CHAPTER: 6

QSAR AND RETROSYNTHESIS OF GEDUNIN AND MODIFIED GEDUNIN

6.1 Introduction

In India, snakebites have become a public threat. In India, an average of more than 2,50,000 snakebites are recorded in just one year. Snakes in India have a great variety of species in terms of length and body weight. Snakes occupy deserts, forests, swampy places, lakes, streams, and rivers in rugged terrains **(Longbottom *et al.*, 2018).** Bites not only lead to death but are also responsible for severe paralysis, irreversible kidney failure, permanent amputation of limbs, and disability. It is vital to explore all possible antidote candidates and study their biochemical characteristics using QSAR (Quantitative structure-activity relationships) and cell viability assays. QSAR stands for the quantitative structure-activity relationship of the molecule and is an *in silico* technique developed to model the biological activity of molecules based on the chemical structure of a molecule by using statistics and mathematical knowledge to predict data concerning our desired molecule. A basic QSAR model has molecular descriptors, which are numerical representations of the chemical structure of the molecule and the biological activity predicted by that model. The biological properties of $_{IC50}$ and log P of the inhibitor molecule were considered for the prediction, where $_{IC50}$ is the half-maximum inhibitory concentration required for 50% inhibition of the enzyme in vitro, and log P is the measure of the Hydrophobicity that affects the metabolism affecting absorption, etc. The approach to creating QSAR models is based on machine learning methods using

the online chemical modelling environment (OCHEM) platform. The OCHEM platform has mainly two subsystems: a modelling framework to carry out the workflow and a database of experimentally measured properties to search through the data **(Kovalishyn *et al.*, 2018).** The model can be either linear or non-linear. The model used was based on machine learning methods in which a deep neural network (DNN) with a descriptor set CDK2 - Java-based descriptor calculation tool, which calculates the geometric, topological, charge-based, and constitutional descriptors, was used to calculate complex nonlinear relationships. The random forest method for predicting Log P using star drop software is a collection of predictors in the form of a tree, and each tree value depends on the independently sampled random vector with the same distribution for trees **(Chan *et al.,* 2016).** Machine learning methods are better for QSAR predictions than conventional methods in terms of prediction accuracy, training duration, prediction time, sensitive parameters, and interpretability of the models. The DNN method has advantages over the random forest and other machine learning methods in terms of the multi-task and generative models and the DNN‟s ability to automatically generate new chemical features **(Chan *et al.*, 2016).**

Retrosynthesis is an analytical technique that produces beginning material, commonly dubbed "synthon," to deconstruct or divide a target organic molecule. A specific breakdown pattern produces fragments. This is called retro synthesis, a reversible chemistry synthesis method. In his book The Logic of Chemical Synthesis, **E. J. Corey** presented this notion. Some terms are included, such as Disconnect, which signifies atom bonds' disconnection. The moieties used in chemical synthesis are those reactants. Synthon ions or radicals following the breakup. Retrosynthesis is a technique to design organic synthesis by

disconnecting a target molecule into a precursor matter. Retrosynthetic analysis is a method of analytics based on the desired product rather than a synthetic form of organic chemistry. Smaller pieces of the target molecule are broken down. The actual synthesis can then be predicated on Retrosynthesis, which is exceptionally successful but requires a good grasp of chemical compounds, compound classes, chemical processes, and reaction conditions **(Guo *et al.,* 2020).**

The qsar and retrosynthesis of gedunin and modified gedunin involve QSAR-based log p and $_{IC50}$ prediction of gedunin and modified gedunin, whereas Retrosynthesis of modified gedunin has also been shown here using software such as stardrop and ochem platforms.

6.2 Experimental

6.2.1 QSAR

6.2.1.1. Data Sets

The training set of 75 molecules with 85% similarity to the test molecule was obtained from the OCHEM database for the log $_{IC50}$ model. The validation set of 10 molecules was excluded from the training set to validate the predicted model from log $_{IC50,}$ whereas, for the prediction from a log-P model, the training set of 66 molecules with 80% structural similarity was provided by the Star-Drop software Modeller **(T'jollyn *et al.,* 2011)** and a set of 14 validation molecules together with a test set including inhibitor molecules.

6.2.1.2. Molecular Descriptors

The descriptors provide information on the physicochemical properties of the molecule. The theoretical molecular descriptors for the log $_{IC50}$ model were calculated using the OCHEM platform, considering log P, log S, E-state, and

CDK2.0. CDK 2.0 contains constitutional descriptors that provide information about the chemical composition of molecules, electrostatic descriptors such as dipole moment, polarisability, etc., predict the crystal density of the molecule, topological descriptors that predict the surface properties of molecules, such as the permeability of molecules and their solubility, biological activity, and orientation of the molecule in space is predicted by geometric descriptors. In contrast, hybrid descriptors form the prediction models **(Kratochvíl *et al.,* 2018).** MW (molecular weight) log P, log S, log $_{IC50,}$ and TPSA (topological surface area) were considered for predicting the log P model using the Stardrop Modeller.

6.2.1.3. Machine Learning Methods

To generate QSAR models, descriptors and machine learning algorithms such as random forest and deep neural network (DNN) were used. The parameters used were also optimized with Open Babel, and the validation of the generated model was a 5-fold cross-validation on the OCHEM platform. Training sets with DNN and random forest **(Deo *et al.,* 2015)** were used to generate the model. The DNN algorithm was used to predict the log $_{IC50}$ of a molecule with set CDK2 descriptors, as DNN has the advantage that it incrementally modifies the training data representation and thus leads to a super accuracy of the model **(Wainberg *et al*., 2018) (Tavanaei *et al*., 2019**). The random forest method was used to predict the log P of the inhibitor molecule using the Star-Drop software platform. This machine-learning method is based on an ensemble of decision trees. The training set had several features that were used to create a tree of molecules using randomization and stochastic recursive partitioning. First, randomizes the tree and selects a subset of features, then divides it into homogeneous groups so that decision tree nodes are a group of molecules with similar predictable properties

(Wainberg *et al.,* 2018). Descriptors are the parameters selected for each node and tree structure with the assignment of the last node. The second option was to assign a bootstrap value (0.37 or 37%) to the training data, which was used as the internal validation rate for the tree. RF (Random forest) is generally a robust method with an increased test set error rate in contrast to DNN **(Kriegeskorte N and Golan T, 2019).**

6.2.1.4.Validation and Cross-Validation

After generating a model using known data, it is crucial to validate the model to generalize unknown data. This is performed by dividing the entire dataset into a training set and an internal test set. The model was generated using a training set with known property values and tested using a test set with unknown property values, the value of which was predicted using the random forest or DNN algorithms **(Maestro *et al.,* 2018).** The DNN algorithm model generated on the OCHEM platform was also validated five times. Here, the data are layered in five separate folds so that fold 1 is a training set, and fold 2 is a validation set. Each convolution is a test set once, so the test set error is reduced. There is no randomization of properties and estimates to make the accuracy correct**(Deo *et al*., 2015) (Li *et al.,* 2017).**

6.2.2. Retrosynthesis

Retrosynthesis was performed using stardrop software **(Pahler *et al*., 2013)**, where the inhibitor molecule was the reagent, and the products were the initial raw compounds that could be used to synthesize the inhibitor. (a) Reacting the phorbol derivative with C2H2ClNO2S (cyanomethanesulfonyl chloride) at room temperature for three hours using solvent dichloromethane followed by water first

at pH2 then after at pH 8 to obtain an intermediate; contacting the intermediate of step (a) with C4H5NO (oxazine) in the presence of $B_2(pin)_2$, *t*$BuC_6H_4C(O)Cl$, $Pd_2(dba)_3$ (6 mol %), CuTC (16 mol %), $P(OPh)_3$ (21 mol %) thereof at room temperature to obtain the phorbol derivative with amine and oxygen groups, and (c) isolating inhibitor $C_{26}H_{31}N_2O_6F$.

6.3 Statistical analysis

Data collection and interpretation are statistical analyses used to detect patterns and trends. Statistical analysis may be employed in the circumstances such as obtaining research interpretations, statistical modelling, or conceiving surveys and studies. The obtained data were analyzed on the OCHEM platform **(Kovalishyn *et al.*, 2018)** for QSAR studies using DNN machine learning methods. The log P test was performed using the ICM tool (version 9). The MTT assay was performed in a laboratory using cell lines, electron microscopy, and glassware.

6.4 . Results and Discussion

Log P was predicted using the ICM tool for novel inhibitor molecule **(Table 6.1.,6.2.,6.3.)**, and the predicted log P of Inhibitor with model accuracy of 0.84 came to be 2.97 kow. **(Figure 6.1.,6.2.,6.3.).** To predict the IC_{50} of the novel inhibitor compound and gedunin, the OCHEM platform **(Table 6.4.)** was used, and the predicted log IC_{50} of the inhibitor came was 7.17 (-log M) and 0.70 (-log M) of gedunin **(Figure 6.4.).** For the validation study of QSAR, an MTT assay was performed using hepG2 cell lines, and the predicted IC_{50} of gedunin was 35.95 µg/ml in 24h **(Figure 6.5., 6.6.).** Retrosynthesis was performed using stardrop software to synthesize inhibitor molecules **(Figure 6.7.).** QSARPredicted

log P of Inhibitor with model accuracy of 0.84 is 2.97 kow whereas Predicted log IC_{50} of inhibitor is 7.17 (-log M) and 0.70 (-log M) of gedunin. The training set has great molecular diversity and is used to generate QSAR models and validate their accuracy. A validation set of molecules excluded from the training set was used. The test set was intended to estimate the prediction error in an unbiased manner for the model. Plotting the data together is a better model when the measured and predicted values are close. The regression for the test set was greater than that of the training set, owing to the non-identity of both datasets **(Kato *et al.,* 2020).** The correlation coefficient, that is, R2 in the case of the training set (called q2 in the case of the validation set), predictive and observed values take into account the model fitness in the dataset so that higher values of R2 are preferred; if it reaches 1, the model is perfect. **(Buhlmann and Reymond, 2020).** Mean squared error (MSE) measures the absolute model fitness, while RMSE gives the standard deviations of the model residuals as a measure of dispersion and considers the usefulness of models.

6.4.1. Good Model Criteria **(Verma *et al.,* 2010).**

6.4.1.1. Regressing on the predicted Test set **(Verma *et al*., 2010).**

- $R^2 > 0.6$ to ensure the fitness of the model
- R^2 passes through the origin
- regression slope close to 1

6.4.1.2. q^2 of the training set > 0.5, with cross-validation **(Bokulich *et al.,* 2018).**

6.4.1.3. RMSE of Test set < 10% **(Bokulich *et al*., 2018).**

6.4.2. Log P Model

Table 6.1. lists log P is measured, predicted, and experimental values developed using the random forest algorithm. The data included a training set of 66 organic molecules selected using Chembl's Star Drop Modeller. The training set was further subdivided into a validation set (14 molecules) and a test set, as shown in **Table 6.2. and 6.3.**, respectively, with validation molecules excluded from the training set to accurately model the generated set. The molecular descriptor was chosen for this model. The test kit shows our inhibitor molecule, which has a predicted log P-value of 2.97 Kow. The predicted model of log P accuracy depends on the square of the value of the correlation coefficient R2R2, which measures the chances that the model will replicate the result, or the ratio of the variation of a dependent variable to an independent variable with a value between 0 and 1. The R2 of the model is 0.84, as shown in **Figure. 6.1**., while 1 for the training rate and 0.82 for the model's validation rate, as shown in **Figure 6.2. and 6.3**., there is clear evidence that the regression model is best, as the value is closer to 1.

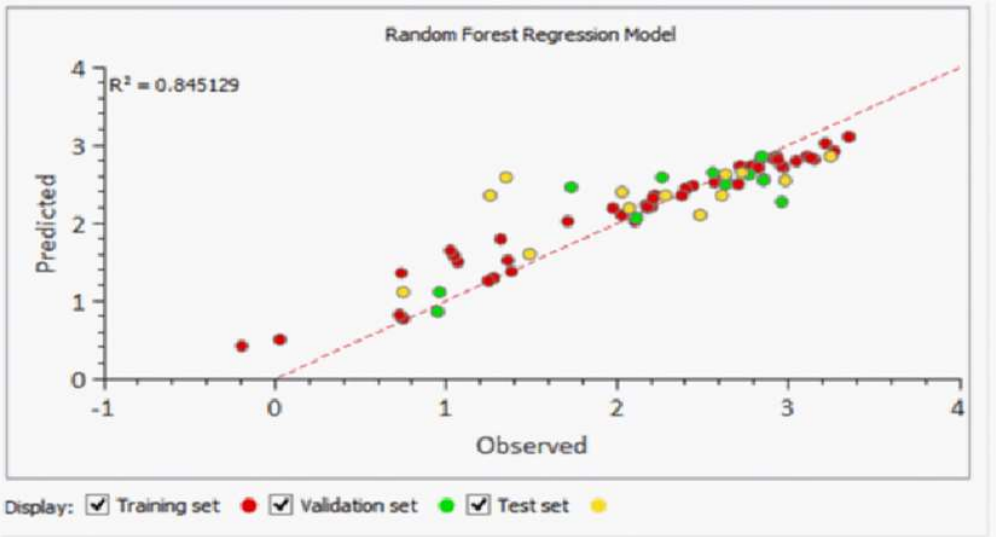

Figure 6.1. Log P model generated on Stardrop **(Tjollyn *et al*., 2011)** platform using Random forest algorithm showing the regression coefficient of 0.845129 with predicted log P-Value on Y-axes and measured log P value on X-axes. The training set is in Red, the Validation set is in Green color, and the Test set is in Yellow color.

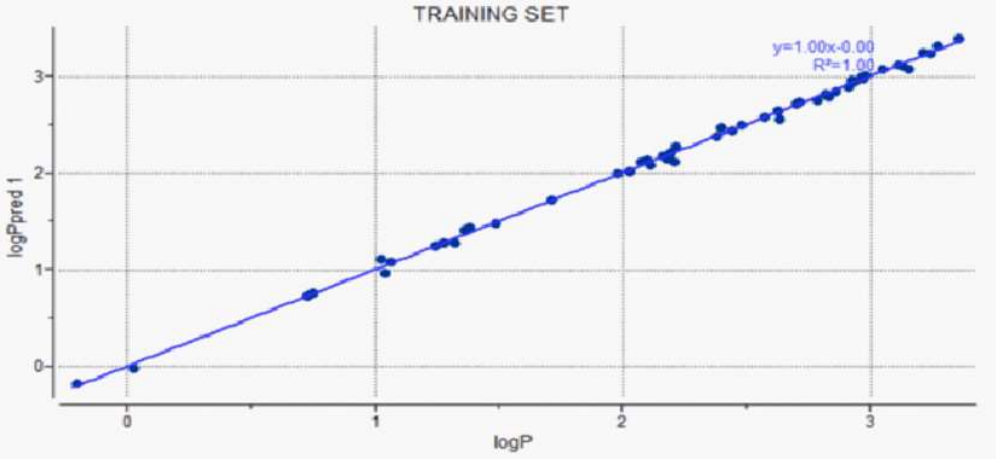

Figure 6.*2*. Log P Model **(Tjollyn *et al*.,2011)**: Traning set showing Regression coefficient of 1.0 with predicted log P-Value on Y-axes and measured log P value on X-axes

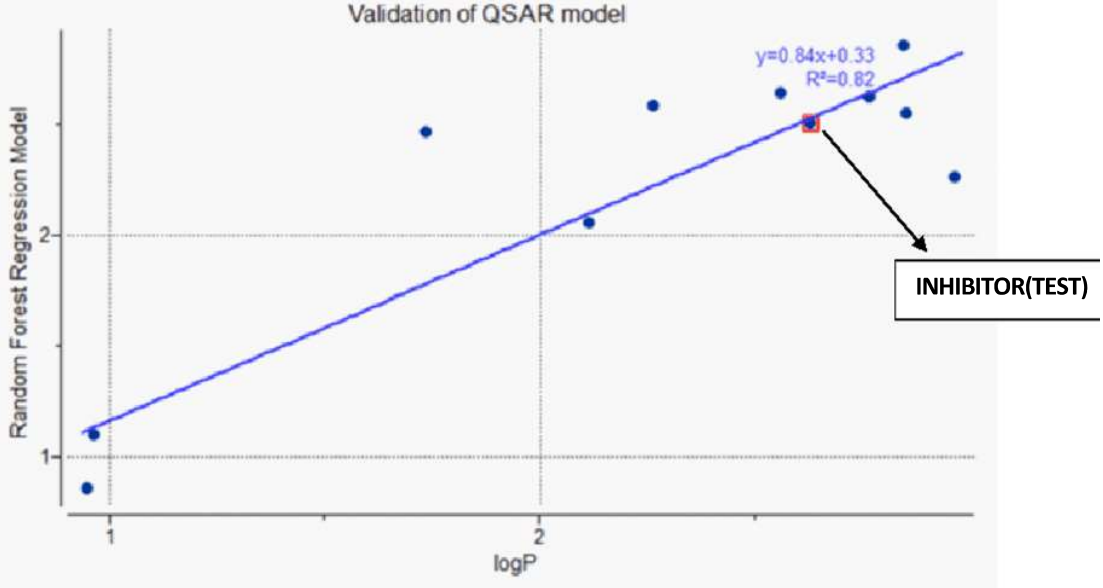

Figure 6.3. Log P Model **(T'jollyn *et al.*,2011)**: Validation set showing Regression coefficient of 0.82 with predicted log P-Value on Y-axes and measured log P value on X-axes and set showing Inhibitor molecule (test) with 2.97 predicted value of log P on Y-axes.

6.4.3. Log $_{IC50}$ Model

The proposed log $_{IC50}$ model was implemented through a deep neural network, and H. generated a DNN algorithm on the OCHEM platform. The descriptors used were CDK 2.0 constitutional, hybrid, topological, electrostatic, and geometric. A training set of 310 molecules with similarity to the test molecule was selected, as shown in **Table 6.4.,** and further subdivided into a validation set of 69 molecules, as shown in **Table 6.5. Figure 6.4.** shows the linear graph of the log IC_{50} model with a training set R2 value of 0.74 + 0.03 kow. The model accounted for 74% of the variance in log $_{IC50}$ for the training set, and it was best to adapt it because the value of the regression coefficient was close to 1. q2 value of 0.73+ 0.03 is comparable to the R2 states predictive relevance of the model, which we

obtained after applying the QSAR model to the training set. RMSE is the square root of the variance and predicts the absolute fit of the model, or model accuracy, by considering the closeness between the observed and predictive values. The RMSE value for the training set is 0.69 + 0.04, and a lower value indicates a good fit. The mean absolute error (MAE) shows the predictive performance of the qsar models, which for the training set is 0.49 + 0.03, while an MAE close to zero is a perfect scalar model. A 5-fold cross-validation was performed using the 10-molecule set, and the resulting R2 was 0.87 + 0.04, indicating a good fit of the model, while the q2 value was 0.85 + 0.04, which is similar to the R2 of the validation set; thus, it is evident that the model is a good fit for the test set to be used. The root means square error (RMSE) for the validation set was 0.6 + 0.1, and the mean absolute error (MAE) was 0.42 + 0.06, which is close to zero, suggesting a perfect qsar model.

6.4.4. Graph Analysis of Log $_{IC50}$ model using Stardrop

The log $_{IC50}$ model with a training set has a regression coefficient of 0.74, with the predicted log $_{IC50}$ on the Y axes and the measured log $_{IC50}$ on the X-axis. The model showed a 74% of variance in $_{IC50}$ for the training set, making it the model with the best fit. Log $_{IC50}$ model cross-validation set a regression coefficient of 0.87, with the predicted log $_{IC50}$ and the measured log $_{IC50}$ following the most appropriate model. Cross-validation is 87%. The predicted log $_{IC50}$ value was 7.17 -log M for the inhibitor molecule, whereas, for gedunin, it was 0.70 –Log. 0.74 with the predicted log $_{IC50}$ value and measured logarithmic $_{IC50}$ value. **(Figure 6.4., 6.5., 6.6.)**

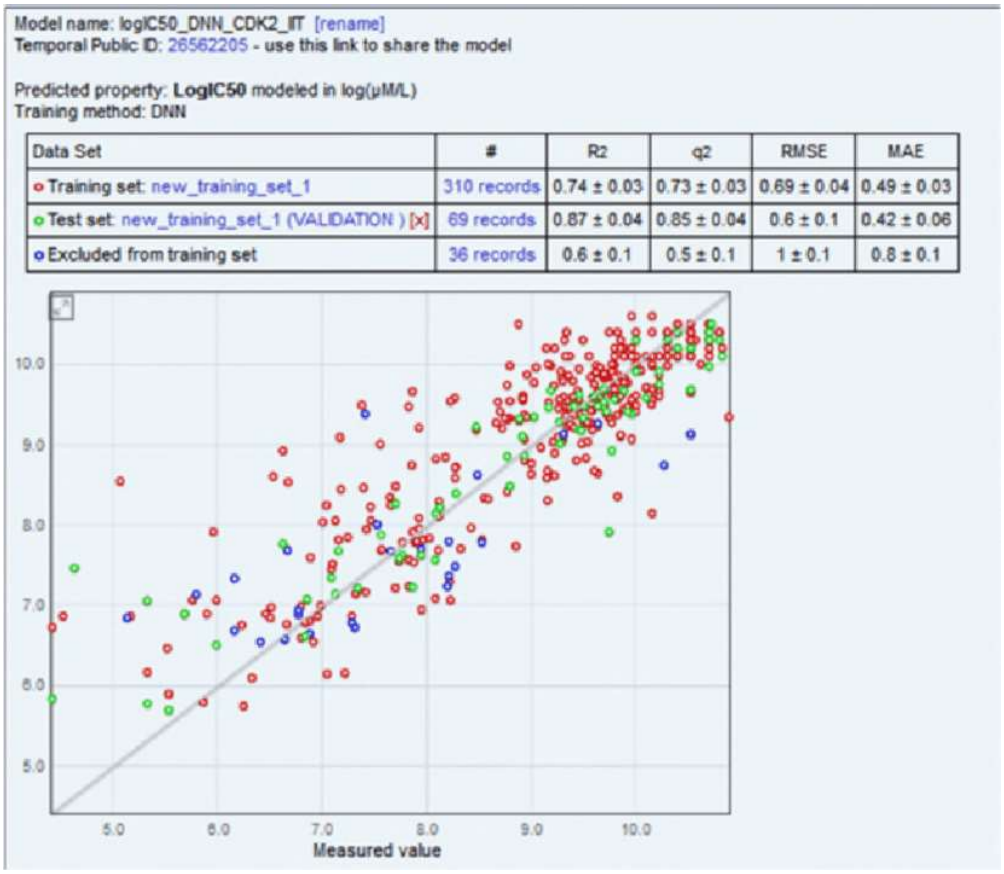

Figure 6.4. Log $_{IC50}$ model generated on the OCHEM platform **(Li *et al.*,2017)** using DNN (deep neural network)algorithm with predicted log $_{IC50}$ Value on Y-axes and measured log $_{IC50}$ value on X-axes. The training set is in Red; the Validation set is in Green color where as the blue color depicts excluded molecules from the generated library.

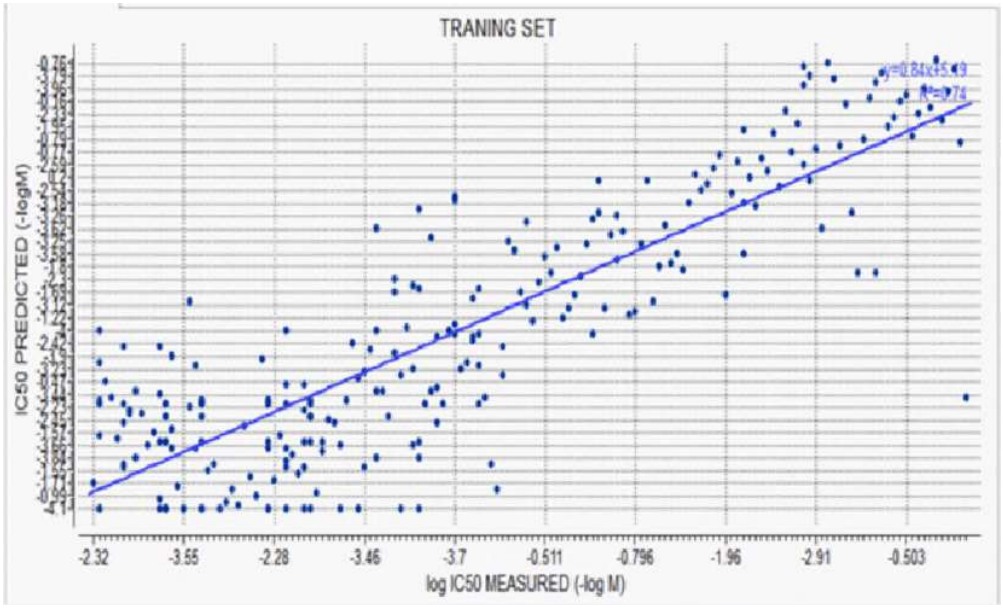

Figure 6.5. Log $_{IC50}$ model **(Li*et al.*, 2017)**: Training set with Regression Coefficient 0.74 where predicted log $_{IC50}$ Value on Y-axes and measured log IC_{50} value on X-axes.

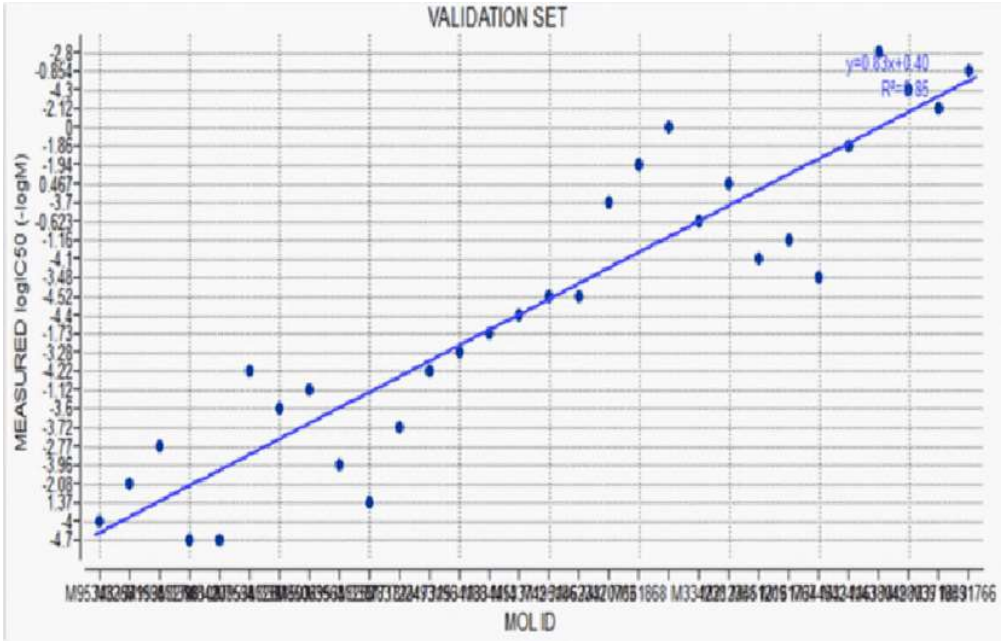

Figure 6.6. Log $_{IC50}$ model **(Li *et al.*, 2017)**: Validation set with Regression Coefficient 0.85 where predicted log $_{IC50}$ Value on Y-axes and measured log IC_{50} value on X-axes.

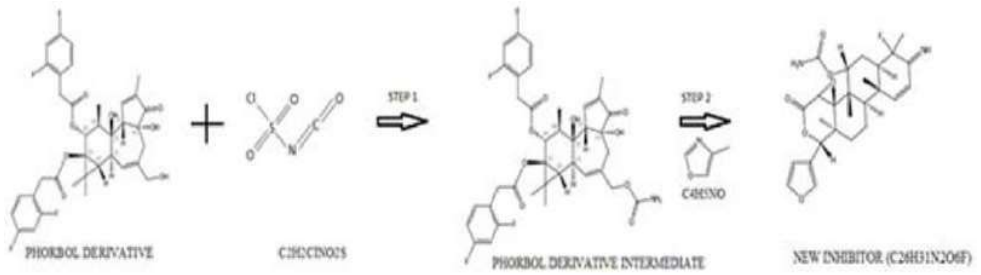

Figure 6.7. Retrosynthesis pathway for the formation of $C_{26}H_{31}N_2O_6F$ inhibitor Molecu*l*e.

Table 6.1. Training Set of 66 molecules for QSAR prediction of log P Model depicting measured log P, Predicted log P (log P pred), and experimentally predicted log P value (logPprex) developed using Random forest algorithm on Stardrop Modeller**(T'jollyn *et al*., 2011)**

Training Set for QSAR			
Id of Molecule	Log P	logPpred_1	logPprex_1
MOL ID 1	2.195	2.204705	2.21233
MOL ID 2	1.065	1.075012	3.953819
MOL ID 3	2.82	2.808868	2.807411
MOL ID 6	1.383	1.434635	0.933455
MOL ID 7	2.981	3.003877	1.886763
MOL ID 8	2.399	2.464688	2.455819
MOL ID 11	2.926	2.931093	2.766559
MOL ID 12	1.489	1.475707	2.973529
MOL ID 13	2.03	2.013026	2.002546
MOL ID 14	3.243	3.228829	1.352993
MOL ID 15	2.182	2.178203	2.104262
MOL ID 17	0.7514	0.753973	0.774199
MOL ID 19	2.971	2.968238	2.45952

MOL ID 20	2.628	2.63882	1.697035
MOL ID 21	3.05	3.068896	2.738283
MOL ID 22	1.278	1.272162	1.247167
MOL ID 23	0.7281	0.722587	0.721414
MOL ID 24	2.703	2.713222	2.710154
MOL ID 25	1.245	1.240723	1.282824
MOL ID 26	2.03	2.013026	2.031905
MOL ID 28	3.113	3.114334	1.360316
MOL ID 29	0.7514	0.753973	0.752277
MOL ID 30	3.272	3.311323	2.48012
MOL ID 31	1.361	1.405386	1.381493
MOL ID 32	3.155	3.069502	1.285728
MOL ID 33	2.195	2.204705	2.220485
MOL ID 35	2.18	2.143285	1.105012
MOL ID 38	1.982	1.991274	2.306565
MOL ID 39	2.715	2.730452	1.357853
MOL ID 44	2.913	2.884496	2.908908
MOL ID 45	2.076	2.110018	2.948019
MOL ID 46	2.635	2.551809	1.045272
MOL ID 47	2.574	2.572034	1.62554
MOL ID 49	1.715	1.716383	1.294909
MOL ID 50	1.042	0.958886	1.773425
MOL ID 52	2.833	2.783679	1.175918
MOL ID 53	2.165	2.172641	1.381673
MOL ID 55	2.965	2.988725	3.291439
MOL ID 56	1.322	1.271959	2.320146
MOL ID 59	2.628	2.637969	2.499268
MOL ID 60	2.925	2.958056	0.526438
MOL ID 62	2.443	2.433344	1.702667
MOL ID 63	2.18	2.143285	1.105012

MOL ID 64	2.216	2.271109	3.718797
MOL ID 65	3.357	3.384725	3.366265
MOL ID 67	0.029	-0.023577	1.876707
MOL ID 68	3.357	3.384725	3.345509
MOL ID 69	2.861	2.83696	0.792936
MOL ID 70	2.787	2.747436	1.696683
MOL ID 71	2.399	2.464688	2.401561
MOL ID 73	0.7363	0.735474	2.252285
MOL ID 74	2.211	2.11129	2.935341
MOL ID 75	2.481	2.492061	3.195806
MOL ID 77	2.101	2.131754	1.322759
MOL ID 79	2.628	2.63882	1.697035
MOL ID 80	2.82	2.808868	2.809107
MOL ID 81	1.278	1.272162	1.35358
MOL ID 82	2.381	2.373829	3.057157
MOL ID 83	1.245	1.240723	1.247228
MOL ID 85	3.214	3.238956	4.47337
MOL ID 88	3.13	3.103463	2.192053
MOL ID 89	-0.2	-0.18224	2.755896
MOL ID 90	0.7281	0.722587	0.73447
MOL ID 91	2.941	2.94306	1.998598
MOL ID 93	2.114	2.077585	2.366496
MOL ID 94	1.025	1.105654	2.749936
MOL ID 95	2.703	2.713222	2.708382

Table 6.2.The validation set of 14 molecules for the Log P Model depicting measured log P and predicted log P values using the Random forest algorithm on Stardrop Modeller**(T'jollyn *et al*., 2011)**

Validation Table for QSAR		
Id of Molecule	Log P	Random Forest Regression Model predicted log P value
MOL ID 4	2.628	2.505
MOL ID 5	2.766	2.623
MOL ID 10	2.628	2.505
MOL ID 18	1.736	2.465
MOL ID 34	2.114	2.054
MOL ID 36	2.965	2.261
MOL ID 42	2.263	2.583
MOL ID 48	2.56	2.64
MOL ID 51	2.846	2.855
MOL ID 57	2.766	2.623
MOL ID 61	2.852	2.549
MOL ID 84	0.9473	0.8575
MOL ID 86	0.9473	0.8575
MOL ID 92	0.9631	1.099

Table 6.3. Test set including Inhibitor molecule for Log P Model depicting measured log P, Experimentally predicted log P (logPprex), and predicted log P (logPpred) values using Random forest algorithm on Stardrop Modeller **(T'jollyn *et al.*, 2011)**

The test set for QSAR			
Id of Molecule	Log P	logPpred	logPprex
MOL ID 9	3.243	1.818537	1.551188
MOL ID 16	1.489	1.972312	1.831624
MOL ID 27	1.26	1.616524	1.563422
MOL ID 37	2.733	2.496129	2.338939
MOL ID 40	1.352	1.758023	1.736135
MOL ID 41	2.281	2.667522	2.490457
MOL ID 43	0.7465	2.018868	1.96778
MOL ID 54	2.076	2.052265	1.725932
MOL ID 58	1.26	1.616524	1.573543
MOL ID 66	2.628	2.491805	2.412945
MOL ID 72	2.481	2.169213	2.078671
MOL ID 76	2.031	2.195364	2.248183
MOL ID 78	2.981	2.172859	1.941746
MOL ID 87	2.608	2.264297	2.037676
Inhibitor $C_{26}H_{31}N_2O_6F$	2.589	2.97	3.21

Table 6.4. A training set of 310 molecules for the $_{IC50}$ Model depicting measured and predicted log $_{IC50}$ values in (-log(M)) unit using the DNN algorithm on the OCHEM platform. **(Li *et al.*, 2017)**

No.	Molecule for the Training set	Log_IC_50 PREDICTED (-log(M))	Log_IC_50 MEASURED (-log(M))
1	M333646	-1.71	-2.32
2	M335289	-2.73	-2.28
3	M460283	-1.79	-1.76
4	M192993	-2.15	-4.15
5	M334414	-1.55	-1.73
6	M332760	-2.23	-1.46
7	M467743	-1.15	-1.32
8	M672297	-3.56	-3.96
9	M3381750	-3.87	-3.82
10	M180781	-3.87	-2.96
11	M450467	-3.84	-3.57
12	M450860	-4.1	-3.37
13	M95343249	-3.37	-2.93
14	M95343259	-3.61	-2.92
15	M95343258	-4.1	-4
16	M3381752	-3.2	-2.72
17	M85063952	-1.97	-2.42
18	M83571995	-2.83	-2.08
19	M324965	-2.81	-3.43
20	M638166	-3.06	-3.3
21	M325505	-2.9	-3.22
22	M162231	-1.57	-1.83
23	M367004	-3.07	-2.94
24	M325120	-3.98	-2.8
25	M418192	-2.25	-1.04

26	M325762	-3.66	-1.86
27	M195554	-4.1	-4.52
28	M534188	-3.04	-3.28
29	M533742	-4.2	-4.4
30	M83738256	-2.77	-3
31	M451895	-4.1	-4.15
32	M454773	-3.83	-3.89
33	M195555	-3.89	-3.62
34	M450386	-4.1	-3.55
35	M95343271	-1.17	-1.41
36	M648555	-4.1	-4.7
37	M209179	-2.45	-1.18
38	M225063	-4.5	-4.7
39	M672289	-0.99	-0.793
40	M377754	-4.5	-4.4
41	M3387724	-2.61	-0.53
42	M376847	-4.4	-4.3
43	M225005	-4.2	-4.3
44	M429344	-4.4	-4.3
45	M475432	-4.3	-4.3
46	M326413	-0.47	0.48
47	M228163	-4.3	-4.3
48	M473196	-4.1	-4.22
49	M371901	-0.87	0.83
50	M377753	-4.4	-4
51	M339177	-3.77	-3.92
52	M431583	-4.4	-3.82
53	M470950	-4.1	-3.8
54	M3391516	-3.52	-3.41
55	M228048	-4.2	-3.8
56	M83723926	-3.47	-3.33
57	M472176	-4.4	-3.74

58	M522677	-4.4	-3.34
59	M95343274	-2.42	-2.77
60	M642143	-4.5	-2.88
61	M464796	-4.3	-4.7
62	M84207595	-4.3	-4.52
63	M83722698	-2.11	-2.11
64	M465323	-4.1	-4.4
65	M366646	-2.81	-2.94
66	M633372	-3.27	-2.66
67	M3398108	-4.5	-4.7
68	M3398105	-4	-4.62
69	M341043	-4.1	-3.33
70	M3398106	-4.3	-4.57
71	M3398117	-3.9	-3.62
72	M465488	-4.1	-4
73	M501249	-4.4	-4.7
74	M195128	-3.98	-4.15
75	M291435	-4.5	-4.52
76	M537284	-4.4	-4.52
77	M506544	-4.4	-4.4
78	M3398107	-4.4	-4.8
79	M292870	-4.4	-4.3
80	M504990	-4.3	-4.7
81	M291040	-4.6	-4.15
82	M291410	-4.6	-3.96
83	M84207593	-4.5	-4.7
84	M84207594	-4.2	-4.7
85	M84207586	-4.1	-4.52
86	M84207591	-4.4	-4.4
87	M84207592	-4.2	-4.4
88	M215990	-4.1	-4.1
89	M84207590	-4.3	-4

90	M3381760	-3.7	-4.1
91	M3381762	-3.92	-4.05
92	M3381748	-3.88	-3.74
93	M3381756	-3.82	-3.7
94	M3381749	-3.94	-3.46
95	M64283	-3.23	-3.35
96	M3381755	-3.33	-2.8
97	M3381753	-3.41	-2.7
98	M189970	-3.56	-3.82
99	M85063957	-3.21	-1.92
100	M340248	-1.96	-1.92
101	M434294	-0.87	-0.955
102	M533413	-0.1	-0.328
103	M450974	-3.44	-3.89
104	M445046	-3.74	-3.8
105	M189512	-3.84	-3.77
106	M451420	-3.87	-3.68
107	M420482	-3.66	-3.64
108	M453139	-3.23	-3.64
109	M453004	-3.27	-3.57
110	M420484	-3.04	-3.54
111	M450939	-3.48	-3.51
112	M445363	-3.25	-3.38
113	M192979	-3.46	-3.37
114	M85063948	-3.83	-3.29
115	M190472	-4.2	-3.22
116	M449314	-3.55	-2.92
117	M85063956	-2.09	-1.92
118	M85063958	-1.79	-1.88
119	M85063959	-2.25	-1.65
120	M85063962	-0.87	-1.28
121	M85063963	-2.06	-1.12

122	M85063964	-0.6	-0.796
123	M85063966	-0.98	-0.511
124	M162076	-1.82	-2.52
125	M434891	-0.95	-1.95
126	M429343	-1.8	-1.9
127	M162144	-1.22	-1.7
128	M432349	-2.35	-1.65
129	M119385	-1.45	-1.09
130	M162152	-1.07	0.242
131	M672298	-3.37	-3.82
132	M9472	-2.65	-3.64
133	M637700	-3.98	-3.62
134	M672299	-3.21	-3.41
135	M672303	-3.79	-3.3
136	M3397217	-2.34	-2.55
137	M672301	-3.09	-1.17
138	M95343263	-3.34	-4.89
139	M95343260	-3.66	-4.22
140	M95343255	-3.77	-3.89
141	M83577592	-2.84	-3.52
142	M3386440	-3.26	-3.52
143	M83738249	-3.27	-3.51
144	M3382335	-3.38	-3.46
145	M83577587	-3.37	-3.41
146	M1234	-3.09	-3.25
147	M95343251	-0.9	-0.46
148	M633036	-3.49	-4
149	M372860	-3.3	-2.89
150	M422395	-3.58	-2.27
151	M210995	-2.3	-2.11
152	M538662	-1.84	-2.02
153	M371573	-1.69	-1.56

154	M211818	-2.47	-1.39
155	M536946	-1.6	-0.886
156	M212657	-0.9	0.103
157	M95343257	-3.07	-3.96
158	M95343253	-3.56	-3.96
159	M3391507	-3.11	-3.85
160	M83738247	-3.12	-3.85
161	M3391505	-3.62	-3.77
162	M83723931	-3.69	-3.74
163	M3389886	-3.33	-3.74
164	M83722493	-3.38	-3.72
165	M324998	-3.53	-3.66
166	M3391503	-3.75	-3.6
167	M95343254	-3.31	-3.49
168	M15407	-3.56	-3.46
169	M83738949	-3.46	-3.46
170	M83738951	-3.59	-3.31
171	M83738222	-3.34	-2.96
172	M95343252	-3.74	-2.77
173	M95343264	-3.6	-3.85
174	M95343265	-3.27	-3.77
175	M95343268	-3.67	-3.7
176	M95343280	-3.96	-4
177	M95343286	-2.75	-1.85
178	M83738265	-3.18	-2.47
179	M95343283	-3.31	-2.77
180	M83738246	-3.52	-3.17
181	M3391499	-3.04	-3.21
182	M348129	-2.54	-0.673
183	M371208	-3.88	-3.72
184	M448557	-1.52	-1.1
185	M3382334	-3.47	-1.82

186	M368768	-0.81	-0.879
187	M672304	-3.47	-3.92
188	M319779	-3.86	-4.22
189	M434410	-0.77	-0.658
190	M161767	-1.82	-1.16
191	M342313	-3.57	-2.91
192	M85063954	-1.82	-1.96
193	M450275	-3.18	-3.59
194	M452830	-3.6	-3.8
195	M95343262	-2.55	0.929
196	M85063949	-2.59	-3.15
197	M83736823	-3.65	-4.52
198	M3384132	-3.33	-2.74
199	M3381761	-3.51	-3.7
200	M3381763	-3.85	-3.7
201	M84207588	-4.2	-3.89
202	M84207587	-4.1	-4.3
203	M84207585	-4.1	-4.3
204	M3398120	-4.3	-3.32
205	M95343279	-4.2	-4.82
206	M3398119	-4.3	-3.49
207	M84207596	-4.4	-4.15
208	M3398116	-4.5	-4.72
209	M95343278	-3.01	-1.55
210	M95343275	-4.1	-3.89
211	M208678	-0.17	0.67
212	M154491	0.2	0.138
213	M95343272	0.25	-0.246
214	M224173	-4.3	-4.52
215	M95343270	-1.3	-2.21
216	M452372	-4.1	-4.4
217	M3384033	-2.49	-1.7

218	M321408	-3.98	-3.19
219	M465121	-1.09	-2.08
220	M365709	-3.59	-3.4
221	M3347641	-2.64	-3
222	M83739113	-3.57	-2.92
223	M95343285	-3.58	-3.59
224	M95343282	-3.59	-3.66
225	M334238	-2.93	-0.623
226	M332706	-1.85	-1.24
227	M335540	-1.54	-1.87
228	M3381625	-4.1	-4.4
229	M3381759	-3.69	-4.15
230	M83736821	-3.56	-4.15
231	M672302	-3.5	-4.15
232	M335042	-1.7	-1.57
233	M451209	-3.71	-4.1
234	M187417	-3.61	-4
235	M335649	-2.04	-1.01
236	M85063946	-2.36	-3.82
237	M3397216	-3.99	-3.8
238	M344542	-3.02	-3.48
239	M83724194	-3.77	-3.03
240	M672305	-2.62	-3.22
241	M451622	-4.2	-3.14
242	M3381757	-3.96	-3.05
243	M95343250	-2.85	-2.17
244	M85063950	-1.74	-2.85
245	M381064	-3.62	-4.12
246	M3381754	-3.53	-2.68
247	M322895	-3.77	-3.82
248	M364291	-2.33	-2.59
249	M322415	-4	-4.15

250	M189501	-1.07	-2.22
251	M380807	-4.3	-3.85
252	M3384034	-3.54	-2.22
253	M334143	-1.92	-1.86
254	M520642	-0.16	-1.21
255	M95343269	-3.69	-3.8
256	M83738254	-3.56	-3.27
257	M83596819	-2.97	-3.09
258	M455817	-3.98	-4.3
259	M83718093	-2.59	-2.27
260	M455118	-3.84	-3.8
261	M83573551	-1.95	-1.42
262	M196877	-3.91	-3.28
263	M371498	-0.79	-0.836
264	M83583537	-1.92	0.041
265	M475851	-4	-4
266	M83583567	-0.73	1.58
267	M474020	-4.2	-3.96
268	M83738936	-3.43	-3.59
269	M638042	-3.54	-2.8
270	M95343277	-1.24	-1.82
271	M462342	-4.2	-4.52
272	M372353	-4.1	-4.4
273	M381870	-3.73	-3.21
274	M3398110	-4.4	-4.8
275	M370763	-3.88	-3.7
276	M84207589	-4.2	-3.85
277	M470563	-4.5	-4.7
278	M498339	-4.5	-4.4
279	M3398118	-3.79	-3.4
280	M498039	-4.3	-4.3
281	M615817	-4.4	-4.52

282	M3381758	-3.36	-3.33
283	M3381751	-3.45	-3.26
284	M449768	-4	-3.77
285	M453102	-3.91	-3.77
286	M446204	-3.96	-3.66
287	M452009	-3.28	-3.54
288	M445045	-3.67	-3.44
289	M450481	-3.86	-3.36
290	M85063970	-1.07	0.013
291	M85063967	-0.85	-0.503
292	M162382	-1	-0.983
293	M672300	-2.68	-3.15
294	M367142	-2.72	-2.27
295	M371859	-1.69	-2.11
296	M370409	-3.6	-4.22
297	M211042	-2.68	-3.6
298	M211683	-4	-3.16
299	M373236	-2.31	-3.15
300	M374245	-2.06	-1.46
301	M210637	-0.55	-0.914
302	M209173	-0.76	-0.23
303	M212358	0.1	0.467
304	M208666	-0.87	1.48
305	M95343256	-3.46	-3.7
306	M95343284	-3.42	-3.57
307	M95343266	-3.41	-4
308	M95343267	-4.1	-3.96
309	M95343281	-0.15	-1.05
310	M95343287	-3.49	-1.38

Table 6.5. The validation set of 69 molecules and Inhibitor and gedunin Test Molecule depicting measured and predicted log $_{IC50}$ values in a log-M unit for the $_{IC50}$ Model using the DNN algorithm on the OCHEM platform**(Li *et al.*,2017)**

Sr. No	Molecule for the Validation set	Log $_{IC50}$ MEASURED (-Log M)
1	M335289	-2.28
2	M460283	1.76
3	M334414	-1.73
4	M95343258	-4
5	M324998	-3.64
6	M83571995	-2.08
7	M195186	-4.52
8	M534188	-3.28
9	M533742	-4.4
10	M454773	-3.89
11	M473196	-4.22
12	M377753	-4
13	M3391516	-3.41
14	M95343274	-2.77
15	M366646	-2.94
16	M3398108	-4.7
17	M501249	-4.7
18	M84207594	-4.7
19	M84207592	-4.4
20	M3381749	-3.46
21	M3387718	1.37
22	M445046	-3.8
23	M451420	-3.68
24	M85063963	-1.12

25	M433765	-1.34
26	M119385	-1.09
27	M163152	0.316
28	M95343260	-4.22
29	M83738249	-3.51
30	M3382335	-3.46
31	M213028	-3.74
32	M372860	-2.89
33	M371573	-1.56
34	M95343253	-3.96
35	M83722493	-3.72
36	M3391503	-3.6
37	M95343254	-3.49
38	M95343265	-3.77
39	M83738265	-2.47
40	M83738246	-3.17
41	M672304	-3.92
42	M161767	-1.16
43	M342313	-2.91
44	M3387717	0.672
45	M3381763	-3.7
46	M95343279	-4.82
47	M3398116	-4.72
48	M208678	0.67
49	M3384033	-1.7
50	M321408	-3.19
51	M465121	-2.08
52	M334238	-0.623
53	M451209	-4.1
54	M344542	-3.48
55	M83724194	-3.03
56	M334143	-1.86

57	M83738254	-3.27
58	M371859	-2.12
59	M371498	-0.836
60	M83583567	1.58
61	M638042	-2.8
62	M462342	-4.52
63	M3398110	-4.8
64	M370763	-3.7
65	M498039	-4.3
66	M431766	-0.854
67	M85063970	0.013
68	M161868	-1.94
69	M212358	0.467
	INHIBITOR PREDICTED VALUE	7.17 (-logM)
	Gedunin	0.70 (–logM)

6.5 Conclusion

In this study, QSAR models were generated with machine learning algorithms such as random forest and DNN (deep neural network) together with 5-fold cross-validation to predict log P and log $_{IC50}$ using descriptors from CDK2.0 in the case of DNN and molecular weight, log P in the case of random forest, and with a molecular data set that is similar to the inhibitor molecule. The new inhibitor molecule predicted Log P IS 2.97 kow, and Log $_{IC50}$ is 7.17 (-log M unit) and 0.70 (-log M unit) of gedunin. The log-P and Log-IC$_{50}$ models fit best, with an R2 close to 1 and a low RMSE value. Retrosynthesis of modified gedunin showed a

novel pathway of synthesis without the involvement of gedunin as a reagent or intermediate