

# Exploring Molecular Descriptors and Fingerprints to Predict mTOR Kinase Inhibitors using Machine Learning Techniques

Chetna Kumari, Muhammad Abulaish, SMIEEE and Naidu Subbarao

**Abstract**—Mammalian Target of Rapamycin (mTOR) is a Ser/Thr protein kinase, and its role is integral to the autophagy pathway in cancer. Targeting mTOR for therapeutic interventions in cancer through autophagy pathway is challenging due to the dual roles of autophagy in tumor progression. The architecture of mTOR reveals two complexes – mTORC1 and mTORC2, each having multiple protein subunits. mTOR kinase inhibitors target the structurally and functionally similar catalytic subunits of both mTORC1 and mTORC2. In this paper, we have explored two different categories of molecular features – *descriptors* and *fingerprints* for developing predictive models using machine learning techniques. Random Forest *variable importance measures* and autoencoders are used to identify molecular descriptors and fingerprints, respectively. We have built various predictive models using identified features and their combination for predicting mTOR kinase inhibitors. Finally, the best model based on the Mathew correlation co-efficient value over the validation dataset is selected for screening kinase SARfari bioactivity dataset. In this study, we have identified twenty best performing descriptors for predicting mTOR kinase inhibitors. To the best of our knowledge, it is the first study on integrating traditional machine learning and deep learning-based approaches for feature extraction to predict mTOR kinase inhibitors.

**Index Terms**—Drug Discovery, Kinase, mTOR, Autophagy, Molecular Descriptor, Fingerprints, Machine Learning, Deep Learning.



## 1 INTRODUCTION

Kinases belong to one of the prominent family of drug targets in human [1]. The discovery of phosphatidylinositol 3-kinase related kinases (PIKK) family in 1990 dramatically changed the landscape of stress-induced field. PIKK includes six members in human, such as ataxia telangiectasia mutated (ATM), ataxia telangiectasia- and RAD3-related (ATR), the catalytic subunit of DNA-dependent protein kinase (DNA-PK), human suppressor of morphogenesis in genitalia-1 (hSMG-1), transformation/transcription domain-associated protein (TRRAP), and mammalian Target of Rapamycin (mTOR) [2]. mTOR is a Ser/Thr protein kinase as it transduces the cellular signals through phosphorylation of serine/threonine residues of substrate molecules in upstream and/or downstream pathways. The structure of mTOR reveals two complexes – mTOR complex1 (mTORC1) and mTOR complex2 (mTORC2), having two and three unique protein subunits, respectively; while four protein subunits are common in both of the complexes [3]. A summarized view of the topological variations of mTOR complexes are shown in Table 1. The crystal structures of mTOR are deposited in Protein Data Bank (PDB) [4]. A 3D view of mTOR with one of its inhibitors, Torin-2 (PDB id:

4JSX), is visualized using Pymol software [5] and shown in Figure 1.

TABLE 1: Protein subunits of mTOR complexes (Laplante and Sabatini [3])

Abbreviations	Protein subunits Full name	mTOR Complexes	
		mTORC1	mTORC2
Raptor	Regulatory protein associated with mTOR	✓	✗
PRAS40	Proline-rich Akt substrate 40 kDa	✓	✗
Rictor	Rapamycin-insensitive component of mTOR	✗	✓
mSin1	Stress-activated mapk-interacting protein 1 in mammal	✗	✓
Protor1/2	Protein with rictor 1 & 2	✗	✓
mTOR	Core catalytic subunit	✓	✓
mLST8	mammalian Lethal with sec-13 protein 8	✓	✓
DEPTOR	DEP domain comprising mTOR-interacting protein	✓	✓
	Tti1/Tel2 complex	✓	✓

The mTOR has clinical interventions in various human diseases, such as cancer, type II diabetes, cardiovascular diseases, obesity, autoimmunity, neurodegeneration, and aging [6]. It also plays a vital role in male fertility [7]. The altered status of mTOR in various diseases, its functional role in initiation, maintenance or progression of the diseases, and its resistance to conventional chemotherapy make it drug-gable. Therefore, mTOR is culminated as a well-established pharmacological target.

Autophagy is a dynamic cellular process for recycling intracellular nutrients which helps the eukaryotic cells to adjust metabolism to survive during the adverse growth conditions [8]. Autophagy is primarily known to be regulated by mTORC1; however, there are evidences of its

- C. Kumari is currently with the Department of Computer Science, Jamia Millia Islamia (A Central University), Delhi, India. E-mail: k.chetna@gmail.com
- M. Abulaish (corresponding author) is currently working as an Associate Professor at the Department of Computer Science, South Asian University, Delhi, India. E-mail: abulaish@sau.ac.in
- N. Subbarao is currently working as an Associate Professor at the School of Computational and Integrative Biology, Jawaharlal Nehru University, Delhi, India. E-mail: nsrao@jnu.ac.in



Fig. 1: A 3D visualization of mTOR with Torin-2 (PDB id: 4JSX) inhibitor using Pymol software

regulation by mTORC2 as well [9]. The role of mTOR is integral to the autophagy pathway in cancer. Autophagy acts as suppressor during tumor initiation, whereas it acts as promotor during tumor progression by enabling tumor cells to adapt to the changes in nutrient availability. The modulation and regulation of autophagy pathway through mTOR, and dual roles of autophagy in tumor cells make mTOR a challenging and promising anticancer drug target [10], [11].

### 1.1 Generations of mTOR Inhibitors

A molecule which binds to the pharmacological target and decreases its activity is known as *inhibitor*. Till date, three generations of mTOR inhibitors are known [12]. Clinically approved inhibitor Rapamycin (or sirolimus) binds to the site other than the catalytic subunit of mTORC1. As a result, it is also known as allosteric inhibitor of mTOR. Rapamycin and its analogs (rapalogs) are the first generation mTOR inhibitors. Few known ATP-competitive mTOR inhibitors or mTOR kinase inhibitors (mTOR KIs), such as AZD-8055, CC-223, Torin 2 bind to the catalytic subunits of both mTOR complexes which are structurally and functionally similar. These are the second generation mTOR inhibitors, and considered more efficient than rapamycin and rapalogs. The cumulative advantages of combining rapamycin and AZD-8055 due to their binding site proximity on mTOR give rise to the third generation mTOR inhibitor – RapaLink-1 [13]. The second generation mTOR inhibitors (or mTOR KIs) have therapeutic advantages over the first generation mTOR inhibitors, and the third generation mTOR inhibitor has just entered the drug development pipeline. Moreover, few of the known mTOR KIs targeting cancer are undergoing different stages of clinical trials [14]. This motivated us to develop computational approaches to predict new clinical compounds like mTOR KIs. A partial list of six known mTOR kinase inhibitors that are in clinical trial for cancer are shown in Figure 2.

### 1.2 Existing Computational Approaches

Computational approaches to discover novel mTOR kinase inhibitors are extensively reviewed in literatures [21]. The existing approaches are broadly categorized as (i) comparative modeling, (ii) structure-guided virtual screening or molecular docking, (iii) quantitative structure-activity relationship (QSAR), (iv) pharmacophore-based modeling, (v) molecular dynamics simulations, (vi) machine learning, and (vii) similarity-based searching of hit and lead molecules. On limiting our search to machine learning for the development of mTOR kinase inhibitors, we found that a series of classification models have been developed using naive Bayes and recursive partitioning classifiers to predict mTOR kinase inhibitor-like compounds [22]. In a hierarchical study, the classifier-based predictions are integrated to molecular docking and *in vitro* enzyme assays to discover novel mTOR kinase inhibitors [23].

In this study, we have also identified computational approaches to design inhibitors using other targets. These approaches may guide us to predict mTOR kinase inhibitor-like compounds. Random forest algorithm is used for selecting optimal molecular descriptors for ligands of thymidine kinase and other targets [24]. A novel method is developed to predict compound-protein interactions using positive and unlabeled samples, and compared using existing classifiers [25]. Computational intelligence methods in drug discovery has now progressed from machine learning to deep learning using big data platform [26], [27]. In 2012, Merck-sponsored Kaggle competition on chemical compound activity prediction revealed deep learning as a potent tool in drug design [28]. The application of deep learning approaches outside bioactivity prediction is shown by an integrative approach of data analysis on cancer dataset extracted from multiple platforms [29].

In 2015, a drug combination prediction challenge was launched as a part of the DREAM 10 challenge in collaboration with AstraZeneca and the Sanger Institute, and the outcome has recently been reported [30]. The developed computational approaches may help combat short-liveness of cancer targeted therapy. Moreover, the study may provide useful guidelines to predict promising anticancer drug combinations for targeting autophagy pathway, as the autophagy modulators (inhibitor/inducer) are often used in combination with standard treatment. Few of such drug combinations are already in clinical or pre-clinical trial for the treatment of various types of cancer [31].

### 1.3 Our Contributions

Though machine learning technique are extensively used in biomedical domain, limited research efforts has been directed towards predicting mTOR inhibitors. We noticed that the existing approaches for predicting mTOR inhibitors are based on few selective classifiers and limited number of molecular descriptors that are mainly derived from experimental data. In this study, we explore a wide variety of molecular descriptors and fingerprints to predict mTOR inhibitors using four different machine learning techniques. The key contributions of this study can be summarized as follows:

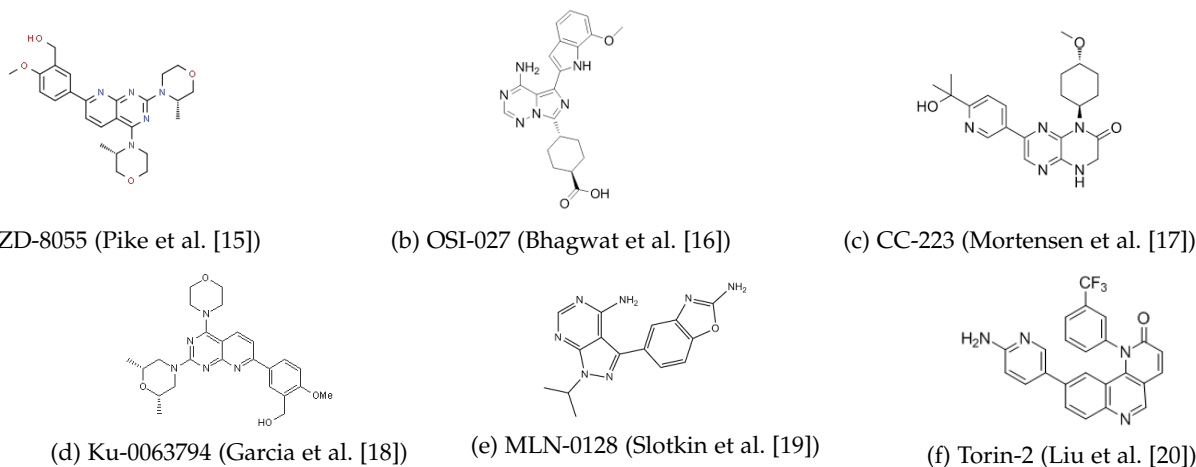


Fig. 2: A partial list of six known mTOR kinase inhibitors that are in clinical trial for cancer

- Exploration of diverse features including *molecular descriptors* and *molecular fingerprints* of mTOR bioactive molecules from ChEMBL database and development of various classification models using four classifiers, namely random forest, support vector machine, decision tree, and neural network to predict mTOR kinase inhibitor-like molecules.
- Application of the variable importance measure to rank and select the molecular descriptors using random forest classifier.
- Application of the neural network-based autoencoders to reduce 166-bits MACCS (Molecular Access System) fingerprints.
- A comparative study of the descriptors-based and fingerprints-based classification models to predict mTOR kinase inhibitors.
- Screening of the kinase SARfari compounds ( $\sim 40,922$  compounds) using the best performing classification model for predicting mTOR kinase inhibitor-like compounds.

## 2 METHODS

### 2.1 Pipeline Implemented

The pipeline of our proposed approach mainly consists of the following steps: (i) mTOR bioactivity data collection, (ii) partition of data into training and validation sets, (iii) molecular feature (descriptors and fingerprints) extraction and dimension reduction, (iv) classification model learning and evaluation, (v) comparative analysis, (vi) best classification model selection, and (vii) Kinase dataset screening. Figure 3 presents the visualization of these steps.

### 2.2 Dataset

The mTOR bioactivity dataset is downloaded from ChEMBL [32] and unique compounds are selected considering only human mTOR kinase inhibition data. As a result, a total number of 1804 unique compounds with  $IC_{50}$  values between  $0.07nM$  and  $50000nM$  (that is,  $0.07nM \leq IC_{50} \leq 50000nM$ ) associated with target id ChEMBL2842 are selected. The 2D structures of the compounds are converted

into 3D structures using CORINA v2.64 software, and molecules are saved in 3D-sdf format. Out of 1804 compounds, 1590 compounds with  $IC_{50}$  values less than  $10\mu M$  are considered as *active* and remaining 214 compounds with  $IC_{50}$  values  $\geq 10\mu M$  are considered as *inactive*. As reported in [22],  $10\mu M$  cut-off value is considered as a reasonable starting point for hit-to-lead activity.

As shown in Table 2, out of the complete dataset of 1804 molecules, 80% (1444) molecules are considered as the training set, whereas remaining 20% (360) molecules are considered as the validation set. The training dataset is again divided into five parts using random sampling technique, such that four parts are used to train classification model, while keeping aside one part to test the model. The training and testing processes are repeated five times such that test sets differ in each iteration, and every molecule gets its participation in training and testing at least once. After validating the trained classification models, the best performing model on the basis of Mathews correlation coefficient (MCC) values over validation set is used for screening a large compound dataset ( $\sim 40,922$  compounds) retrieved from the kinase SARfari database to predict mTOR kinase inhibitor-like compounds.

TABLE 2: Statistics of the mTOR bioactivity dataset extracted from ChEMBL database

ChEMBL id.	Dataset	#Active molecules	#Inactive molecules	Total
ChEMBL2842	Training set	1273	171	1444
	Validation set	317	43	360

### 2.3 Feature Extraction and Dimension Reduction

Prior to splitting the mTOR bioactivity dataset into training and validation sets, 1D/2D/3D descriptors and 166-bits MACCS fingerprints (FPs) are calculated using PaDEL descriptor software [33]. Since the calculated 3D descriptors are sparse, they are excluded from the list of descriptors for further study. A total number of 1444 1D/2D descriptor columns and 166-bits MACCS FPs are filtered separately to remove the redundant columns. As a result, 1171 informative descriptor columns and 133-bits MACCS FPs are retained for further analysis.

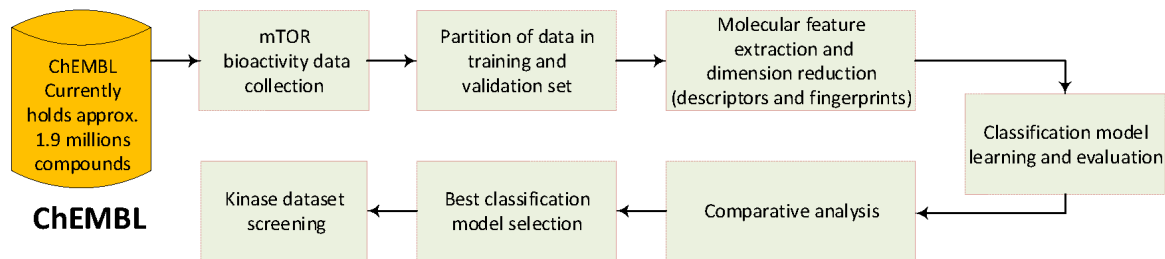


Fig. 3: A visualization of the pipeline of our proposed approach for predicting mTOR inhibitors

### 2.3.1 Variable Importance Measures

Variable Importance Measures (VIMs) are calculated to rank, select, and prioritize the molecular descriptors. Unlike other algorithms, random forest (RF) is not prone to overfitting as the variable selection is randomized during each tree-split. Therefore, VIMs are calculated based on prediction error and splitting criteria using RF classification model. By default, RF takes two-third of the dataset to train the model and keeps one-third of the dataset for testing during regression task, and  $\sim 70\%$  of the data as training set and  $\sim 30\%$  of the data to test the model during classification task. The dataset separated for testing is called out-of-bag (OOB) sample, and it is used to calculate OOB error. The OOB error calculated during model training represents the performance over test dataset, and it is equivalent to the error captured by  $n$ -fold cross-validation. Hence, training can be terminated on stabilization of OOB error [34]. Prediction error is calculated using *the percentage of misclassification* in classification task, and *mean square error* (MSE) in case of regression task which assigns mean decrease accuracy (MDA) values. The prediction error is finally calculated as the difference between out-of-bag (OOB) error calculated from newly generated dataset by randomly changing the order of the selected predictors, and the OOB error calculated from the original dataset. Shifting an important predictor generally increases the OOB error, which ultimately increases the MDA values. On the other hand, decrease in gini index (DGI) is used as splitting criteria to assign mean decrease gini (MDG) values to multiple descriptors present during the construction of the trees (the lowest the Gini index, the purest the split). Hence, selection of an important predictor for splitting is based on high decrease in Gini index, resulting in high MDG. In this study, each of the 1171 molecular descriptors of the training dataset are assigned MDA values based on the *percentage of misclassification* and MDG values using RF classification task.

### 2.3.2 Deep Neural Network-Based Autoencoder

Neural networks with more than one hidden layer are known as deep neural network (DNN). DNN has achieved great success in diverse areas of research, primarily in natural language processing, computer vision, and more recently in bioactivity predictions [28]. The architecture of autoencoders (*aka* autoassociative network) are derived from DNN. Autoencoders have adjustable, multilayer encoder network to generate low-dimensional code by transforming the high-

dimensional input data, and a similar decoder network to retrieve the inputs from the code. Thus, autoencoders receive self-supervised training to extract features with minimal information loss, and are non-linear abstraction of *principal component analysis* [35]. In this study, simple architecture of autoencoder with three hidden layers and each having equal number of nodes are used for dimension reduction of MACCS fingerprints (FPs) using `h2o` library in R. Finally, best performing descriptors set and dimensionally reduced MACCS FPs are integrated to generate hybrid feature set.

## 2.4 Machine Learning Techniques for Compound Classification

Supervised machine learning technique for compound classification for predicting bioactivities needs a labeled training dataset which is designed using known annotated compounds with specific activities. The training set is then used to develop classifiers that assign class labels to unseen samples. In this study, molecular descriptors and/or fingerprints based models are built to predict mTOR kinase inhibitors using four popular classifiers – Random Forest (RF) [36], Support Vector Machine (SVM) [37], Decision Tree (DT), and Neural Network (NN) [38].

RF [36] is an ensemble learning algorithm in which predictions are made either by collecting majority votes or by calculating average predictions of the ensembles. RF builds and averages a large collection of mutually related trees. One of the important features of RF is the use of OOB sample, which is used to calculate unbiased classification error during generation of large number of trees. RF is capable of showing excellent performance in presence of large number of predictors, much higher than the number of observations. Moreover, RF is a well-established data analysis tool in bioinformatics [39], [40] which we have implemented using `randomForest` library in R.

SVM [37] is a supervised learning algorithm, which performs both classification and regression task for prediction. SVM uses maximum-margin to perform the classification task. First, the data vectors are mapped into an  $d$ -dimensional space, then the algorithm finds a hyperplane for the new space with maximal margin between positive (*active*) and negative (*inactive*) class samples in the training dataset. In our implementation,  $d$  represents the number of molecular descriptors whose values are represented by particular co-ordinates. The closest samples on either side of

the hyperplane are known as *support vectors*. In this study, *sigest* function is used to choose the penalty parameter  $c$  and kernel parameter  $g$  through cross-validation method. We have used `e1071` package of R for implementing the LIBSVM [41], and performed classification task using the radial basis function (RBF) kernel.

DT [42] is a supervised learning algorithm, which incorporates both numeric and categorical variables, as well as missing values. DT uses tree-like structure, which consists of a root node, decision nodes, and terminal or leaf nodes to support the decision. Removal of a decision node from the decision tree is called pruning. We have used the CART [42] algorithm which is implemented through `RPART` library in R. CART reduces the size of an overly large grown tree to minimize the estimated misclassification error. The size of the decision tree is controlled by the complexity parameter ( $cp$ ), and it is used to select an optimal tree size. By default, 10-fold cross-validation is employed in CART for model evaluation.

ANN [38] is conceptualized from the architecture of human brain. The simplest unit of a neural network is called neuron. The neurons are organised into layers, called input layers, hidden layers, or output layers depending on their position in the network. ANN is categorized based on the arrangements of neurons in certain topology and connections to each other. In this study, we have used *nnet* function of *caret* library in R to train neural network models. Two hyper-parameters – *size* and *decay* are optimized using cross-validation to generate optimal classification models. The *size* parameter describes the number of units in hidden layer, and *decay* parameter is the standardized parameter which is used to avoid overfitting of the classification model.

### 3 RESULTS

#### 3.1 Prioritization of Molecular Descriptors

In order to identify most discriminative molecular descriptors, the MDA and MDG values are calculated for each of the 1171 descriptors of the training dataset using RF classification model. The error rates during model training are calculated as percentage of the misclassification of active molecules, inactive molecules, and OOB datasets, wherein OOB error is considered to be crucial for selecting the final model. In this study, OOB error is first calculated under the default parameter settings, keeping the number of trees (*ntree*) as 500 and the number of input descriptors to be used in each node (*mtry*) as 34, which is roughly equivalent to the square root of the number of descriptors. As shown in Figure 4, we observe that the OOB error rate remains almost constant after generation of 100 trees. However, in order to avoid overfitting, total 300 trees are generated and four different combinations of *ntree* with *mtry* are used to calculate OOB error values using the `tuneRF` function in R. The OOB error values are shown in Table 3 and visualized in Figure 5, and based on the lowest OOB error, the corresponding values of *ntree* and *mtry* parameters are determined as 300 and 23, respectively to calculate RF-VIMs (MDA and MDG values).

Since there is no any universal rule to prefer one VIM over another, molecular descriptors are selected using both MDA and MDG values. RF automatically takes care of

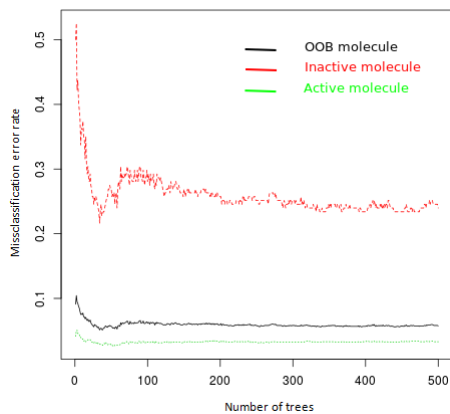


Fig. 4: Misclassification error rate vs. number of trees using default values of the *mtry* and *ntree* parameters

TABLE 3: Tuning *mtry* parameter to calculate RF-VIMs

# Tree	<i>mtry</i>	OOB error
300	16	6.16%
	23	5.89%
	34	6.23%
	51	6.3%

correlation between predictors, and the highly correlated predictors in a pair receive smaller VIMs than the uncorrelated predictors. The correlation coefficient (*aka* Pearson’s product-moment correlation coefficient) between a pair of descriptors  $x$  and  $y$  is calculated using equation 1, where  $\bar{x}$  and  $\bar{y}$  represent the mean of  $x$  and  $y$ , respectively, and  $\sigma_x$  and  $\sigma_y$  represent the standard deviations of  $x$  and  $y$ , respectively, and  $n$  is the number of molecules. Correlation plot of MDA- and MDG-ranked top-30 descriptors are shown in Figure 6. In order to avoid any bias in descriptor selection, only those top-ranked descriptors that have correlation coefficient value  $<0.75$  with other descriptors within the set are selected as features. Table 4 presents the number of features selected from each top- $p$  descriptors based on MDA

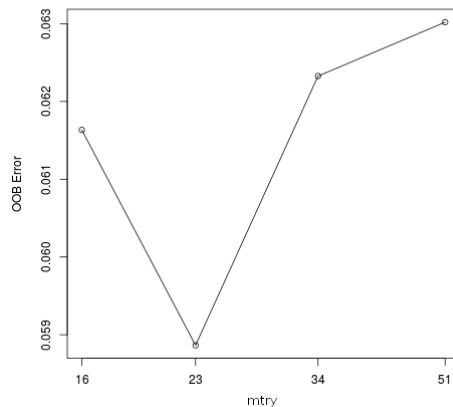


Fig. 5: Visualization of the OOB errors for different *mtry* values

and MDG values, where  $p$  is taken as 10, 20, 30, and 40. As a result, total eight sets of descriptors based on different cut-off values of  $p$  are selected for classification model learning and analysis. Files containing all 1171 descriptors along with their MDA and MDG values (mTORdescriptor\_mda.csv & mTORdescriptor\_mdg.csv) have been uploaded at [www.github.com/chetna-kumari/mTOR\\_Inhibitor](http://www.github.com/chetna-kumari/mTOR_Inhibitor).

$$\rho_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n-1)\sigma_x\sigma_y} \quad (1)$$

### 3.2 Dimension Reduction of Molecular Fingerprints

We have used simple architecture of deep neural network (DNN) based autoencoders to reduce the dimension of 166-bits MACCS FPs. Although DNN allows different number of neurons in each layer, we have considered same number of neurons (nodes) in each hidden layer of the DNN which limits the number of adjustable parameter combinations. In this study, the parameter settings for generating autoencoders are as follows:  $no.ofhiddenlayers(h) = 3$ ,  $no.ofnodes = 100$ ,  $no.ofepochs = 100$ ,  $costfunction(l1) = 0.00001$ ,  $activationfunction = tanh$ . We executed several autoencoders with different sets of neurons (assuming same in each layer), and evaluated their performance using four primary metrics – mean squared error (MSE), root-mean squared error (RMSE), mean absolute error (MAE), symmetric mean absolute percentage error (SMAPE), and one composite metric –  $R^2$  (coefficient of determination) [43]. The calculated values of MSE, RMSE, MAE, SMAPE,  $R^2$ , and number of nodes (#Nodes) using autoencoders are given in Table 5 and visualized in Figure 7. Comparatively lower values of MSE, RMSE, MAE and SMAPE, higher value of  $R^2$ , and higher number of nodes are preferred for dimension reduction using autoencoder. In this study, the final architecture of autoencoder is selected on the basis of MSE and RMSE values. It can be observed from Table 5 that the MSE and RMSE values decrease with increasing number of nodes. On analysis, we found that autoencoder with 100 nodes in each of the three layers achieves small MSE and RMSE values, and do not change significantly after further increasing the number of nodes. At this point, the MAE and SMAPE values are also reasonably small, whereas  $R^2$  value is considerably high (that is, 0.995). Therefore, we have chosen three hidden layers with 100 nodes in each layer in autoencoder architecture and selected the nodes from the middle layer.

### 3.3 Classification Model Learning and Evaluation

The classification models are trained using eight sets of descriptors, one set of autoencoder-reduced FPs, and a hybrid feature set, as shown in Table 4. The descriptors- and fingerprints-based models are trained using four classifiers – RF, SVM, DT, and NN, while the hybrid feature-based model is trained using the best performing classifier. In RF model, the OOB errors calculated during model training represent the performance over the test dataset. However, to make sure that all samples get a chance to be a part of the training as well as test, cross-validation is crucial

for model evaluation and comparison with other models. Hence, 5-fold cross-validation over the training dataset is used for model evaluation. The performance of the developed models are evaluated using various metrics – accuracy (ACC), sensitivity (SE), specificity (SP), F-measure (F1), Mathews correlation coefficient (MCC) and area under curve (AUC) values, such as Receiver Operating Characteristic AUC (ROC-AUC) and Precision Recall AUC (PR-AUC). The accuracy, sensitivity, specificity, F-measure, and Mathews correlation coefficient values are calculated using equations 2, 3, 4, 7, and 8 respectively. In these equations, TP (true positives) represents the number of active molecules that are classified as active, FP (false positives) represents the number of inactive molecules that are classified as active, FN (false negatives) represents the number of active molecules that are classified as inactive, and TN (true negatives) represents the number of inactive molecules that are classified as inactive. Table 6 presents the evaluation results of different classification models in terms of these metrics.

$$Accuracy(ACC) = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

$$Sensitivity(SE) = \frac{TP}{TP + FN} \quad (3)$$

$$Specificity(SP) = \frac{TN}{TN + FP} \quad (4)$$

$$Precision(\Pi) = \frac{TP}{TP + FP} \quad (5)$$

$$Recall(\rho) = \frac{TP}{TP + FN} \quad (6)$$

$$F\text{-measure}(F1) = \frac{2 \times \Pi \times \rho}{\Pi + \rho} \quad (7)$$

$$MCC = \frac{TN \times TP - FN \times FP}{(TN + FP) \times (FP + TP) \times (TP + FN) \times (FN + TN)} \quad (8)$$

Two common graphical representations to evaluate the performance of classifiers are ROC and PR curves, as shown in Figure 8. ROC curve represents the relationship between true positive rate (sensitivity) and false positive rate (1 - specificity), while the PR curve represents the relationship between precision (positive predictive value) and recall (true positive rate or sensitivity) at different cut-off values. PR-AUC value is preferred over ROC-AUC value, while MCC is most important measure for evaluation of binary classifier when the model training is done using class-imbalanced dataset [44], [45].

### 3.4 Comparative Analysis

In order to establish the efficacy of our proposed approach, we have compared and contrasted it with one of the existing state-of-the-art methods proposed by Wang et al. [22]. Wang et al. developed a series of *in silico* models using Recursive Partitioning (RP) and Naive Bayes (NB) algorithms to predict mTOR inhibitors. They considered thirteen molecular descriptors, atom centre fragments (ACFs) descriptor, and two types of fingerprints (FPs) namely extended-connectivity fingerprints and path-based fingerprints for model development, and evaluated the classification models in terms of various metrics, such as accuracy, sensitivity (recall), specificity, F-measure, AUC, and MCC values, and

TABLE 4: Feature sets used in this study to build classification models

Feature set id.	Feature category	Feature selection technique	#Top-ranked descriptors	#Selected features
$f_1$	Descriptors	MDA	10	5
$f_2$			20	10
$f_3$			30	16
$f_4$			40	20
$f_5$		MDG	10	5
$f_6$			20	9
$f_7$			30	11
$f_8$			40	17
$f_9$	Fingerprints (FPs)	Autoencoder		100
$f_{10}$	Hybrid (Descriptors + FPs)	MDA & Autoencoder	40	20+100

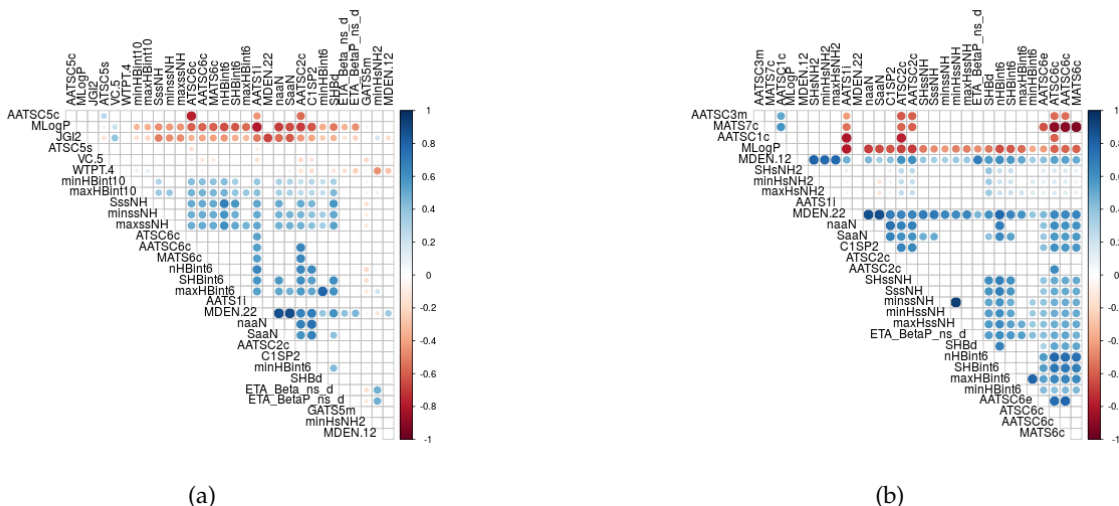


Fig. 6: Correlation plot of top-30 molecular descriptors ranked on the basis of (a) MDA, and (b) MDG values

TABLE 5: Dimension reduction of 166-bits MACCS FPs using autoencoders

#Nodes ( $h = 3$ )	MSE	RMSE	MAE	SMAPE	$R^2$
10, 10, 10	0.035	0.186	0.102	1.375	0.965
20, 20, 20	0.021	0.144	0.083	1.357	0.984
30, 30, 30	0.014	0.116	0.066	1.345	0.988
40, 40, 40	0.009	0.097	0.056	1.341	0.991
50, 50, 50	0.006	0.080	0.046	1.335	0.993
60, 60, 60	0.005	0.069	0.041	1.332	0.993
70, 70, 70	0.003	0.054	0.032	1.328	0.994
80, 80, 80	0.002	0.050	0.031	1.327	0.993
90, 90, 90	0.002	0.046	0.029	1.326	0.994
100, 100, 100	0.002	0.042	0.027	1.324	0.995
110, 110, 110	0.002	0.040	0.026	1.325	0.996
120, 120, 120	0.002	0.040	0.027	1.325	0.996
130, 130, 130	0.002	0.039	0.026	1.324	0.996

finally compared the performance on the basis of MCC values. MCC is the most important indicator for evaluating the classification models, when the training is done using class-imbalanced datasets. The classification model based on ACFs descriptor using an *in-house* program ACFs-NB shows best performance in terms of MCC value over validation set, and achieves the MCC value of 0.777.

In this study, we have developed various classification models based on the selected feature sets (that is, eight sets of descriptors, one type of fingerprints, and a hybrid feature set), and using four classifiers such as RF, SVM, DT and NN. The models are evaluated using several performance metrics, as shown in Table 6. However, the final comparison is based on MCC values, due to imbalanced training datasets. Our best performing model is based on a prioritized set of 20 descriptors (correlation coefficient  $< 0.75$ ) out of MDA-ranked top 40 and NN classifier, and achieves MCC value of 0.815 over validation set. Table 7 presents the comparative analysis results of our proposed model with Wang et al. [22] model. It can be observed from this table that our proposed model performs better to classify mTOR kinase inhibitor like compounds.

### 3.5 Best Classification Model Selection

Classification models based on the best performing set of twenty descriptors, reduced FPs, and hybrid feature set are compared for final model selection. A list of twenty best performing descriptors is shown in Table 8. It can be observed from this table that all twenty descriptors belong to nine different descriptor categories in which three categories are most prominent, as seven descriptors belong to the *atom type electrotopological state*, five belong to the *autocorrelation*, and

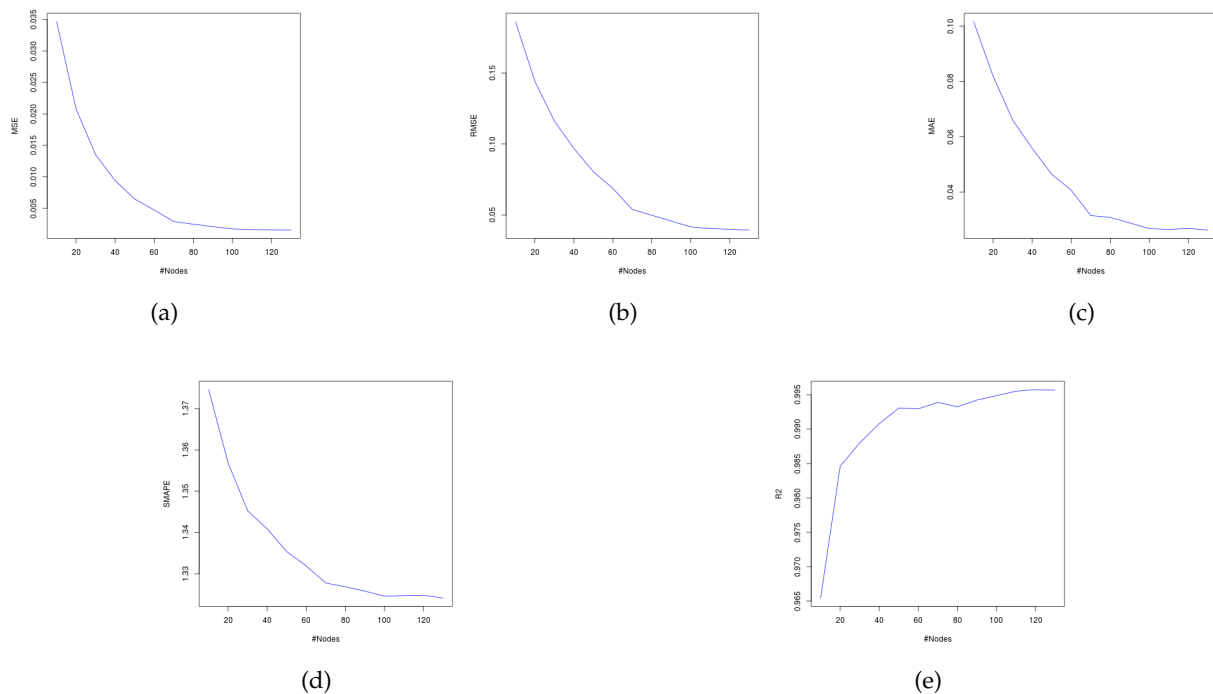


Fig. 7: Performance evaluation plot of autoencoder using (a) MSE, (b) RMSE, (c) MAE, (d) SMAPE, (e)  $R^2$ , and number of nodes (#Nodes)

two belong to *molecular distance edge* category. We observe that the classification models based on NN and twenty best performing descriptor set ( $f_4$ ), reduced FPs ( $f_9$ ), and the hybrid feature set ( $f_{10}$ ) achieve MCC values of 0.815, 0.752 and 0.731, respectively. The model based on  $f_4$  and NN also achieves reasonably good ROC-AUC and PR-AUC values that are 0.928 and 0.989 respectively, as shown in Figure 8. Therefore, we have selected best performing NN model for screening kinase bioactivity dataset.

### 3.6 *In silico* Screening of Kinase Bioactivity Datasets

The current release of ChEMBL (version 25) marks a rapidly growing public repository of approx. 1.9 million distinct compounds, more than 15 million bioactive drug-like molecules, approx. 1.1 million bioassays, and more than 12 thousand drug targets, 11 thousand drugs, and 72 thousand publications (<https://www.ebi.ac.uk/chembl/>) [46]. Kinase SARfari (current version 5.01), an offshoot of ChEMBL, is an integrated repository of chemogenomics data derived from high throughput screening, sequence, alignment, and structural information mainly focussed on kinases [47]. In this study, we have retrieved 40,922 compounds from kinase SARfari database which include bioactivity datasets of more than 500 kinases, to screen mTOR kinase inhibitor-like molecules. For screening, we require at least those features present in the data to be screened, based on which the classification models are trained. Therefore, out of total 40,922 compounds retrieved from the kinase SARfari database, only 26,421 compounds are considered for screening using the best classification model, and the screening result is shown in Table 10. It can be observed from this table that our predictive model is able to discriminate *active* and *inactive*

mTOR molecules, and this model can be used for predicting mTOR kinase inhibitor-like compounds in preliminary stages of the virtual screening.

## 4 DISCUSSION

mTOR is a well established pharmacological target due to its druggability and intervention in many human diseases including cancer. In this study, we have developed a series of *in silico* classification models to predict mTOR kinase inhibitors-like compounds with intent to therapeutically intervene autophagy pathway in cancer. The molecular descriptors and molecular fingerprints of mTOR bioactivity dataset from ChEMBL database are calculated using pADEL software. The complete dataset is divided into training set and validation set (4:1) by applying random sampling. The training dataset is used to build the classification models, which are internally validated using 5-fold cross validation. Validation set is used for external validation of the trained classification models. We have used RF classifier for two purposes – to calculate variable importance measures, and as a classifier. The molecular descriptors are selected based on *percentage of misclassification* and *decrease in gini index* values of variable importance measures using RF classification model. The selected molecular descriptors are further filtered on the basis of correlation coefficients, and deep neural network-based autoencoders are used to extract the molecular fingerprints. Simple architecture of autoencoder with three hidden layers and 100 neurons in each layer reduces the dimension of 166-bits MACCS fingerprints. Since no single feature is sufficient to discriminate inhibitors from non-inhibitors, different cut-off values of the number of top-ranked descriptors, reduced FPs and hybrid feature set



TABLE 6: Performance evaluation metrics of classifiers using prioritized feature sets

Classifier	Feature set id.	Training set (5-fold cv)						Validation set							
		ACC	SE	SP	F1	ROC-AUC	PR-AUC	MCC	ACC	SE	SP	F1	ROC-AUC	PR-AUC	MCC
RF	$f_1$	0.937	0.966	0.719	0.964	0.928	0.987	0.694	0.950	0.972	0.791	0.972	0.943	0.991	0.762
	$f_2$	0.940	0.964	0.760	0.966	0.918	0.987	0.715	0.952	0.972	0.814	0.973	0.942	0.991	0.778
	$f_3$	0.950	0.972	0.784	0.972	0.922	0.988	0.760	0.956	0.975	0.814	0.975	0.907	0.985	0.789
	$f_4$	0.948	0.969	0.790	0.9705	0.914	0.986	0.753	0.950	0.975	0.767	0.972	0.933	0.989	0.758
	$f_5$	0.936	0.962	0.737	0.963	0.922	0.989	0.694	0.953	0.978	0.767	0.973	0.9373	0.991	0.769
	$f_6$	0.943	0.969	0.754	0.968	0.930	0.989	0.727	0.956	0.978	0.791	0.975	0.937	0.990	0.785
	$f_7$	0.945	0.972	0.737	0.969	0.916	0.987	0.728	0.953	0.975	0.791	0.973	0.918	0.987	0.773
	$f_8$	0.940	0.970	0.719	0.966	0.910	0.984	0.708	0.953	0.978	0.767	0.973	0.923	0.988	0.769
	$f_9$	0.938	0.976	0.655	0.965	0.938	0.991	0.685	0.953	0.981	0.744	0.964	0.904	0.983	0.765
SVM	$f_1$	0.936	0.967	0.702	0.964	0.932	0.989	0.685	0.939	0.972	0.698	0.966	0.913	0.986	0.698
	$f_2$	0.941	0.972	0.7135	0.967	0.917	0.986	0.709	0.928	0.968	0.628	0.959	0.936	0.990	0.637
	$f_3$	0.9398	0.977	0.661	0.966	0.914	0.986	0.692	0.947	0.978	0.721	0.970	0.936	0.990	0.738
	$f_4$	0.943	0.9725	0.725	0.968	0.914	0.986	0.720	0.953	0.978	0.767	0.973	0.933	0.989	0.769
	$f_5$	0.940	0.967	0.743	0.966	0.926	0.988	0.713	0.939	0.972	0.698	0.966	0.9399	0.991	0.698
	$f_6$	0.939	0.964	0.754	0.965	0.932	0.989	0.711	0.936	0.968	0.698	0.964	0.913	0.986	0.687
	$f_7$	0.938	0.967	0.719	0.965	0.914	0.986	0.697	0.942	0.972	0.721	0.967	0.903	0.982	0.715
	$f_8$	0.938	0.971	0.696	0.965	0.895	0.982	0.694	0.931	0.965	0.674	0.961	0.933	0.990	0.660
	$f_9$	0.940	0.970	0.714	0.966	0.913	0.986	0.704	0.956	0.978	0.791	0.975	0.928	0.988	0.783
DT	$f_1$	0.941	0.969	0.731	0.967	0.926	0.989	0.713	0.936	0.965	0.721	0.964	0.929	0.990	0.693
	$f_2$	0.929	0.958	0.714	0.960	0.940	0.991	0.662	0.939	0.959	0.791	0.965	0.936	0.9905	0.722
	$f_3$	0.934	0.961	0.731	0.962	0.925	0.988	0.685	0.953	0.972	0.814	0.973	0.948	0.992	0.778
	$f_4$	0.927	0.958	0.696	0.958	0.919	0.987	0.650	0.947	0.965	0.814	0.970	0.916	0.986	0.757
	$f_5$	0.938	0.966	0.731	0.965	0.916	0.987	0.702	0.936	0.965	0.721	0.964	0.950	0.992	0.693
	$f_6$	0.931	0.955	0.754	0.961	0.940	0.991	0.684	0.936	0.965	0.721	0.964	0.936	0.991	0.693
	$f_7$	0.934	0.956	0.772	0.962	0.925	0.988	0.699	0.958	0.987	0.744	0.977	0.930	0.987	0.791
	$f_8$	0.931	0.954	0.754	0.961	0.928	0.988	0.682	0.936	0.965	0.721	0.964	0.899	0.982	0.693
	$f_9$	0.931	0.969	0.643	0.961	0.914	0.986	0.651	0.936	0.972	0.674	0.964	0.886	0.982	0.682
NN	$f_1$	0.938	0.961	0.772	0.965	0.928	0.989	0.713	0.942	0.968	0.744	0.967	0.946	0.992	0.720
	$f_2$	0.941	0.967	0.748	0.967	0.928	0.989	0.717	0.950	0.972	0.791	0.972	0.933	0.990	0.762
	$f_3$	0.948	0.972	0.772	0.971	0.932	0.990	0.749	0.953	0.978	0.767	0.973	0.892	0.983	0.769
	$f_4$	0.949	0.974	0.760	0.971	0.926	0.989	0.750	0.961	0.978	0.837	0.978	0.928	0.989	0.815
	$f_5$	0.944	0.967	0.772	0.968	0.912	0.986	0.733	0.950	0.972	0.791	0.972	0.918	0.988	0.762
	$f_6$	0.940	0.966	0.743	0.966	0.928	0.989	0.711	0.950	0.968	0.814	0.972	0.933	0.990	0.767
	$f_7$	0.937	0.963	0.743	0.964	0.931	0.989	0.700	0.936	0.965	0.721	0.964	0.920	0.987	0.693
	$f_8$	0.940	0.966	0.743	0.967	0.926	0.989	0.711	0.939	0.975	0.674	0.966	0.903	0.983	0.693
	$f_9$	0.940	0.968	0.737	0.966	0.931	0.989	0.712	0.947	0.968	0.791	0.970	0.943	0.992	0.752

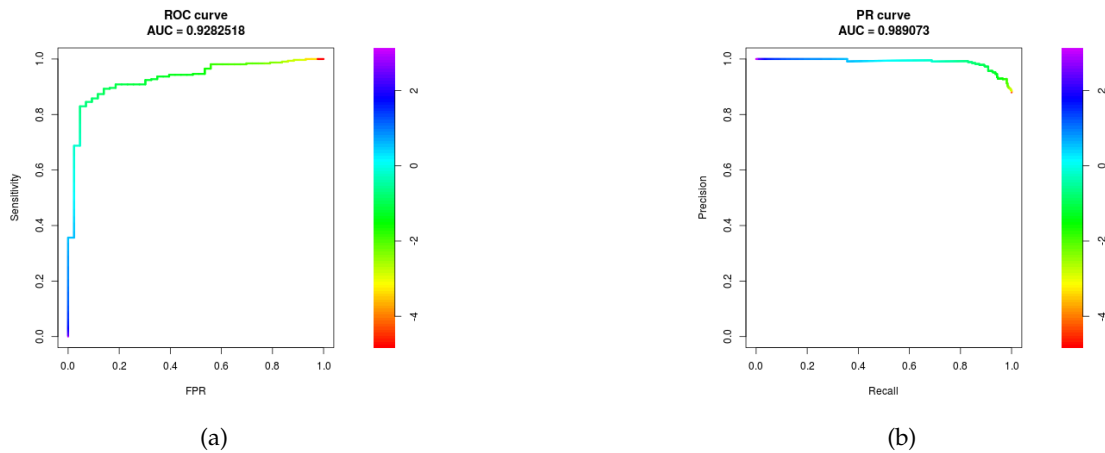


Fig. 8: (a) ROC and (b) PR curves with corresponding AUC values of NN classifier showing best performance over the validation dataset.

TABLE 7: Comparison of *in silico* models predicting mTOR inhibitors

Criteria	Wang et al., 2014	Proposed work
<b>(i) Dataset</b>		
a. Source	ChEMBL & BindingDB	ChEMBL
b. Class distribution	Imbalanced	Imbalanced
c. Balancing technique	not used	not used
<b>(ii) Features extraction</b>		
a. Tool used	Discovery Studio	PaDEL
b. #Descriptors	13+1 (ACF)	1171
c. #Fingerprints (FPs) type	2	1
<b>(iii) Features selection</b>		
a. Descriptors	not used	RF-VIM
b. FPs	not used	Autoencoder
<b>(iii) Classifiers selection</b>		
	RP, NB, ACF-NB	RF, SVM, DT, NN
<b>(iv) Best model Selection</b>		
a. Classifier	ACF-NB	NN
b. Feature set	ACF	$f_4$
c. Feature selection	not used	MDA
d. Performance over validation set		
Accuracy	92%	96.1%
Sensitivity	92.4%	97.8%
Specificity	90.3%	83.7%
Precision	97.3%	97.8%
PR-AUC	not calculated	0.989
<b>MCC</b>	<b>0.777</b>	<b>0.815</b>

TABLE 8: List of best performing twenty descriptors

Descriptor type	Descriptor	Rank	MDA
Atom type electrotopological state	maxHBin6	7	4.158
	SHBd	11	4.031
	maxssNH	20	3.405
	maxHBin10	24	3.272
	maxHBa	32	3.088
	minHBin5	33	3.057
Autocorrelation	minaaN	38	3.004
	MATS6c	15	3.813
	GATS5m	18	3.518
	ATSC5s	26	3.208
	AATSC2c	30	3.157
Molecular distance edge	AATSC3m	31	3.151
	MDEN.12	1	5.263
Extended topochemical atom	MDEN.22	3	4.529
	ETA_BetaP_ns_d	6	4.448
Topological charge	JGI2	9	4.117
Mannhold LogP	MlogP	16	3.616
Carbon types	C1SP2	25	3.249
Chi cluster	VC.5	27	3.179
Weighted path	WIPT.4	28	3.176

(combined set of best performing descriptors and reduced FPs) are used to build several classification models using four different classifiers – RF, SVM, DT, and NN. Performance measures are calculated using accuracy, sensitivity, specificity, F-measure, AUC (ROC-AUC and PR-AUC), and MCC values.

We observe that the classification models based on descriptor sets perform better than reduced MACCS fingerprints-based models in most of the cases. On evaluating the performance using MCC over the validation dataset, the model based on prioritized set of 20 descriptors (out of MDA-ranked top-40) and NN classifier outperform other models. Finally, we have selected this model to screen a compound dataset from kinase SARfari database, which is a specialized repository of bioactivity data for highly

conserved catalytic site in protein kinases. The predictive model is biased towards prediction of *active* class, yet capable of discriminating *active* and *inactive* molecules of mTOR bioactivity dataset. The reason for this biasness may be the disproportionate distribution of *active* and *inactive* classes in ChEMBL datasets as well as the compounds extracted from kinase SARfari database. These datasets are mainly extracted from literatures, and are more biased towards active compounds. This biasness may be due to more incentive given for publishing active compounds than publishing inactive molecules, which are mostly discarded. However, for building the predictive models, information on inactive molecules are equally important. The bioactivity dataset of mTOR in ChEMBL faces the same problem, and shows under-representation of *inactive* class in training dataset. As a result, the predictive models based on such skewed datasets are biased towards predicting majority class instances. There are various ways to deal with the class-imbalance problem in bioactivity data, such as using undersampling, oversampling, or using suitable performance metrics like PR-AUC and MCC values for evaluation of the models. Synthetic Minor Oversampling Technique (SMOTE) [48] is widely used method for balancing the class-imbalanced datasets. In another study reported in [49], we have discussed the handling of class-imbalance problem in mTOR bioactivity dataset using SMOTE. The main findings of this study can be summarized as follows:

- Two best performing models using MCC values – one over the validation dataset (MCC=0.815) and another over the training dataset (MCC=0.760), use MDA for descriptor selection. Hence, MDA may be preferred over MDG for prioritization and selection of descriptors.
- The best performing classifier based on the MCC value over the validation dataset is NN, which uses the prioritized set of 20 descriptors, out of the top-40 descriptors ranked by MDA.
- Three most important descriptor categories for predicting mTOR kinase inhibitor-like compounds are *atom type electrotopological state*, *autocorrelation*, and *molecular distance edge*.
- Screening of compound datasets from kinase SARfari database shows that our classification model is capable of discriminating *active* and *inactive* molecules of mTOR. Hence, this model seems useful for predicting mTOR kinase inhibitor-like compounds in preliminary stages of the virtual screening.

## 5 CONCLUSION AND FUTURE WORK

Machine learning-based compound classification is one of the ligand-based virtual screening approach, which is considered in this study to predict mTOR kinase inhibitor-like compounds from ChEMBL database. mTOR, a serine-threonine kinase, has implications in several diseases including cancer. Based on the existing literatures, it is found that mTOR has dual role in cancer through the process of autophagy. Though virtual screening methods (either structure- or ligand-based) cannot be solely applied to design a new drug, their application in the preliminary stages of the drug development process dramatically reduces the

TABLE 9: Performance evaluation metrics of NN classifier based on best performing descriptor set ( $f_4$ ), reduced FPs ( $f_9$ ) and hybrid feature set ( $f_{10}$ )

Feature set id.	Training set (5-fold cv)							Validation set						
	ACC	SE	SP	F1	ROC-AUC	PR-AUC	MCC	ACC	SE	SP	F1	ROC-AUC	PR-AUC	MCC
$f_4$	0.949	0.974	0.760	0.971	0.926	0.989	0.750	0.961	0.978	0.837	0.978	0.928	0.989	0.815
$f_9$	0.940	0.968	0.737	0.966	0.931	0.989	0.712	0.947	0.9685	0.791	0.970	0.943	0.992	0.752
$f_{10}$	0.945	0.969	0.760	0.969	0.931	0.989	0.733	0.944	0.972	0.744	0.9685	0.943	0.992	0.731

TABLE 10: Screening results of kinase SARfari compounds using the best performing classification model

<b>Classifier</b>	NN
<b>Feature set id.</b>	$f_4$
<b>MCC over validation set</b>	0.815
<b>#compounds extracted</b>	40922
<b>#compounds selected</b>	26421
<b>#Active molecules</b>	23719
<b># Inactive molecules</b>	2702

time and cost investment to complete a drug development cycle. On analysis, it is found that the classification models based on prioritized set of molecular descriptors generally outperform MACCS FPs-based models. We have shown that the classification model based on MDA-ranked descriptors shows best performance based on MCC values over the training and validation datasets. Hence, MDA may be preferred over MDG for descriptor selection and prioritization using mTOR bioactivity dataset. Although the model based on hybrid feature set under-perform prioritized descriptor set and fingerprints-based models using NN classifier, the fact cannot be generalized for all the classifiers used in this study, and there is a scope to use different combinations of molecular descriptors and FPs for evaluating hybrid features-based models using different classifiers.

The limitations of this study include the choice of the dataset which should ideally be stage-specific for the study of autophagy regulating kinases, as autophagy plays dual role in cancer (suppresses tumor in initial stage of tumorigenesis while promotes tumor in matured tumor cells). Handling class-imbalance problem in datasets, detailed evaluation of the hybrid feature-based models, and including more autophagy regulating kinases as targets to build a common model for screening large compound library seem promising future directions of research.

## REFERENCES

- [1] R. Santos, O. Ursu, A. Gaulton, A. P. Bento, R. S. Donadi, C. G. Bologa, A. Karlsson, B. Al-Lazikani, A. Hersey, T. I. Oprea, and J. P. Overington, "A comprehensive map of molecular drug targets," *Nature Reviews Drug discovery*, vol. 16, no. 1, p. 19, 2017.
- [2] R. T. Abraham, "Pi 3-kinase related kinases: 'big' players in stress-induced signaling pathways," *DNA Repair*, vol. 3, no. 8, pp. 883–887, 2004.
- [3] M. Laplante and D. M. Sabatini, "mTOR signaling in growth control and disease," *Cell*, vol. 149, no. 2, pp. 274–293, 2012.
- [4] H. Yang, D. G. Rudge, J. D. Koos, B. Vaidialingam, H. J. Yang, and N. P. Pavletich, "mTOR kinase structure, mechanism and regulation," *Nature*, vol. 497, no. 7448, pp. 217–223, 2013.
- [5] W. L. DeLano *et al.*, "Pymol: An open-source molecular graphics tool," *CCP4 Newsletter on protein crystallography*, vol. 40, no. 1, pp. 82–92, 2002.
- [6] R. A. Saxton and D. M. Sabatini, "mTOR signaling in growth, metabolism, and disease," *Cell*, vol. 168, no. 6, pp. 960–976, 2017.
- [7] P. F. Oliveira, C. Cheng, and M. G. Alves, "Emerging role for mammalian target of rapamycin in male fertility," *Trends in Endocrinology & Metabolism*, vol. 28, no. 3, pp. 165–167, 2017.
- [8] D. J. Klionsky and S. D. Emr, "Autophagy as a regulated pathway of cellular degradation," *Science*, vol. 290, no. 5497, pp. 1717–1721, 2000.
- [9] P. Sini, D. James, C. Chresta, and S. Guichard, "Simultaneous inhibition of mTORC1 and mTORC2 by mTOR kinase inhibitor azd8055 induces autophagy and cell death in cancer cells," *Autophagy*, vol. 6, no. 4, pp. 553–554, 2010.
- [10] Y. C. Kim and K.-L. Guan, "mTOR: A pharmacologic target for autophagy regulation," *The Journal of Clinical Investigation*, vol. 125, no. 1, pp. 25–32, 2015.
- [11] X. Li, H.-I. Xu, Y.-x. Liu, N. An, S. Zhao, and J.-k. Bao, "Autophagy modulation as a target for anticancer drug discovery," *Acta Pharmaceutologica Sinica*, vol. 34, no. 5, pp. 612–624, 2013.
- [12] E. Rad, J. Murray, and A. Tee, "Oncogenic signalling through mechanistic target of rapamycin (mTOR): A driver of metabolic transformation and cancer progression," *Cancers*, vol. 10, no. 1, p. 5, 2018.
- [13] V. S. Rodrik-Outmezguine, M. Okaniwa, Z. Yao, C. J. Novotny, C. McWhirter, A. Banaji, H. Won, W. Wong, M. Berger, E. de Stanchina *et al.*, "Overcoming mTOR resistance mutations with a new-generation mTOR inhibitor," *Nature*, vol. 534, no. 7606, p. 272, 2016.
- [14] F. Chiarini, C. Evangelisti, J. A. McCubrey, and A. M. Martelli, "Current treatment strategies for inhibiting mTOR in cancer," *Trends in Pharmacological Sciences*, vol. 36, no. 2, pp. 124–135, 2015.
- [15] K. G. Pike, K. Malagu, M. G. Hummersone, K. A. Menear, H. M. Duggan, S. Gomez, N. M. Martin, L. Ruston, S. L. Pass, and M. Pass, "Optimization of potent and selective dual mTORC1 and mTORC2 inhibitors: the discovery of azd8055 and azd2014," *Bioorganic & Medicinal Chemistry Letters*, vol. 23, no. 5, pp. 1212–1216, 2013.
- [16] S. V. Bhagwat, P. C. Gokhale, A. P. Crew, A. Cooke, Y. Yao, C. Mantis, J. Kahler, J. Workman, M. Bittner, L. Dudkin, D. M. Epstein, N. W. Gibson, R. Wild, L. D. Arnold, P. J. Houghton, and J. A. Pachter, "Preclinical characterization of OSI-027, a potent and selective inhibitor of mTORC1 and mTORC2: Distinct from rapamycin," *Molecular Cancer Therapeutics*, vol. 10, no. 8, pp. 1394–1406, 2011.
- [17] D. S. Mortensen, S. M. Perrin-Ninkovic, G. Shevlin, J. Zhao, G. Packard, S. Bahmanyar, M. Correa, J. Elsner, R. Harris, B. G. Lee *et al.*, "Discovery of mammalian target of rapamycin (mTOR) kinase inhibitor CC-223," *Journal of Medicinal Chemistry*, vol. 58, no. 13, pp. 5323–5333, 2015.
- [18] J. M. García-Martínez, J. Moran, R. G. Clarke, A. Gray, S. C. Cosulich, C. M. Chresta, and D. R. Alessi, "KU-0063794 is a specific inhibitor of the mammalian target of rapamycin (mTOR)," *Biochemical Journal*, vol. 421, no. 1, pp. 29–42, 2009.
- [19] E. K. Slotkin, P. P. Patwardhan, S. D. Vasudeva, E. de Stanchina, W. D. Tap, and G. K. Schwartz, "MLN0128, an ATP-competitive mTOR kinase inhibitor with potent in vitro and in vivo antitumor activity, as potential therapy for bone and soft-tissue sarcoma," *Molecular Cancer Therapeutics*, vol. 14, no. 2, pp. 395–406, 2015.
- [20] Q. Liu, C. Xu, S. Kirubakaran, X. Zhang, W. Hur, Y. Liu, N. P. Kwiatkowski, J. Wang, K. D. Westover, P. Gao *et al.*, "Characterization of torin2, an ATP-competitive inhibitor of mTOR, ATM, and ATR," *Cancer Research*, vol. 73, no. 8, pp. 2574–2586, 2013.
- [21] Y. Luo and L. Wang, "Discovery and development of ATP-competitive mTOR inhibitors using computational approaches," *Current Pharmaceutical Design*, vol. 23, no. 29, pp. 4321–4331, 2017.

- [22] L. Wang, L. Chen, Z. Liu, M. Zheng, Q. Gu, and J. Xu, "Predicting mtor inhibitors with a classifier using recursive partitioning and naïve bayesian approaches," *PLoS One*, vol. 9, no. 5, p. e95221, 2014.
- [23] L. Wang, L. Chen, M. Yu, L.-H. Xu, B. Cheng, Y.-S. Lin, Q. Gu, X.-H. He, and J. Xu, "Discovering new mtor inhibitors for cancer treatment through virtual screening methods and in vitro assays," *Scientific Reports*, vol. 6, p. 18987, 2016.
- [24] G. Cano, J. Garcia-Rodriguez, A. Garcia-Garcia, H. Perez-Sanchez, J. A. Benediktsson, A. Thapa, and A. Barr, "Automatic selection of molecular descriptors using random forest: Application to drug discovery," *Expert Systems with Applications*, vol. 72, pp. 151–159, 2017.
- [25] Z. Cheng, S. Zhou, Y. Wang, H. Liu, J. Guan, and Y.-P. P. Chen, "Effectively identifying compound-protein interactions by learning from positive and unlabeled examples," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2016.
- [26] E. Gawehn, J. A. Hiss, and G. Schneider, "Deep learning in drug discovery," *Molecular Informatics*, vol. 35, no. 1, pp. 3–14, 2016.
- [27] L. Zhang, J. Tan, D. Han, and H. Zhu, "From machine learning to deep learning: Progress in machine intelligence for rational drug discovery," *Drug Discovery Today*, 2017.
- [28] J. Ma, R. P. Sheridan, A. Liaw, G. E. Dahl, and V. Svetnik, "Deep neural nets as a method for quantitative structure–activity relationships," *Journal of Chemical Information and Modeling*, vol. 55, no. 2, pp. 263–274, 2015.
- [29] M. Liang, Z. Li, T. Chen, and J. Zeng, "Integrative data analysis of multi-platform cancer data with a multimodal deep learning approach," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 12, no. 4, pp. 928–937, 2015.
- [30] M. P. Menden, D. Wang, M. J. Mason, B. Szalai, K. C. Bulusu, Y. Guan, T. Yu, J. Kang, M. Jeon, R. Wolfinger *et al.*, "Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen," *Nature communications*, vol. 10, no. 1, p. 2674, 2019.
- [31] N. Chen and V. Karantz, "Autophagy as a therapeutic target in cancer," *Cancer biology & therapy*, vol. 11, no. 2, pp. 157–168, 2011.
- [32] A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani *et al.*, "ChEMBL: A large-scale bioactivity database for drug discovery," *Nucleic Acids Research*, vol. 40, no. D1, pp. D1100–D1107, 2012.
- [33] C. W. Yap, "Padel-descriptor: An open source software to calculate molecular descriptors and fingerprints," *Journal of Computational Chemistry*, vol. 32, no. 7, pp. 1466–1474, 2011.
- [34] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edition, ser. Springer series in statistics. Springer, 2009.
- [35] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.
- [36] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [37] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.
- [38] S. Haykin, "Neural networks a comprehensive foundation: Prentice hall international," *Inc., Englewood Cliffs*, 1999.
- [39] A.-L. Boulesteix, S. Janitza, J. Kruppa, and I. R. König, "Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics," *WIREs: Data Mining and Knowledge Discovery*, vol. 2, no. 6, pp. 493–507, 2012.
- [40] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston, "Random forest: A classification and regression tool for compound classification and qsar modeling," *Journal of Chemical Information and Computer Sciences*, vol. 43, no. 6, pp. 1947–1958, 2003.
- [41] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [42] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees*. CRC press, 1984.
- [43] A. Botchkarev, "A new typology design of performance metrics to measure errors in machine learning regression algorithms." *Interdisciplinary Journal of Information, Knowledge & Management*, vol. 14, 2019.
- [44] D. Chicco, "Ten quick tips for machine learning in computational biology," *BioData mining*, vol. 10, no. 1, p. 35, 2017.
- [45] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets," *PLoS one*, vol. 10, no. 3, p. e0118432, 2015.
- [46] A. Gaulton, A. Hersey, M. Nowotka, A. P. Bento, J. Chambers, D. Mendez, P. Mutowo, F. Atkinson, L. J. Bellis, E. Cibrián-Uhalte *et al.*, "The ChEMBL database in 2017," *Nucleic Acids Research*, vol. 45, no. D1, pp. D945–D954, 2016.
- [47] A. Bender, "Databases: Compound bioactivities go public," *Nature Chemical Biology*, vol. 6, no. 5, p. 309, 2010.
- [48] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.
- [49] C. Kumari, M. Abulaish, and N. Subbarao, "Using smote to deal with class-imbalance problem in bioactivity data to predict mtor inhibitors," in *Proceedings of the International Conference on Adaptive Computational Intelligence (ICACI), Mysuru, India, July 18-19, 2019*, pp. 1–12.



**Chetna Kumari** received M.Tech. degree in Computation and Systems Biology from Jawaharlal Nehru University (JNU), New Delhi, India. She is currently pursuing PhD in Computational Biology from the Department of Computer Science, Jamia Millia Islamia (A Central University), Delhi, India. She has qualified GATE and CSIR-UGC-NET exam in Life Sciences. Her research interests include Data-driven and Structure-based Drug Design, Biological Data Mining, and Machine Learning.



**Muhammad Abulaish** received PhD degree in Computer Science from Indian Institute of Technology (IIT) Delhi, India in 2007. He is currently an Associate Professor at the Department of Computer Science, South Asian University, Delhi, India. His research interests span over the areas of Data Analytics, Biological Data Mining, and Social Computing. He is a senior member of the IEEE, ACM, and CSI. He has published over 100 research papers in reputed journals and conference proceedings.



**Naidu Subbarao** received PhD in Chemistry from IIT Kanpur and Jawaharlal Nehru Centenary Common Wealth Postdoctoral fellow at Dept. of Biochemistry and Molecular Biology, University of Leeds, UK. He is currently an Associate Professor at the School of Computational and Integrative Sciences, JNU, New Delhi. His research focuses on Structural Bioinformatics and Molecular Recognition studies. His group has developed a pairwise and multiple structural alignment program and protein-protein docking algorithm using graph theoretic methods.