3D-QSAR and relative binding affinity estimation of focal adhesion kinase inhibitors

Suparna Ghosh[1] and Seung Joo Cho[1,2,*]

[1] *Department of Biomedical Sciences, College of Medicine, Chosun University, Gwangju 501-759, Republic of Korea;* s.ghosh@chosun.kr (S.G); chosj@chosun.ac.kr (S.J.C)

[2] *Department of Cellular Molecular Medicine, College of Medicine, Chosun University, Gwangju 501-759, Republic of Korea*

[*] *Correspondence:*

Address: College of Medicine, Chosun University, 375 Seosuk-dong, Dong-gu Gwangju 501-759, Republic of Korea;

E-mail: chosj@chosun.ac.kr;

Telephone: +82-62-230-7482 or +82-11-5479-1010

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 November 2022 doi:10.20944/preprints202211.0356.v1

Abstract

Precise binding affinity predictions are essential for structure-based drug discovery (SBDD). Focal adhesion kinase (FAK) is a member of the tyrosine kinase protein family and is overexpressed in a variety of human malignancies. Inhibition of FAK using small molecules is a promising therapeutic option for several types of cancer. Here, we conducted computational modeling of FAK targeting inhibitors using 3-dimensional structure-activity relationship (3D-QSAR), molecular dynamics (MD), and hybrid topology-based free energy perturbation (FEP) methods. The structure-activity relationship (SAR) studies between the physicochemical descriptors and inhibitory activities of the chemical compounds were performed with reasonable statistical accuracy using CoMFA and CoMSIA. These are two well-known 3D-QSAR methods based on the principle of supervised machine learning (ML). Essential information regarding residue-specific binding interactions was determined using the MD and MM-PB/GBSA methods. Finally, physics-based relative binding free energy ($\Delta\Delta G_{RBFE}^{A\rightarrow B}$) values of analogous ligands were estimated using the alchemical FEP simulation. An acceptable agreement was observed between the experimental and computed relative binding free energies. The overall results using ML and physics-based hybrid approaches could be useful for the rational optimization of accessible lead compounds with similar scaffolds targeting the FAK receptor.

Keywords: Focal adhesion kinase, 3D-QSAR, Molecular Dynamics, MM-PB/GBSA, Free energy perturbation

1. Introduction

Overexpression of the FAK receptor is known for its pivotal role in cell division, proliferation, migration, adhesion, and angiogenesis through its enzymatic activities in different types of cancer progression in humans[1]. FAK, also known as protein tyrosine kinase 2 (PTK2), comprises an N-terminal four-point-one, ezrin, radixin, moesin (FERM) domain, a catalytic kinase domain, and a C-terminal domain (Fig 1)[2]. The FERM domain is further divided into smaller subdomains (F1, F2, and F3), directly bound to the intercellular part of the transmembrane receptor proteins and the binding site for the growth factor receptors, C-Met, p53, and mouse double minute 2 (MDM2) proteins[3]. The highly conserved kinase domain (residue 300-650) participates in the catalytic activity. On the other hand, the C-terminal domain comprises a focal adhesion targeting (FAT) domain and two proline-rich region (PPR) motifs. There are six tyrosine residues as phosphorylation sites (Y397, Y407, Y576, Y577, Y861, and Y925) that are located throughout the FAK receptor and have been identified as critical phosphorylation sites upon binding to signaling proteins[4, 5].

ATP-competitive inhibitors targeting the kinase domain are promising therapeutic interventions for several types of cancers, and many are currently being studied in advanced clinical trials. However, throughout the lead optimization process, there was a persistent dilemma between selectivity and efficacy, demanding more collaborative efforts using computational modeling and medicinal chemistry[6].

Because the binding of inhibitor compounds to target receptors involves contributions of entropy and enthalpy, biophysical and biochemical methods are frequently used to determine binding affinity. However, these procedures are costly, time-consuming, and limited to technical challenges. On the contrary, with the advent of CPU, GPU resources and improved force fields, computational methods have shown dramatic improvement in determining the binding affinity between biomolecules[7, 8]. Methods such as molecular docking, molecular dynamics, MM-PBSA binding free energy, umbrella sampling, free energy perturbation (FEP), and thermodynamic integration (TI) have been developed and effectively used for binding affinity assessment in kinase drug design[9].

In our current work, we conducted the molecular modeling study by taking 125 analogous compounds as FAK inhibitors, which exhibited a wide spectrum of inhibitory activities [10-14]. These compounds are ATP-competitive inhibitors with high structural similarity to TAE226 or TAE molecule. Therefore, the compounds were expected to interact with FAK in a similar manner to TAE226 (PDB: 4D58 and 2JKK)[15, 16]. We developed CoMFA and CoMSIA, two well-known 3D-QSAR methods, to establish the structure-activity relationship of the compounds in the dataset. Unlike 2D-QSAR, 3D-QSAR includes quantum chemical descriptors, unique molecular scaffolds, substituent constants, surface and volume descriptors, and autocorrelation descriptors. This provides richer information and better reflects the non-bonded interaction properties between the receptor and ligands. Additionally, the key structural features of the inhibitors were graphically represented as contour polyhedrons in descriptive color schemes, which are useful for designing new chemical compounds by scaffold hopping or molecular probing. The SAR investigation study was integrated with the residue-specific binding energy profile from the MM-PB/GBSA analysis. The relative binding affinity calculation for a congeneric series of small molecules has gained popularity for lead optimization in the pharmaceutical industry and institutional laboratories over the last decade. We calculated the relative binding free energy ($\Delta\Delta G_{\mathrm{RBFE}}^{\mathrm{A \to B}}$) values by taking 12 compounds and then correlated them with their relative experimental binding free energy ($\Delta\Delta G_{\mathrm{EXP}}^{\mathrm{A \to B}}$) values.

2. Methodology

2.1. Structure preparation

The bis-anilino pyrimidine (BI9)/TAE226 bound FAK complex with the resolution of 1.95 Å was retrieved from the RCSB PDB database (PDB ID 4D58). The crystallographic water molecules and ions were removed, and the missing atoms, side chains, and loops were modeled using the web version of MODELLER in Chimera-1.15, according to our previous studies[17, 18]. SYBYL was used to perform the necessary naming and atom index adjustment of the TAE226 molecule so that it was compatible with the AMBER forcefield during the all-atom MD simulation.

doi:10.20944/preprints202211.0356.v1

2.2. MD simulation and binding energy calculation

The all-atom MD simulation of the protein-ligand complex was conducted by GROMACS version: 2019.5 [19], using the Amber ff03 force field, according to earlier studies[20, 21]. TAE226 or C36 was parameterized using ACEPYPE[22], where atom types were assigned as GAFF types and AM1-BCC partial charge model. The complex was solvated and ionized according to the procedures described in the previous study. Following that, the system was carried out for Minimization, NVT, NPT, and 100 ns of MD production simulations. In the NVT and NPT simulations, a modified Berendsen thermostat and barostat were employed to achieve the 300 K temperature and 1 bar of pressure, respectively. The backbone of protein and the heave atoms of the ligands are restrained during the NVT and NPT ensembles, while they were omitted during the production run. The built-in '*gmx rms*' function was used to calculate the RMSD of the protein and ligand respectively[23]. The MM-GBSA binding energy (ΔG_{bind}), as well as the entropy term (TΔS) between the protein and ligand, was computed using the gmx_MMPBSA[24] package, as described here[25]. The binding energy (ΔG_{bind}) obtained from the MM-PB/GBSA calculation can be expressed as follows:

$$\Delta G_{bind} = \Delta G_{COM} - (\Delta G_{PROT} + \Delta G_{LIG}) \ (1)$$

where ΔG_{COM}, ΔG_{PROT}, and ΔG_{LIG} represent the total free energies of complex, protein, and ligand separately, respectively in the solvent.

2.3. Dataset preparation and molecular modeling

A total of 125 compounds were acquired from previously published literature and their inhibitory activity (IC_{50}) values were translated to -logIC_{50} (pIC_{50}). Compound C36 is already available as bis-anilino pyrimidine (BI9) or TAE226 in high-resolution co-crystallized form bound with FAK (PDB ID 4D58). Besides we employed the MD ensemble to obtain a fully equilibrated protein-ligand structure complex. Therefore, the last frame of C36 from the MD trajectory was considered to be a biological 3D conformer and represented template molecules of the dataset. Based on the template molecule, the rest of the compounds were sketched, minimized, and assigned Gasteiger-Hückel partial charges in SYBYL, as described here[26].

2.4. Development of 3D-QSAR models

The compounds were aligned to the common core using the template molecule (C36) as a reference. The compounds were then classified into low, medium, and high activity classes, and the test set compounds were chosen at random from each class to achieve a final training vs. test set ratio of 3:1. CoMFA and CoMSIA were used to develop 3D-QSAR models. In both methods, the chemical descriptor fields were calculated in a 3D cubic box with a grid spacing of 1 Å. At each grid intersection, a hybridized sp^3 carbon atom with a +1 charge was assigned to compute the steric (S) and electrostatic (E) fields. In CoMSIA, an additional three fields, namely, hydrophobic (H), H-bond acceptor (A), and H-bond donor (D), with a Gaussian function. The partial least squares (PLS) method was used to assess the statistical correlation between the chemical descriptors and inhibitory activities in the CoMFA and CoMSIA models. Leave-one-out and no cross-validation methods were applied to obtain the cross-validation squared correlation coefficient (q^2) and the no cross-validation squared correlation coefficient (r^2) by taking the training set compounds, followed by predicting the pIC_{50} of every compound in the dataset including the test set compounds. The external validation or predictivity of the QSAR models was determined by calculating the predictive squared correlation coefficient or r^2_{pred} values. Additional parameters such as k or k', r_0^2 or r'^2_0, $|r_0^2 - r'^2_0|$, r_m^2 or r'^2_m, $Q^2_{F1,}$ Q^2_{F2}, Q^2_{F3}, and Q^2_{ccc} were also considered for the reliability of the model according to these studies[27, 28]. The applicability domain (AD) of the developed CoMFA and CoMSIA models was evaluated using a distance-based Williams plot according to this study[29]. The field distributions of the descriptors were vividly represented as descriptive colored contours, suggesting favorable and unfavorable chemical substitutions that could increase the inhibitory potency of the lead compounds.

2.5. Relative binding energy calculation

According to this study[30], the relative binding free energy was computed by GENESIS 1.7.1[31] using the hybrid topology approach with the CHARMM36[32] force field. Briefly, C36 and C70 were selected as state-A molecules. On the other hand, compounds C28, C38, C45, C64, C73, C76, C80, C83, C89, and C114 were selected as state-B molecules. These compounds were randomly selected from the dataset based on their variable inhibitory activities. The hybrid ligand's structure, topology, parameters, and input files were generated using CHARMM-GUI[33]. The maximum common substructure (MCS) was applied for overlapping ligands to determine the minimal perturbated atoms between the paired ligands. If such a state-A to state-B mutation is not feasible for a certain ligand, the CGenFFv1.x algorithm discards it automatically and is not considered further. For the simulation setup, two end-state systems were generated for each paired ligand, i.e., the ligand in the solvent and the ligand in the complex. The systems were neutralized and ionized with 0.15 M NaCl counterion. Thereafter, minimization, NVT, and NPT simulations were performed to remove the bad contacts, gradually increasing the temperature from 0.1 K to 300 K and pressure to 1 bar with applying the restraint. Following that, a long 10 ns second NPT simulation was performed without position restraint. Thereafter, the λ-exchange FEP simulations were performed. Twelve λ windows were used to sequentially transform the interactions from state-A to state-B with the surroundings, in which six coupling parameters were used. Finally, the free energy differences were estimated using the Bennett acceptance ratio (BAR) method. The relative binding free energy ($\Delta\Delta G^{A\rightarrow B}_{RBFE}$) between the paired ligands was calculated as the following:

$$\Delta\Delta G^{A\rightarrow B}_{RBFE} = \Delta G^{A\rightarrow B}_{COM} - \Delta G^{A\rightarrow B}_{LIG} \quad (2)$$

doi:10.20944/preprints202211.0356.v1

where $\Delta G_{\mathrm{COM}}^{\mathrm{A}\rightarrow\mathrm{B}}$ and $\Delta G_{\mathrm{LIG}}^{\mathrm{A}\rightarrow\mathrm{B}}$ represent the free energy changes upon the transformation of state-A to state-B in the complex and isolated in solution, respectively.

3. Results and Discussion

3.1. MD simulation analysis and binding energy calculation

The protein-ligand RMSD curves for the 100 ns MD simulation are shown in Figure 1a. Convergence was reached within the initial 5 ns interval, and thereafter both the ligand and the protein maintained a stable plateau at the end of the production run. In the original crystal structure, C36 was stabilized by forming two interatomic H-bonds (Hb-1 and Hb-2) with the keto and amide groups of C502. The H-bond distances were measured through production simulation and were found to be between 2.7-3.5 Å, validating the overall stability of the ligand. Next, we calculated the ligand binding affinity using MM-PB/GBSA end-state binding free energy calculation. The different binding energy (BE) terms are shown in Figure 1b and Table S1. The van der Waals (VDW) and electrostatic (E_{EL}) terms each provided favorable ligand binding energy of -58.85 and -16.96 kcal/mol. The polar (E_{GB}) and non-polar (E_{SURF}) solvation terms are obtained as 29.54 and -6.49 kcal/mol. The ΔTOTAL and interaction entropy (-TΔS) were obtained as -52.76 and 7.51 kcal/mol, respectively. The final binding energy (ΔG_{bind}) was estimated to be -45.25 kcal/mol by deducting the entropy term from ΔTOTAL. Accurate binding energy contributions from active site residues are crucial for structure-guided inhibitor optimization process. In our study, we identified that I428, V436, V884, M499, L501, C502, G505, L553, G563, D564, and L567 were present within the boundary of 4 Å of the ligand atoms and contributed the critical binding affinity to ligand (Table S2). This information was further co-utilized in the 3D-QSAR study.

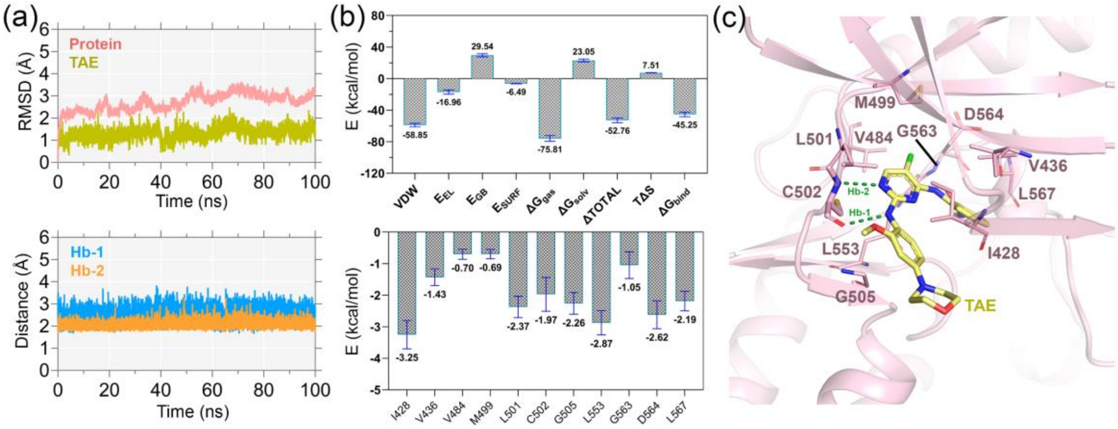

Figure 1. MD simulation and MM-PB/GBSA binding energy calculation. (a) RMSD analysis of the protein backbone and TAE during 100 ns of MD simulation. The distances of the two intermolecular H-bonds (Hb-1 and Hb-2) with the keto and amide groups of C502 are shown during the MD run. (b) Binding affinity calculation and residue-specific binding energy decomposition from the MM-PB/GBSA calculation. (c) Residues within 4 Å distance within the TAE ligand atoms, that contribute critical binding energy to the ligand are shown in the stick representation.

3.2. Statistical analysis of 3D-QSAR models

The receptor-based CoMFA and CoMSIA, two well-known 3D-QSAR models were developed using 125 compounds. Compound C107 has non-specific bio-activity and was discarded from the dataset during model building. The 2D structures and corresponding pIC_{50} values of these compounds are listed in Table S3. Molecular alignment of the compounds was done by superimposing the dataset compounds over the common core of the average MD position of C36. The 3-D alignment of the compounds over C36 inside the binding pocket is shown in Figure 2a. To develop a well-predictive model as well as the model's predictive ability, we split the dataset into a training set and test set compounds by following a 3:1 ratio by employing random sampling methods according to our previous studies[18, 23]. Briefly, the compounds were arranged into three mutually exclusive non-overlapping groups i.e., high, medium, and low activity groups based on their pIC_{50} values. Following that, a random draw was performed from each group in such a way, so that the compounds had an equal chance to be selected in the test set compounds. Using this method, four different training and test sets were developed for the CoMFA study (SET-A to D), as shown in Table S4.

Statistical analyses of the CoMFA models are presented in Table 1. During the model evaluation, we strictly followed the acceptance criterion for each parameter, specified in the 'Threshold value column'. The q^2 and r^2 values for SET-A were 0.593 and 0.839, respectively, at ONC of 5. For SET-B, the q^2 and r^2 values were 0.541 and 0.666 at ONC of 2. The q^2 and r^2 values of SET-C were 0.505 and 0.612 at ONC 2, while SET-D had q^2 and r^2 values of 0.633 and 0.897 at ONC of 6. Higher q^2 and r^2 values in combination with low χ^2 and RMSE values were considered for the internal validation of the proposed model employing the training set compounds. SET-D had the highest q^2 and r^2 with satisfactory χ^2 and RMSE values of 0.325 and 0.356, respectively, which were below the threshold constraint, and was selected as the final CoMFA model among the other datasets. In addition to the above parameters, k or k', r_0^2 or r'^2_0 , $|r_0^2 - r'^2_0|$, r_m^2 or r'^2_m were also computed for internal validation and were found to be in good agreement with the threshold parameters. However, QSAR models are unpredictable without external validation using test set compounds that are not included in the training set during model development. Similar to the internal validation, k or k', r_0^2 or r'^2_0, $|r_0^2 - r'^2_0|$, r_m^2 or r'^2_m parameters were considered to assess the external validation of the model. However, the final selection was done by evaluating the predictive correlation coefficient or r^2_{pred}. Overall, SET-D showed the highest r^2_{pred} value ($r^2_{pred} = 0.911, > 0.6$) and was therefore considered as the final CoMFA model.

We employed the CoMSIA evaluation of SET-D since CoMSIA employed a more comprehensive set of descriptor fields, such as hydrophobic (H), H-bond acceptor (A), and H-bond donor (D), in addition to the steric (S) and electrostatic (E) fields of CoMFA in different permutation-combination processes (Table S5). The highest q^2 and r^2 values were 0.656 and 0.862 at an ONC of 6. The other parameters, such as χ^2 and RMSE, r_m^2 or

r'^2_m followed the well-accepted statistical norms indicating good internal validation. In addition, we obtained an r^2_{pred} of 0.843, indicating excellent predictivity of the CoMSIA model. The actual and predicted pIC_{50} values with the residuals are listed in Table S6, and the PLS correlation plots from CoMFA and CoMSIA are shown in Figures 2b and 2c, respectively.

Overall, SET-D provided statistically significant CoMFA and CoMSIA models with strong internal and external validations, suggesting that both models can predict the inhibitory potential of unknown chemicals with a similar scaffold. Next, we performed the applicability domain (AD) analyses using data obtained from the 3D-QSAR study. Unlike other ML-based methods, 3D-QSAR uses the least squares algorithm to correlate the chemical descriptors and their inhibitory activity thus, QSAR applications are limited but highly efficient for compounds with congeneric series of compounds. The applicability domain is a distance-based graphical prediction method, that determines the uncertainty in the predictability of compounds based on structural similarity. The AD analysis of CoMFA and CoMSIA using the Williams plot is depicted in Figures 2d and 2e in a square area of σ = ±3, in which the standardized residuals of the training and test set compounds are plotted against the leverage values. None of the compounds fell outside the warning leverage (*h**), indicating the reliability and robustness of both 3D-QSAR models.

Table 1. Statistics of CoMFA and CoMSIA models

Statistical parameters	CoMFA				CoMSIA (SED) SET-D	Threshold values	Statistical parameters	CoMFA				CoMSIA SET-D	Threshold values
	SET-A	SET-B	SET-C	SET-D				SET-A	SET-B	SET-C	SET-D		
q^2	0.593	0.541	0.505	0.633	0.656	> 0.5	k_{Test}	0.994	0.979	1.009	1.007	1.011	0.85 ≤ k or k’ ≤ 1.15
ONC	5	2	2	6	6		k'_{Test}	1.002	1.015	0.985	0.991	0.985	
SEP	0.559	0.554	0.612	0.521	0.510		r^2_{Test}	0.578	0.422	0.767	0.922	0.850	
r^2	0.839	0.666	0.643	0.897	0.862	> 0.6	r^2_0 Test	0.494	0.377	0.735	0.915	0.854	≈ r^2
SEE	0.352	0.473	0.277	0.277	0.323	<< 1	r'^2_0 Test	0.540	0.240	0.417	0.886	0.816	
F-value	91.487	90.592	81.911	125.822	89.719		$\lvert r^2_0 - r'^2_0 \rvert$ Test	0.046	0.137	0.317	0.028	0.037	< 0.3
BS- r^2	0.895	0.712	0.699	0.934	0.940		$\frac{r^2-r^2_0}{r^2}$ Test	0.144	0.104	0.317	0.007	-0.003	< 0.1
BS-SD	0.025	0.051	0.050	0.017	0.016		$\frac{r^2-r'^2_0}{r^2}$ Test	0.064	0.430	0.041	0.038	0.039	
χ^2	0.285	0.537	0.507	0.387	0.325	< 1.0	r^2_m Test	0.410	0.333	0.630	0.846	N/A	r^2_m or r'^2_m > 0.5
RMSE	0.333	0.437	0.430	0.382	0.356	< 0.5	r'^2_m Test	0.46	0.24	0.31	0.	0.694	

MAE	0.001	0.001	0.001	0.001	0.001	≈ 0
RSS	14.275	24.554	23.748	15.253	16.28	
k $_{Train}$	0.996	1.003	0.998	0.991	0.997	0.85 ≤ k or k’ ≤ 1.15
k’ $_{Train}$	1.000	0.991	0.996	1.005	0.999	
r_0^2 Train	0.814	0.665	0.597	0.667	0.718	≈ r^2
r'^2_0 Train	0.785	0.396	0.467	0.635	0.662	
$\lvert r_0^2 - r'^2_0 \rvert$ Train	0.028	0.269	0.129	0.041	0.055	< 0.3
$\frac{r^2-r_0^2}{r^2}$ Train	0.029	2.53*10^-5	0.071	0.245	0.167	< 0.1
$\frac{r^2-r'^2_0}{r^2}$ Train	0.063	0.404	0.273	0.291	0.231	
r_m^2 Train	0.706	0.663	0.505	0.476	0.534	r_m^2 or r'^2_m > 0.5
r'^2_m Train	0.644	0.320	0.373	0.438	0.477	

	6	2	3	748		
r^2_{pred}	0.495	0.361	0.724	0.911	0.843	> 0.6
Q^2_{F1}	0.495	0.361	0.724	0.911	0.843	
Q^2_{F2}	0.493	0.353	0.723	0.910	0.842	
Q^2_{F3}	0.495	0.361	0.724	0.911	0.843	
Q^2_{ccc}	0.759	0.655	0.811	0.950	0.916	
S (%)	47.1	47.0	46.9	39.4	18.7	
E (%)	52.9	53.0	53.1	60.6	46.1	
H (%)						
A (%)						
D (%)					35.2	

q^2: squared cross-validated correlation coefficient; ONC: optimal number of components; SEP: standard error of prediction; r^2: squared correlation coefficient; SEE: standard error of estimation; F-value: F-test value; BS-r^2: Bootstrapping squared correlation coefficient; χ^2: chi-square value; RMSE: Root Mean Square Error; MAE: mean absolute error; k: slope of the predicted vs. observed activity at zero intercepts; k’: slope of the observed vs. predicted activity at zero intercepts; r_0^2 : squared correlation coefficient between predicted and observed activity; r'^2_0 : squared correlation coefficient between predicted and observed activity; r_m^2 or r'^2_m : r_m^2 , r'^2_m matrix; r^2_{pred} : predictive correlation coefficient; Q^2_{F1}, Q^2_{F2}, Q^2_{F3}, and Q^2_{ccc}: Q^2_{F1}, Q^2_{F2}, Q^2_{F3} and Q^2_{ccc} matrices. S: steric; E: electrostatic; H: Hydrophobic; A: H-bond acceptor; D: H-bond donor.

doi:10.20944/preprints202211.0356.v1

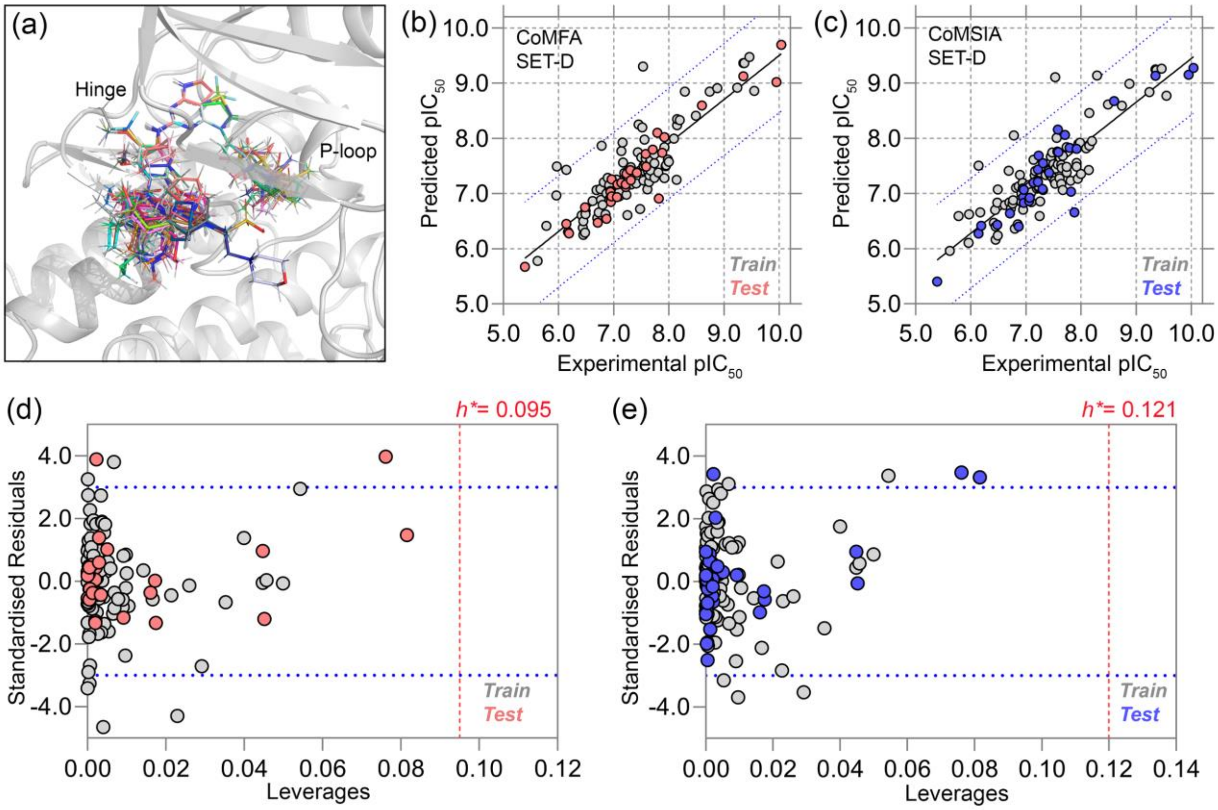

Figure 2. Molecular alignment of the dataset compounds, PLS plots, and applicability domain (AD) analysis. (a) Molecular alignment of the dataset compounds with the common chemical core of C36 (TAE) inside the FAK binding cavity. (b) PLS correlation plots of CoMFA (SET-D) study. (c) PLS correlation plots of the CoMSIA (SET-D) study. (d) and (e) Applicability domain analysis using the distance-based Williams plot using the data obtained from the CoMFA and CoMSIA models. The *h** with dotted lines in red signifies the warning leverage values in both plots.

3.3. Contour map analysis

Following statistical validation, descriptive colored contour maps around the MD structure of C36 were generated from the 3D-QSAR study. The compounds were well aligned on the common core of the *N*-phenylpyrimidine-2-amine moiety inside the ATP pocket (Figure 3a). In the CoMFA analysis, the green and blue contours represent a favorable position for steric and electropositive substitutions, whereas the yellow and red contours did not favor those substitutions (Figures 3b and 3c)[34, 35]. In the steric contour map, a green contour was observed near the R_1 position of the anisole ring and two green contours appeared around the R_2 position of the morpholine ring, indicating that the steric substitution would be preferable for these regions. A yellow contour at the R_3 position near residues D564, V436, and L567 indicates an unfavorable position for bulky steric substitution. Consequently, residue D564 is the part of the DFG motif that contributes -2.62 kcal/mol to ligand binding; thus, a bulky substitution at that position could have the steric hindrance effect and may lead to a decrease in overall binding affinity. Compounds C71, C72, C77, C79, C81, C82, and C84 had steric moieties adjacent to the green contours and exhibited inhibitory (pIC_{50}) more than 9. In the electrostatic contour map, a blue contour near *N*-methylbenzamide and two small red contours near the morpholine ring indicated that positively charged groups would be favorable and unfavorable in that chemical space. Very similar steric and electrostatic contours appeared (Figures 3d and 3e) during the CoMSIA study, although an additional blue contour was present in the *ortho*- position of the six-membered rings at R_2, overall corroborating the CoMFA contours. In the CoMSIA H-bond donor contour, two purple and two cyan contours appeared near R_2 and R_3, indicating the favorable and unfavorable substitutions for the H-bond donor groups, which can increase the overall inhibitory potential of C36. Figure 3f shows an SAR diagram based on the information obtained from the 3D-QSAR analysis. Residues D564, V436, and L567 were proximal (< 4 Å) to the R_3 position of *N*-methylbenzamide, and the critical binding energy decomposed to C36. Furthermore, SAR analysis revealed that non-steric, H-bond donor, and electropositive chemical groups could be advantageous substitutions at R_3 in terms of improving inhibitory effects. Therefore, this chemical space of C36 could serve as a potential site for chemical modification to ameliorate the FAK binding affinity.

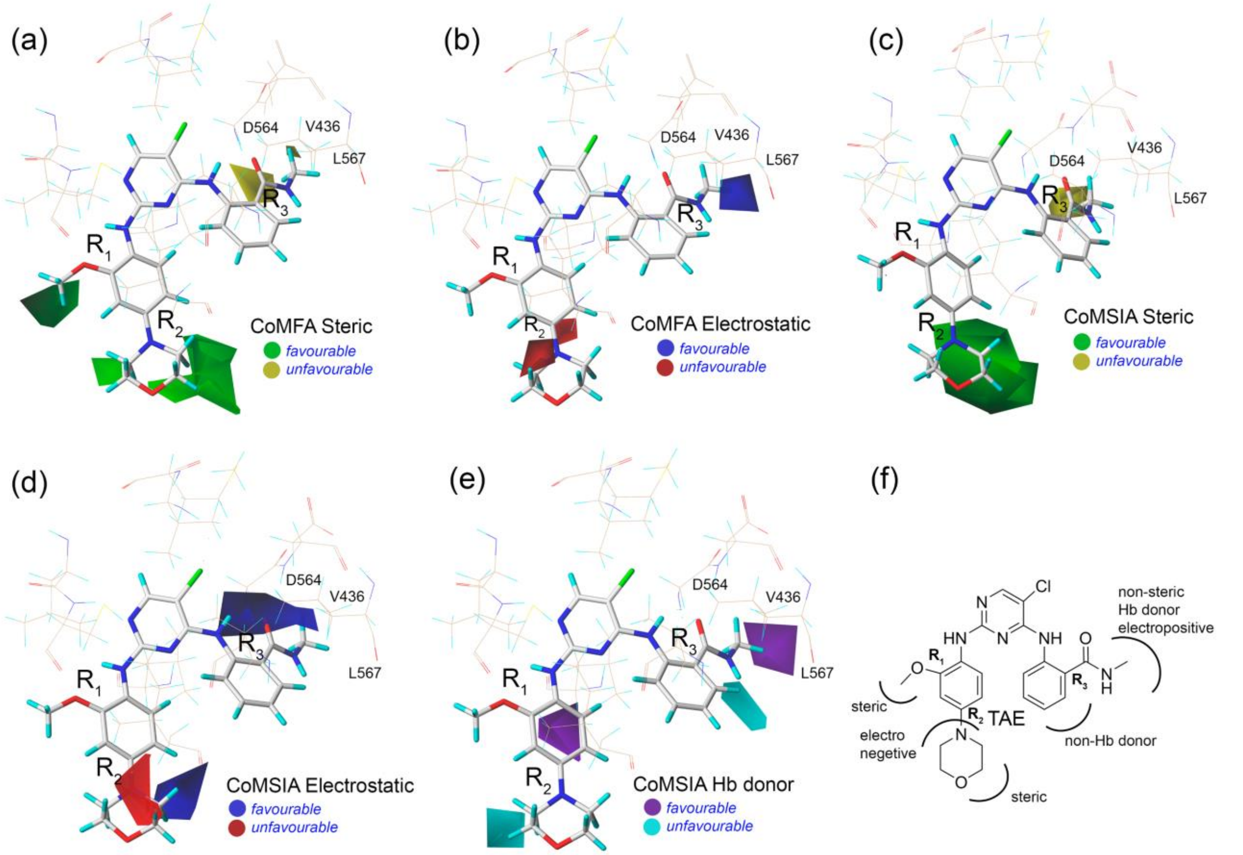

Figure 3. Contour map analysis and structure activity relationship study from 3D-QSAR. (a) Steric contour map and (b) Electrostatic contour map from CoMFA. (c), (d), and (e) are Steric, Electrostatic, and H-bond (Hb) donor contour map from CoMSIA. (f) Implementation of the SAR diagram from CoMFA and CoMSIA analysis by taking TAE (C36) as a reference.

3.4. Relative binding affinity estimation

For relative binding estimation study, we randomly selected 12 compounds from the dataset by varying the degree of inhibitory activity. The experimental binding energy (ΔG_{EXP}) values were deduced from the inhibitory activities of the selected compounds. The partial charges and LJ parameters gradually changed during the alchemical transformation of the ligand from state-A to state-B within the binding pocket in the FEP simulation. These changes were made by implementing a hybrid topology from 0 to 1 in twelve different λ intermediate steps. Figure 4a shows the generalized thermodynamic cycle of the relative binding free energy derivation scheme. In the earlier studies [36, 37], we used an absolute binding free energy estimate in the modeling study of kinase inhibitors and found a satisfactory correlation between the experimental and computed binding free energies, despite the high numerical approximation. Since the entire ligand needs to be perturbed (interactions off or on) corresponding to its surroundings, which requires a large number of λ intermediate states and simulation time. In contrast, only a fraction of the chemical moiety is required to be perturbed to transition from state-A to state-B in fewer λ states. Compounds C36 and C70 were selected as state-A, while compounds C28, C38, C45, C64, C73, C76, C80, C83, C89 and C114 were assigned as state-B. The common and mismatched atoms are shown in black and red in Figure 4b, respectively. A hybrid molecule was generated by superimposing the chemical structures of two ligands. In this hybrid molecule, the common part was assigned as a single topology or the same topology as the first ligand. The remaining hybrid molecules were assigned a single-dual hybrid topology. During the FEP simulation, the dual topology portion was changed (including the LJ parameters, partial charges, and bonds) using the forcefield by 12 alter λ scaling (Table S7). Each alter-λ simulation was run for 1 ns in triplicate to ensure sufficient sampling while overlapping the neighboring windows. In this manner, a total of 72 ns simulation for a single alchemical transformation in complex and isolated forms was performed.

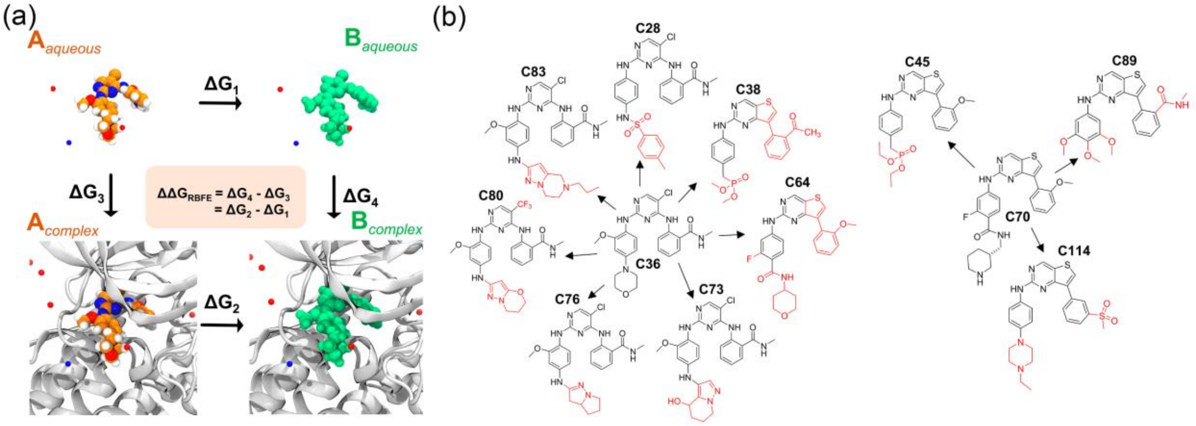

Figure 4. Overview of the FEP scheme and relative binding affinity estimation. (a) Thermodynamic pathway of ligand transformation from State-A to State-B in aqueous and in complex form. The $\Delta\Delta G_{RBFE}^{A\rightarrow B}$ can be deduced from the free energy changes of both states in aqueous and complex systems. (b) Relative binding energy calculation of the ligands through alchemical transformation. The mismatched atoms between the ligand pairs, which need to be perturbated, are shown in red.

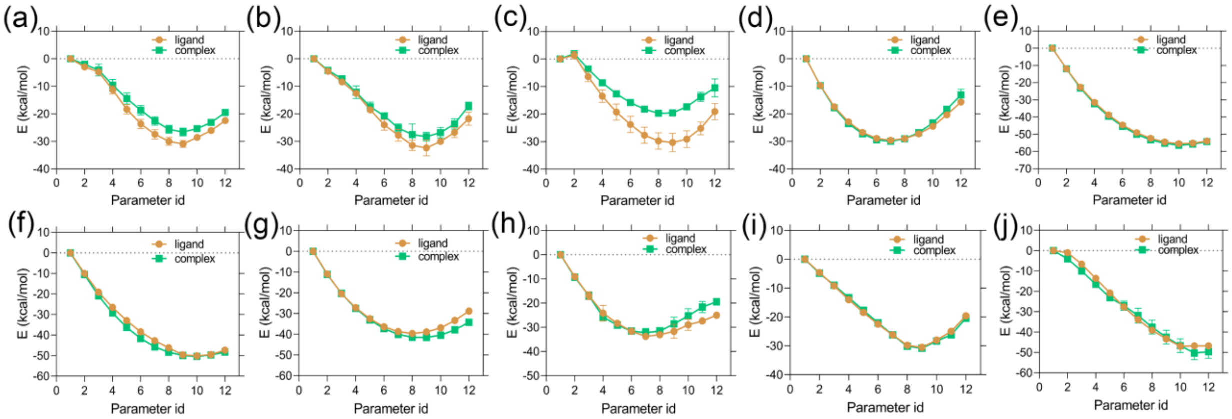

Figure 5. FEP energy convergence plots of (a) C36 → C28, (b) C36 → C38, (c) C36 → C64, (d) C36 → C73, (e) C36 → C76, (f) C36 → C80, (g) C36 → C83, (h) C70 → C45, (i) C36 → C89, (j) C36 → C114 in complex and isolated form.

The results of the alchemical transformation by the FEP methods are shown in Figure 5, as the free energy changes from state-A to state-B through the different λ states in complex and isolated forms. BAR methods were utilized to calculate the energy differences between the neighboring λ windows. In each graph, the total energy difference between the initial (λ=0) and final (λ=1) stages of the ligands in the complex and isolated forms correspond to $\Delta G_{COM}^{A\rightarrow B}$ and $\Delta G_{LIG}^{A\rightarrow B}$, respectively. From equation (2) we derived the $\Delta\Delta G_{RBFE}^{A\rightarrow B}$ from each ligand transformation, which is summarized in Table 2. The computed $\Delta\Delta G_{RBFE}^{A\rightarrow B}$ of C36 → C28, C36 → C38, C36 → C73, C36 → C76, C36 → C83, C36 → C89, and C36 → C114 were found to be 2.94, 4.58, 2.64, -0.23, -0.97, -0.79

and -2.86 kcal/mol with corresponding to their theoretical $\Delta\Delta G_{\mathrm{EXP}}^{\mathrm{A}\rightarrow\mathrm{B}}$ of 1.44, 3.00, 0.93, -1.43, -0.53, -0.60, and -1.29 kcal/mol, respectively, which is a good agreement between experimental and computed relative binding affinity. However, the transformation of C36 → C64, C70 → C45, and C70 → C114 yielded a higher $\Delta\Delta G_{\mathrm{RBFE}}^{\mathrm{A}\rightarrow\mathrm{B}}$ approximation than the $\Delta\Delta G_{\mathrm{EXP}}^{\mathrm{A}\rightarrow\mathrm{B}}$ values. In this case, we anticipated that increasing the number of iterations and λ sampling would reduce the mean statistical approximation. We determined Pearson's correlation coefficient using the computed values and their respective experimental values in Figure S1. A Pearson's R (R_{RBFE}) was obtained as 0.82 and an R^2 of 0.68, indicating the reasonable performance of the physics-based binding affinity calculation. In addition, the correlation statistics can be expressed in a linear equation form:

$$\Delta\Delta G_{\mathrm{EXP}}^{\mathrm{A}\rightarrow\mathrm{B}} = 0.3345\times \Delta\Delta G_{\mathrm{RBFE}}^{\mathrm{A}\rightarrow\mathrm{B}} - 0.4229 \quad (3)$$

The above equation could be useful for FEP-based SAR investigation of TAE/C36 analogs as well as the prediction of $\Delta\Delta G_{\mathrm{EXP}}^{\mathrm{A}\rightarrow\mathrm{B}}$ values with reasonable accuracy.

Table 2. Energy terms of Alchemical binding energy transformation from state-A to state-B

state-A (ΔG_{EXP})	state-B (ΔG_{EXP})	$\Delta\Delta G_{\mathrm{EXP}}^{\mathrm{A}\rightarrow\mathrm{B}}$	ΔG_{COM} ±SD	ΔG_{LIG} ±SD	λ-dependent $\Delta\Delta G_{\mathrm{RBFE}}^{\mathrm{A}\rightarrow\mathrm{B}}$
C36 (-11.33)	C28 (-9.89)	1.44	-19.51 ± 0.87	-22.45 ± 0.99	2.94
	C38 (-8.33)	3.00	-17.12 ± 2.35	-21.77 ± 1.41	4.58
	C64 (-9.67)	1.66	-19.11 ±3.26	-10.48 ±2.95	-8.63
	C73 (-10.40)	0.93	-13.09 ± 0.64	-15.73 ± 2.06	2.64
	C76 (-12.76)	-1.43	-54.20 ± 0.31	-53.97 ± 0.77	-0.23
	C80 (-14.03)	-2.70	-34.22 ± 0.13	-28.88 ± 1.16	-5.34
	C83 (-11.86)	-0.53	-48.29 ± 0.63	-47.34 ± 0.58	-0.97
C70 (-9.53)	C45 (-9.50)	0.03	-19.44 ± 1.02	-25.01 ± 0.26	5.57
	C89 (-10.13)	-0.60	-20.41 ± 0.32	-19.62 ± 0.60	-0.79
	C114 (-10.82)	-1.29	-49.62 ± 1.06	-46.76 ± 3.29	-2.86

ΔG_{EXP}: experimental binding free energy; $\Delta\Delta G_{\mathrm{EXP}}^{\mathrm{A}\rightarrow\mathrm{B}}$: experimental relative binding free energy; ΔG_{COM}: free energy changes in complex; ΔG_{LIG}: free energy changes isolated form; $\Delta\Delta G_{\mathrm{RBFE}}^{\mathrm{A}\rightarrow\mathrm{B}}$: computed relative binding free energy.

Conclusion

In this study, we employed ML and physics-based hybrid modeling approach to study the structure-activity relationship and binding mechanism of *N*-phenylpyrimidine-2-amine based FAK inhibitors. As FAK is one of the most important regulators of growth factor receptor signaling, its overexpression and concomitant drug resistance pose a major challenge to chemists. From the molecular simulation, H-bond analysis and MM-PBSA binding energy calculations were employed to assess the ligand stability, binding affinity, and per-residue binding energy decomposition of the crystal ligand. Residues such as I428, V436, V484, M499, L501, C502, G505, L553, G563, D564, and L567 were identified as important BE contributing residues to the ligand binding. Following that, the statistically reasonable CoMFA and CoMSIA models were developed and both showed excellent predictive capability. Descriptive colored contour maps surrounding compound C36 illustrated that chemical substitutions along these contours would more likely increase the inhibitory activity. This information can be further co-utilized with the residue-specific binding energy profile to aid in molecular probing and ligand design. Finally, we applied the alchemical FEP simulation by taking 12 different ligands to estimate their relative binding affinity. An acceptable agreement was obtained between the experimental relative binding energies and the computed relative binding energies. The molecular modeling techniques employed here in different combinations could be useful for further lead optimization in medicinal chemistry research.

Data availability

Data available within the article or its supplementary information.

Competing interests

The authors have no competing financial interests to declare.

Acknowledgments

This study was supported by research funds from the Chosun University 2022.

References

1. Liao, Y., et al., *ATX/LPA axis regulates FAK activation, cell proliferation, apoptosis, and motility in human pancreatic cancer cells.* In Vitro Cellular & Developmental Biology-Animal, 2022. **58**(4): p. 307-315.
2. Pomella, S., et al., *New Insights on the Nuclear Functions and Targeting of FAK in Cancer.* International Journal of Molecular Sciences, 2022. **23**(4): p. 1998.
3. Le Coq, J., et al., *New insights into FAK structure and function in focal adhesions.* Journal of Cell Science, 2022. **135**(20): p. jcs259089.
4. Zhai, C., et al., *Activation of autophagy induces monocrotaline-induced pulmonary arterial hypertension by FOXM1-mediated FAK phosphorylation.* Lung, 2022. **200**(5): p. 619-631.
5. Yang, J.Y., et al., *Induction of Apoptosis and Effect on the FAK/AKT/mTOR Signal Pathway by Evodiamine in Gastric Cancer Cells.* Current Issues in Molecular Biology, 2022. **44**(9): p. 4339-4349.

6. Spallarossa, A., et al., *The Development of FAK Inhibitors: A Five-Year Update.* International Journal of Molecular Sciences, 2022. **23**(12): p. 6381.
7. Wankowicz, S.A., et al., *Ligand binding remodels protein side-chain conformational heterogeneity.* Elife, 2022. **11**: p. e74114.
8. Singh, V.K., et al., *Docking, ADMET prediction, DFT analysis, synthesis, cytotoxicity, antibacterial screening and QSAR analysis of diarylpyrimidine derivatives.* Journal of Molecular Structure, 2022. **1247**: p. 131400.
9. Castelli, M., et al., *New perspectives in cancer drug development: computational advances with an eye to design.* RSC Medicinal Chemistry, 2021. **12**(9): p. 1491-1502.
10. Wang, R., et al., *Design, synthesis, biological evaluation and molecular modeling of novel 1H-pyrrolo[2,3-b]pyridine derivatives as potential anti-tumor agents.* Bioorganic Chemistry, 2020. **94**: p. 103474.
11. Wang, R., et al., *Discovery of 7H-pyrrolo[2,3-d]pyridine derivatives as potent FAK inhibitors: Design, synthesis, biological evaluation and molecular docking study.* Bioorganic Chemistry, 2020. **102**: p. 104092.
12. Qu, M., et al., *Design, synthesis and biological evaluation of sulfonamide-substituted diphenylpyrimidine derivatives (Sul-DPPYs) as potent focal adhesion kinase (FAK) inhibitors with antitumor activity.* Bioorganic & Medicinal Chemistry, 2017. **25**(15): p. 3989-3996.
13. Wang, R., et al., *Design, synthesis, biological evaluation and molecular docking study of novel thieno[3,2-d]pyrimidine derivatives as potent FAK inhibitors.* European Journal of Medicinal Chemistry, 2020. **188**: p. 112024.
14. Xie, H., et al., *Design, synthesis and biological evaluation of ring-fused pyrazoloamino pyridine/pyrimidine derivatives as potential FAK inhibitors.* Bioorganic & Medicinal Chemistry Letters, 2020. **30**(21): p. 127459.
15. Lietha, D. and M.J. Eck, *Crystal structures of the FAK kinase in complex with TAE226 and related bis-anilino pyrimidine inhibitors reveal a helical DFG conformation.* PloS one, 2008. **3**(11): p. e3800.
16. Zhou, J., et al., *Allosteric regulation of focal adhesion kinase by PIP2 and ATP.* Biophysical journal, 2015. **108**(3): p. 698-705.
17. Ghosh, S., S. Keretsu, and S.J. Cho, *Designing of the N-ethyl-4-(pyridin-4-yl)benzamide based potent ROCK1 inhibitors using docking, molecular dynamics, and 3D-QSAR.* PeerJ, 2021. **9**.
18. Ghosh, S. and S.J. Cho, *Structural Insights from Molecular Modeling of Isoindolin-1-One Derivatives as PI3Kγ Inhibitors against Gastric Carcinoma.* Biomedicines, 2022. **10**(4): p. 813.
19. Abraham, M.J., et al., *GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers.* SoftwareX, 2015. **1**: p. 19-25.
20. Ghosh, S., S. Keretsu, and S.J. Cho, *Computational Modeling of Novel Phosphoinositol-3-kinase γ Inhibitors Using Molecular Docking, Molecular Dynamics, and 3D-QSAR.* Bulletin of the Korean Chemical Society, 2021. **42**(8).
21. Ghosh, S., S. Keretsu, and S.J. Cho, *Molecular Modeling Studies of N-phenylpyrimidine-4-amine Derivatives for Inhibiting FMS-like Tyrosine Kinase-3.* International Journal of Molecular Sciences, 2021. **22**(22): p. 12511.
22. Da Silva, A.W.S. and W.F. Vranken, *ACPYPE-Antechamber python parser interface.* BMC research notes, 2012. **5**(1): p. 1-8.
23. Keretsu, S., S. Ghosh, and S.J. Cho, *Computer aided designing of novel pyrrolopyridine derivatives as JAK1 inhibitors.* Scientific reports, 2021. **11**(1): p. 1-12.

doi:10.20944/preprints202211.0356.v1

24. Valdés-Tresanco, M.S., et al., *gmx_MMPBSA: A New Tool to Perform End-State Free Energy Calculations with GROMACS.* Journal of Chemical Theory and Computation, 2021. **17**(10): p. 6281-6291.
25. Ghosh, S. and S.J. Cho, *Comparative binding affinity analysis of dual CDK2/FLT3 inhibitors.* Bulletin of the Korean Chemical Society.
26. Ghosh, S., S. Keretsu, and S.J. Cho, *3D-QSAR, Docking and Molecular Dynamics Simulation Study of C-Glycosylflavones as GSK-3β Inhibitors.* Journal of the Chosun Natural Science, 2020. **13**(4): p. 170-180.
27. Todeschini, R., D. Ballabio, and F. Grisoni, *Beware of unreliable Q 2! A comparative study of regression metrics for predictivity assessment of QSAR models.* Journal of Chemical Information and Modeling, 2016. **56**(10): p. 1905-1913.
28. Veerasamy, R., et al., *Validation of QSAR models-strategies and importance.* Int. J. Drug Des. Discov, 2011. **3**: p. 511-519.
29. Abdizadeh, R., F. Hadizadeh, and T. Abdizadeh, *QSAR analysis of coumarin-based benzamides as histone deacetylase inhibitors using CoMFA, CoMSIA and HQSAR methods.* Journal of Molecular Structure, 2020. **1199**: p. 126961.
30. Cournia, Z., B. Allen, and W. Sherman, *Relative binding free energy calculations in drug discovery: recent advances and practical considerations.* Journal of chemical information and modeling, 2017. **57**(12): p. 2911-2937.
31. Jung, J., et al., *GENESIS: a hybrid-parallel and multi-scale molecular dynamics simulator with enhanced sampling algorithms for biomolecular and cellular simulations.* Wiley Interdisciplinary Reviews: Computational Molecular Science, 2015. **5**(4): p. 310-323.
32. Huang, J., et al., *CHARMM36: An improved force field for folded and intrinsically disordered proteins.* Biophysical Journal, 2017. **112**(3): p. 175a-176a.
33. Kim, S., et al., *CHARMM-GUI free energy calculator for absolute and relative ligand solvation and binding free energy simulations.* Journal of chemical theory and computation, 2020. **16**(11): p. 7207-7218.
34. Bang, S.J. and S.J. Cho, *Comparative molecular field analysis (CoMFA) and comparative molecular similarity index analysis (CoMSIA) study of mutagen X.* BULLETIN-KOREAN CHEMICAL SOCIETY, 2004. **25**(10): p. 1525-1530.
35. San Juan, A.A. and S.J. Cho, *3D-QSAR study of microsomal prostaglandin E 2 synthase (mPGES-1) inhibitors.* Journal of Molecular Modeling, 2007. **13**(5): p. 601-610.
36. Ghosh, S. and S.J. Cho, *Structure–activity relationship and in silico development of c-Met kinase inhibitors.* Bulletin of the Korean Chemical Society, 2022.
37. Ghosh, S. and S.J. Cho, *Binding Studies and Lead Generation of Pteridin-7 (8H)-one Derivatives Targeting FLT3.* International Journal of Molecular Sciences, 2022. **23**(14): p. 7696.