R.A., Beard, H.S., Frye, L.L., Pollard, W.T., et al. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. J Med Chem 47, 1750–9, https://doi.org/10.1021/jm030644s (2004).

[30] Lyne, P.D., Lamb, M.L., Saeh, J.C. Accurate prediction of the relative potencies of members of a series of kinase inhibitors using molecular docking and MM-GBSA scoring. J Med Chem 49, 4805–8, https://doi.org/10.1021/jm060522a (2006).

[31] Bochevarov, A.D., Harder, E., Hughes, T.F., Greenwood, J.R., Braden, D.A., Philipp, D.M., et al. Jaguar : A High-Performance Quantum Chemistry Software Program with Strengths in Life and Materials Sciences. Int J Quantum Chem 2013, 2110–2142, https://doi.org/10.1002/qua.24481 (2013).

[32] Singh, N., Villoutreix, B.O., Ecker, G.F. Rigorous sampling of docking poses unveils binding hypothesis for the halogenated ligands of L-type Amino acid Transporter 1 (LAT1). Sci Rep 9, 1–20, https://doi.org/10.1038/s41598-019-51455-8 (2019).

[33] Rücker, C., Rücker, G., Meringer, M. Y-randomization and its variants in QSPR/QSAR. J Chem Inf Model 47, 2345–57, https://doi.org/10.1021/ci700157b (2007).

[34] Melagraki, G., Afantitis, A., Sarimveis, H., Koutentis, P.A., Kollias, G., Igglessi-Markopoulou, O. Predictive QSAR workflow for the in silico identification and screening of novel HDAC inhibitors. Mol Divers 13, 301–11, https://doi.org/10.1007/s11030-009-9115-2 (2009).

[35] Yosaatmadja, Y., Patterson, A.V., Smaill, J.B., Squire, C.J. The 1.65 Å resolution structure of the complex of AZD4547 with the kinase domain of FGFR1 displays exquisite molecular recognition. Acta Crystallogr Sect D Biol Crystallogr 71, 525–33,

https://doi.org/10.1107/S1399004714027539 (2015).

[36] Shuyan, D., Zhou, Z., Chen, Z., Xu, G., Chen, Y. Fibroblast Growth Factor Receptors (FGFRs): Structures and Small Molecule Inhibitors Shuyan. Cell 2019, 1–15 (2019).

[37] Yoza, K., Himeno, R., Amano, S., Kobashigawa, Y., Amemiya, S., Fukuda, N., et al. Biophysical characterization of drug-resistant mutants of fibroblast growth factor receptor 1. Genes to Cells 21, 1049–58, https://doi.org/10.1111/gtc.12405 (2016)..

[38] Katoh, M. Fibroblast growth factor receptors as treatment targets in clinical oncology. Nat Rev Clin Oncol 2019. https://doi.org/10.1038/s41571-018-0115-y (2019).

[39] Ryan, M.R., Sohl, C.D., Luo, B., Anderson, K.S. The FGFR1 V561M gatekeeper mutation drives AZD4547 resistance through STAT3 activation and EMT. Mol Cancer Res 17, 532–43, https://doi.org/10.1158/1541-7786.MCR-18-0429 (2019).

Authors' contributions

The authors contributed to this work in the following ways: J.I.Y., S.L.A and A.T.O. performed experiments, data analysis and interpretation; J.I.Y., S.L.A., A.T.O., A.R.A., S.T.A., A.T.A., O.M.O. drafted and critically evaluated the manuscript. All authors read and approved the final manuscript.

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

Supplementary Materials

- All Jupyter notebooks implementing the LSTM-SM, NN-FP, and RF-2D models
- The trained LSTM-SM model used for screening ChemDiv database
- LSTM-SM Model architecture
- All FGFR1 inhibitors used for training, testing, and validation

Figures

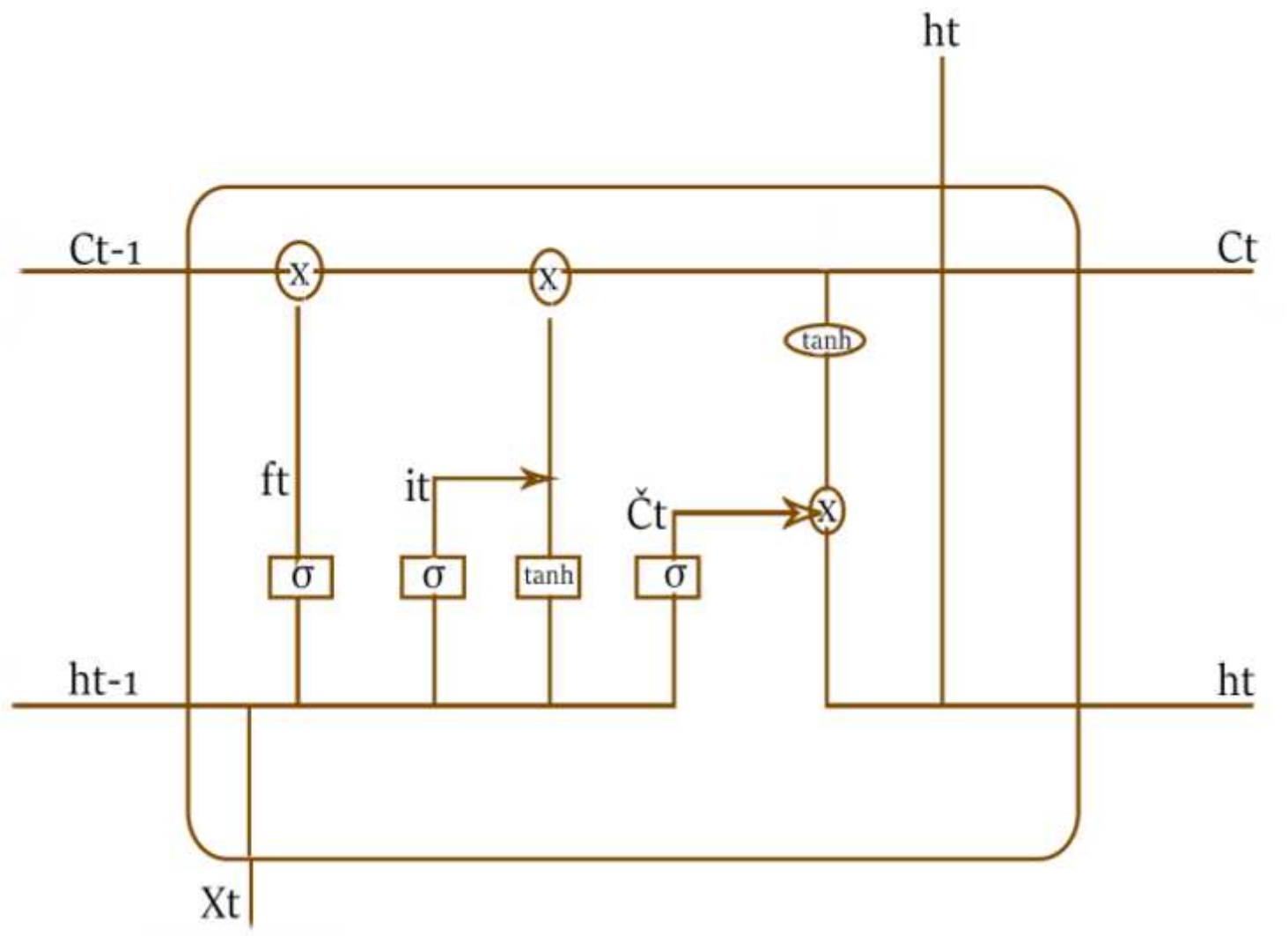

Figure 1

A cell Unit in a LSTM Network(σ: Sigmoid activation Function; Ct-1: previous cell state; ht-1: previous hidden state; ft: forget gate; it: input layer gate; Čt: tanh layer gate; tanh: tanh activation function; ht: new hidden state; Ct: new cell state)

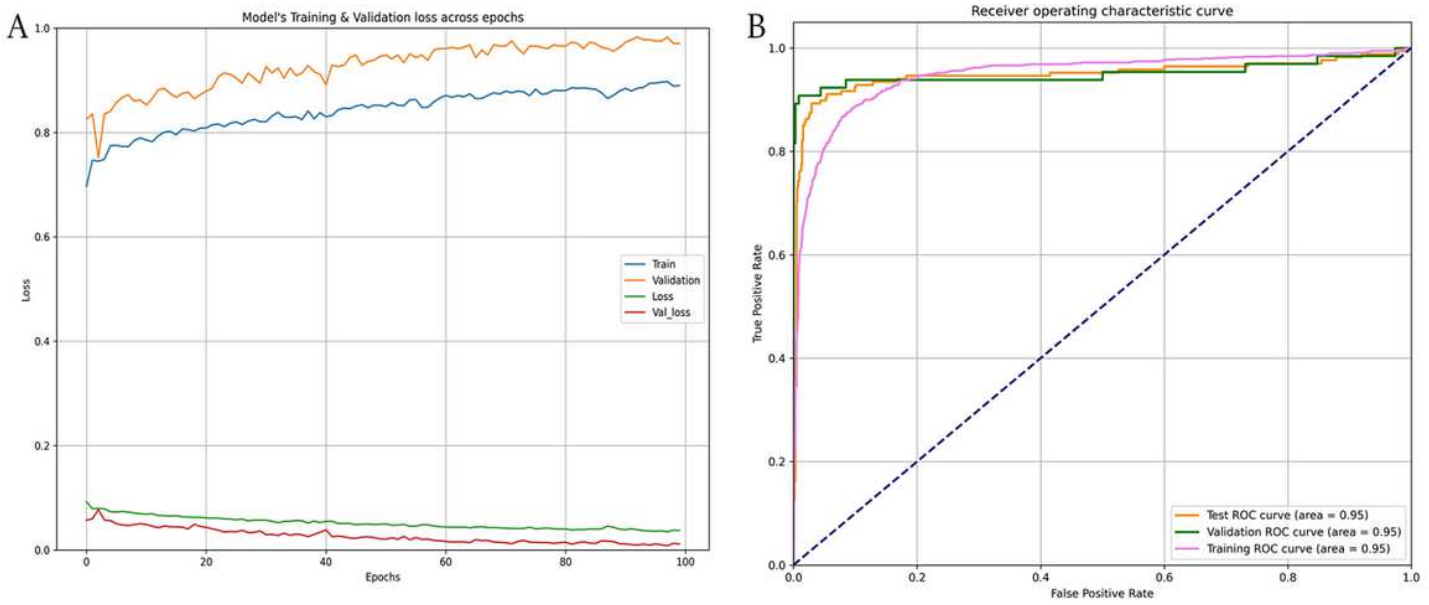

Figure 2

LSTM-SM model evaluation a) accuracy and loss over 100 epoch b) ROC curves: Training-set: AUC 0.95; Test-set: AUC 0.95; Validation-set: AUC 0.95

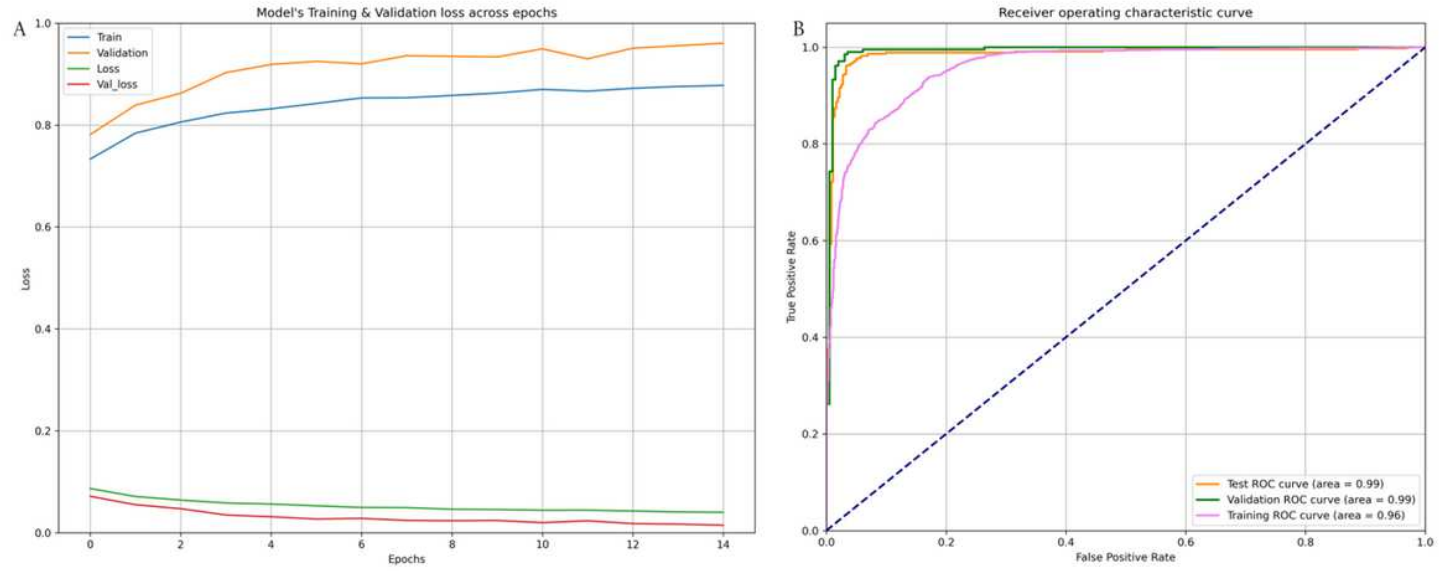

Figure 3

NN-FP model evaluation a) accuracy and loss over 15 epoch b) ROC-CURVEs: Training-set: AUC 0.95; Test-set: AUC 0.99; Validation-set: AUC 0.99

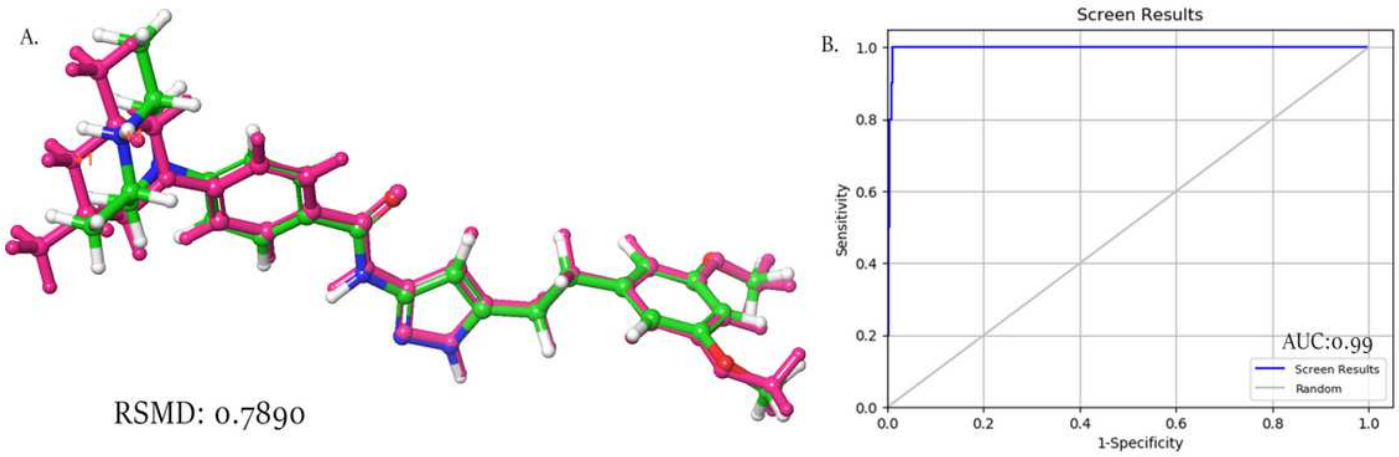

Figure 4

Validation of docking protocol: a) Superimposing of redocked co-crystallized ligand pose (magenta) with the crystalized pose (green) RSMD: 0.789Å b) ROC curve of the docking protocol enrichment study (AUC:0.99).

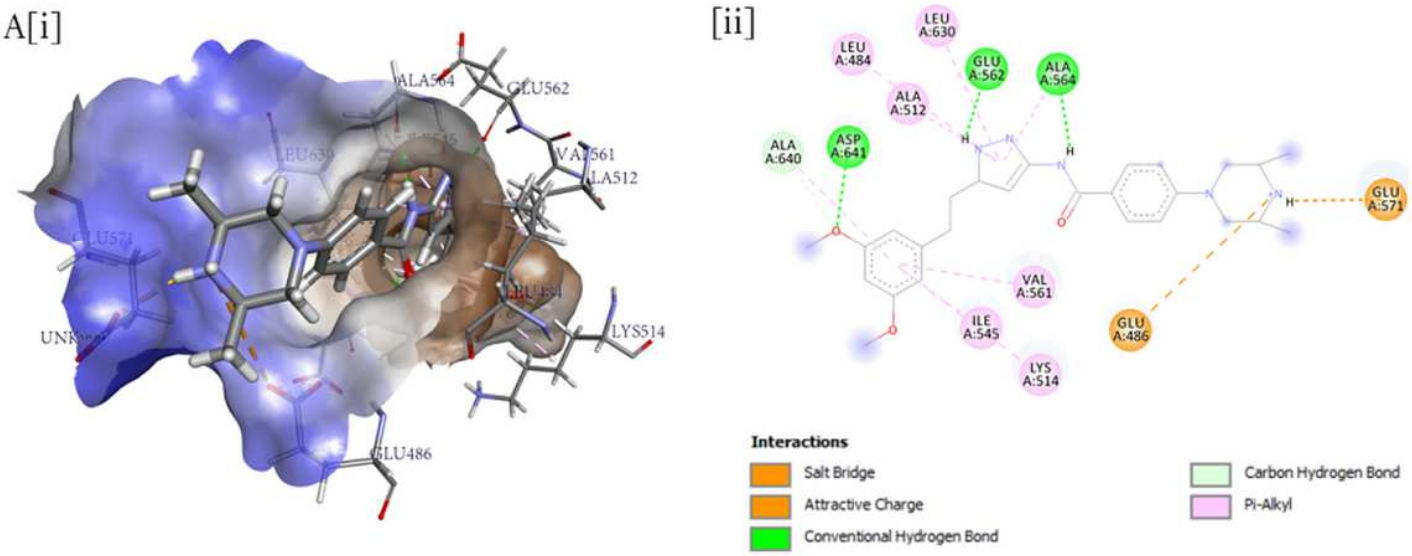

Figure 5

Induced-fit docking of AZD4547: binding affinity -139.25kcal/mol;[i] 3D interactions B[ii] 2D interactions

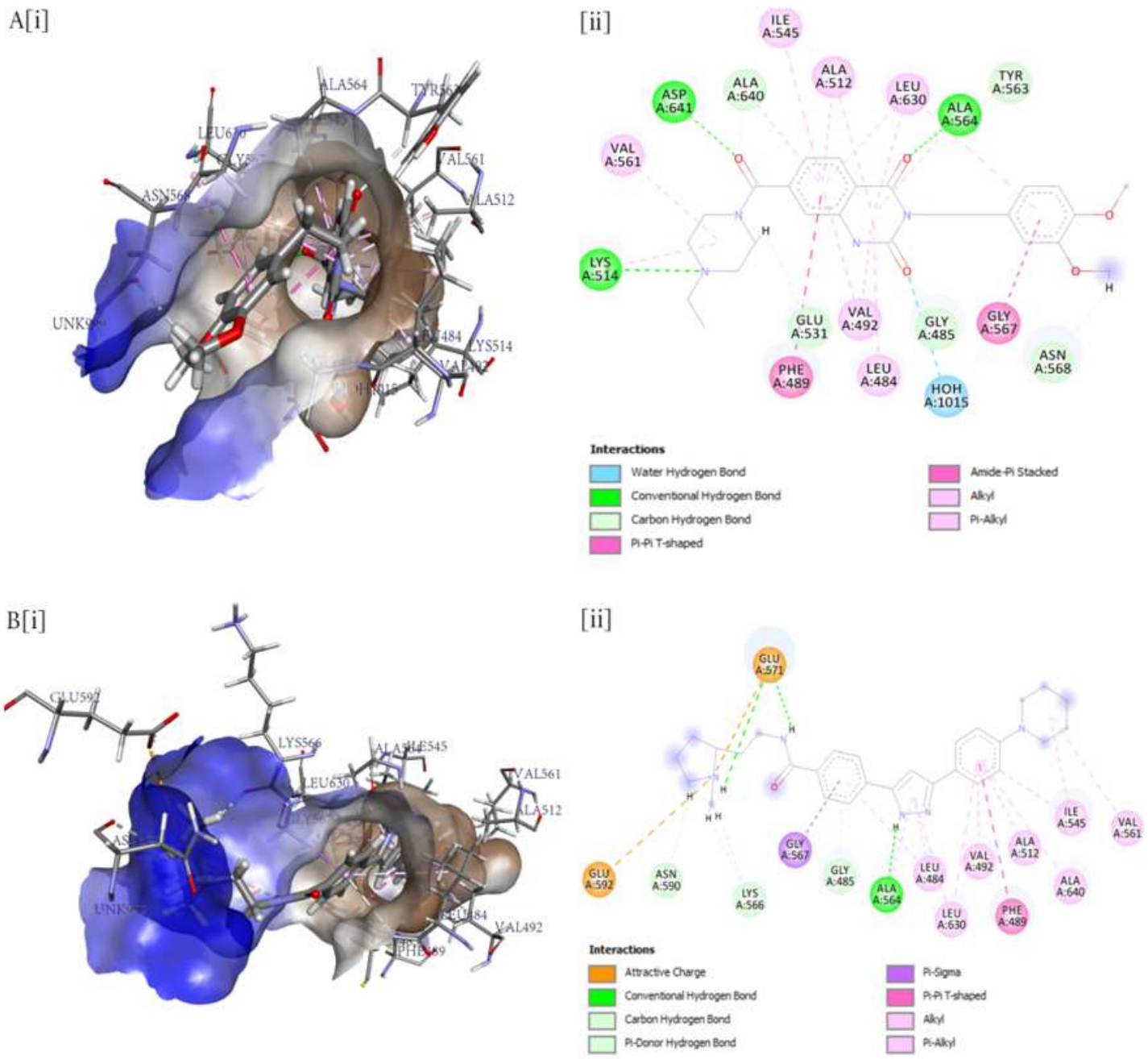

Figure 6

Induced-fit Docking of a)Compound 2912: binding affinity -144.06kcal/mol b) Compound 3448: binding affinity -132.70 kcal/mol; [i] 3D interactions [ii] 2D interactions

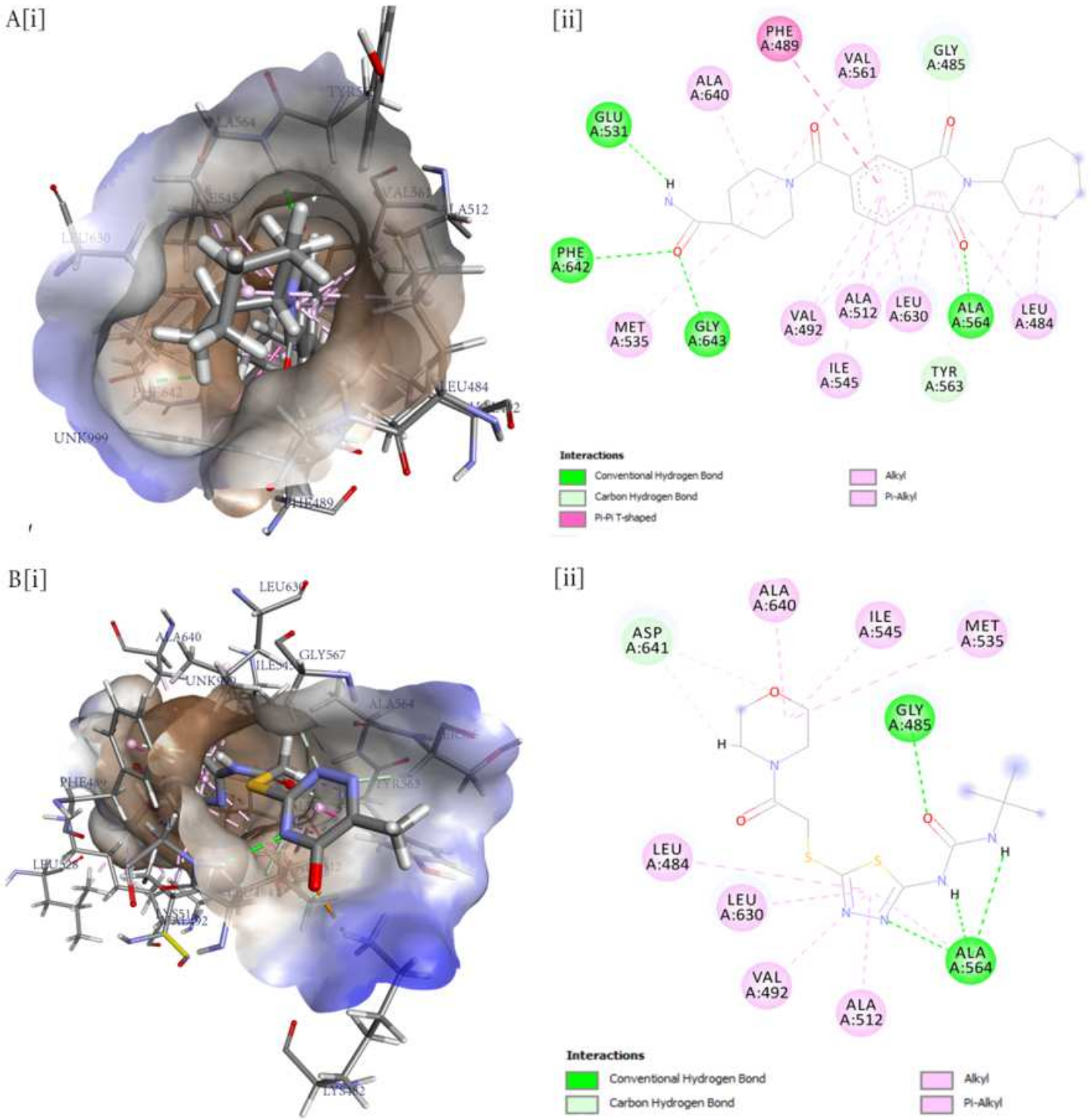

Figure 7

Induced-Fit Docking of a) Compound 5227: binding affinity -125.64kcal/mol b)Compound 1717 binding affinity -124.36 kcal/mol; [i] 3D interactions [ii] 2D interaction

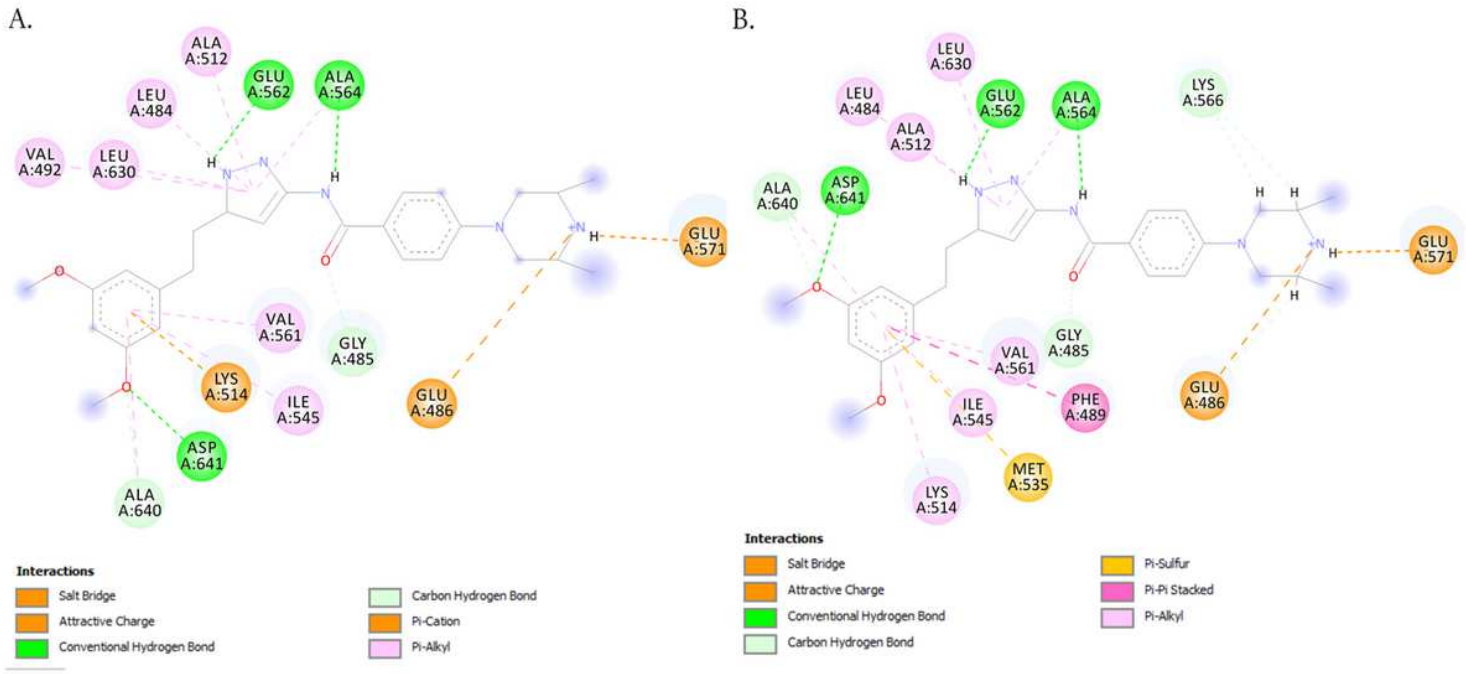

Figure 8

QM-MM Optimization of AZD4547 Induced-fit Poses: a) Un-optimized Induced-fit pose b) Optimized Induced-fit pose

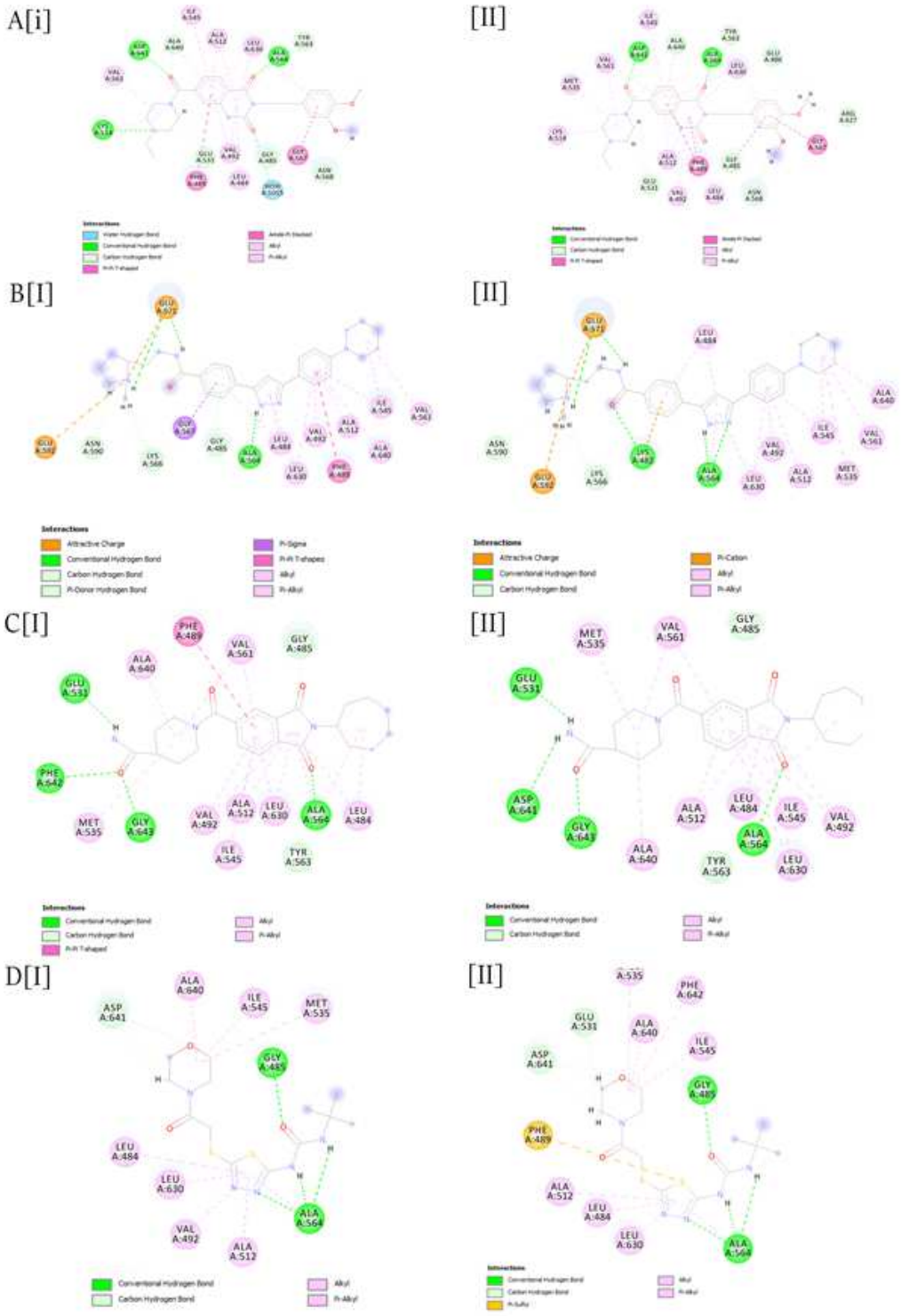

Figure 9

QM-MM Optimization of Induced-fit Poses: a) Compound 2912 b) 3488 c) 5227 d) 1717 (i: Un-optimized Induced-fit pose ii: Optimized Induced-fit pose)

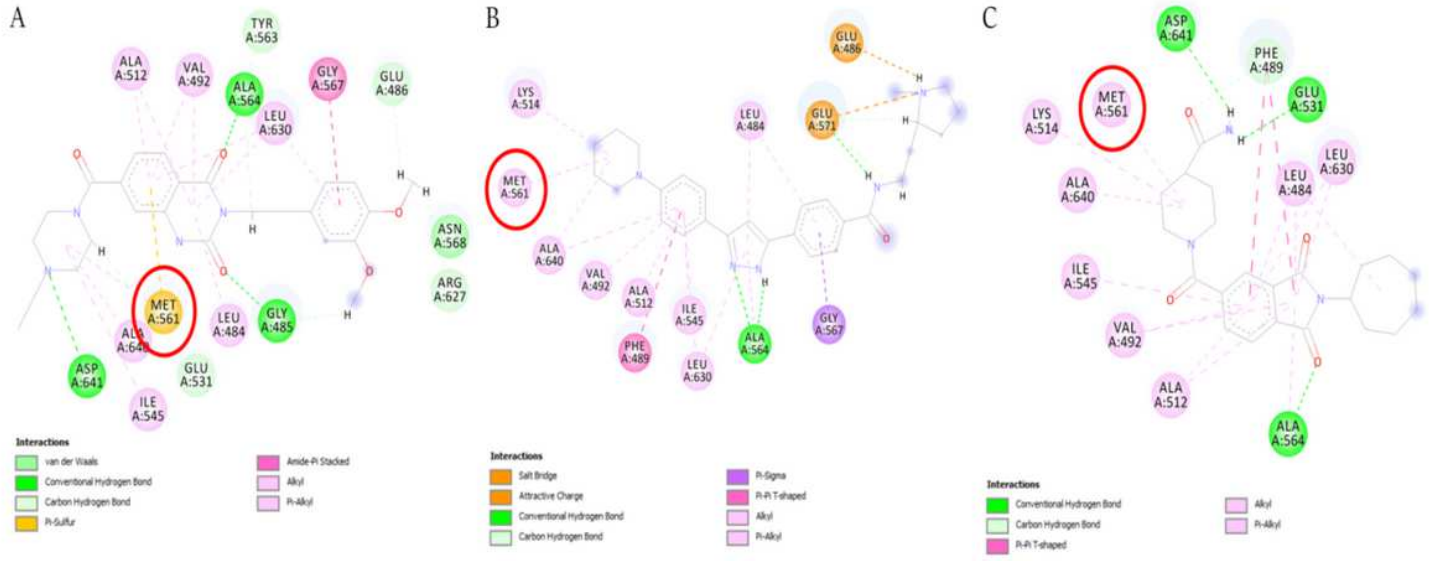

Figure 10

Induced-fit Docking of Mutated FGFR1 Protein (Val531 to Met531): a) Compound 2912 b) Compound 3488 c) Compound 5227