An Interpretable Machine Learning model for Selectivity of Small Molecules against Homologous Protein Family

Sarveswara Rao Vangala, Navneet Bung, Sowmya Ramaswamy Krishnan, Arijit Roy*

TCS Research (Life Sciences Division), Tata Consultancy Services Limited, Hyderabad, 500081, India.

*To whom correspondence should be addressed.

Abstract

Motivation: The primary goal of drug design is to develop potent small molecules that can inhibit the target protein with high selectivity. In the early stage of drug discovery, various experimental and computational methods are used to measure the target-specificity of small molecules against the target protein of interest. The selectivity of the small molecule remains a challenge, especially when the target protein belongs to a homologous family, which can often lead to off-target side effects.

Results: We have developed a multi-task deep learning model for predicting the selectivity of small molecules on the closely related homologs of the target protein. The multi-task model, which can learn from training data of the related tasks has been tested on the Janus kinase (JAK) and dopamine receptor family of proteins. To decipher the model decision on selectivity, the important fragments associated with each homolog protein were identified using SHapley Additive exPlanations (SHAP) method. The performance of the multi-task model was evaluated using various representation of small molecules such as fingerprints (ECFP4) and molecular graph representations. It was observed that the feature-based representation (ECFP4) with the XGBoost performed marginally better when compared to deep neural network models in most of the evaluation metrics. Both the models outperformed the graph-based models. The identification of important fragments associated with each proteins of the homolog family using SHAP method, explains the factors that governed the decision of the multi-task predictive model. The proposed method can be used post hit generation.

Contact: roy.arijit3@tcs.com

1 Introduction

One of the crucial steps for the success of drug discovery is to find a molecule that can bind to the target protein with high affinity and selectivity. The selectivity is often difficult to achieve, especially for the targets that belong to large families of structurally and/or functionally related proteins. Lack of selectivity can lead to off-target side effects, which is one of the reasons for the high attrition rate of drug molecules.

A majority of the current druggable targets in humans are confined to a few protein families. A study in 2017 identified 667 human proteins as druggable targets, among which 44% are from four homologous families alone (Santos et al., 2017). Examples of common druggable homologous protein families include protein kinases, ion channels, Rhodopsin-like G protein-coupled receptors (GPCRs), and nuclear hormone receptors (Santos et al., 2017). A more specific case is of the four kinases, JAK1, JAK2, JAK3 and tyrosine kinase 2 (TYK2), which form the Janus kinase (JAKs) family, and are centrally implicated in the cytokine receptor-mediated cell signaling process. Each of these druggable proteins play

different roles in cytokine-induced cell signaling (Dymock et al., 2013, Dymock et al., 2014) and therefore, selective inhibitors against individual proteins are now a key goal (Dymock et al., 2014). Several selective inhibitors against JAK1 (Norman et al., 2012), JAK2 (Dymock et al., 2014; Dymock et al., 2013), JAK3 (Pei et al., 2018), TYK2 (Norman et al., 2012) have been identified and used for treating specific diseases. For examples, the JAK2-specific inhibitors have been used to treat myeloproliferative neoplasms and are now being extended to treat leukemia, lymphoma and solid tumors (Dymock et al., 2013). The JAK3-specific inhibitors are used in immune-inflammatory diseases, such as rheumatoid arthritis and psoriasis (Pei et al., 2018). Similarly, the proteins of the dopamine receptor family have different functions and there are ongoing attempts to prepare selective inhibitors against the individual proteins (Keck et al., 2019; Mishra et al., 2018).

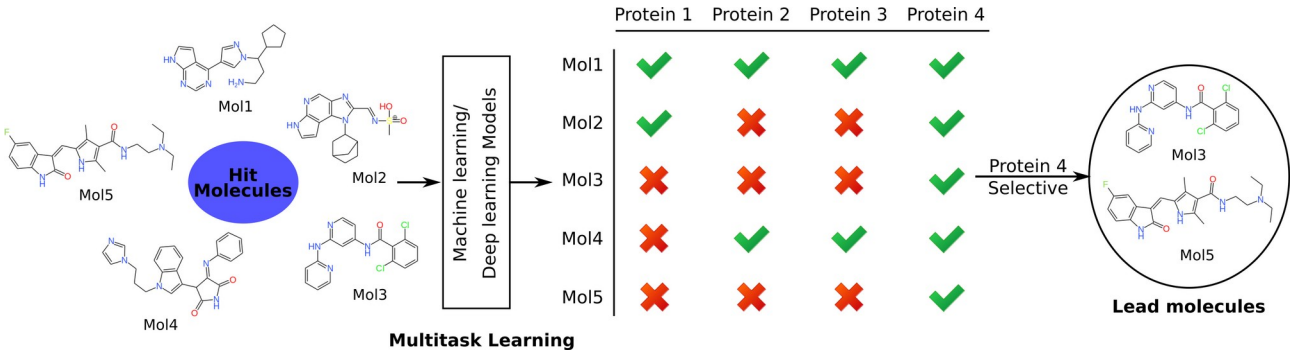


Figure 1. A multi-task machine learning model can screen the hit molecules to check the selectivity against the target protein from the structurally related family of proteins. In the above toy example, all the molecules were found to inhibit protein 4 during experiments, but only Mol2 and Mol5 were found to be selective. By monitoring the selectivity at an early stage, only selective molecules can be considered for further drug development.

Various *in silico* methods have been developed for improving the selectivity and has been extensively discussed in few review articles (Huggins et al., 2012; Chaudhari et al., 2020). There are attempts to develop databases of small molecules associated with their targets so that users can query about a new molecule based on structural similarity (Allaway et al., 2018; Chen et al., 2017). Peón et. al. has developed a webserver, MolTarPred, to predict the targets of a molecule (Peón et al., 2019). Similarly, structure-based approaches like docking or 3D-QSAR methods were also found to be useful for improving the selectivity (Huggins et al., 2012). There are also attempts to develop network-based approaches which can identify off-target effects (Moya-García and Ranea, 2013). While there have been various attempts to address the problem of selectivity, it still remains a challenge.

Recently, artificial intelligence has been used in various fields of science and technology including drug design and development (Schneider et al. 2020). Recent deep learning-based generative models (Krishnan et al., 2021a; Krishnan et al., 2021b; Bung et al., 2021; Olivecrona et al., 2017; Segler et al., 2018; Popova et al., 2018) have helped explore the vast chemical space while optimizing various physico-chemical properties. These methods have helped to drastically reduce the time required for hit identification (Zhavoronkov et al., 2019). However, none of the above approach addresses the challenge of selectivity during molecule generation.

The selectivity of small molecules in *in vitro* experiments is usually addressed by screening them against a subset of proteins, which are part of the same homologous family of the target protein. Kinase inhibitors are most often tested for selectivity due to the presence of a large number of kinases during drug discovery and development (Santos et al., 2017; Li et al., 2019). To mimic the experimental setup, various machine learning models can be trained to predict the effect of small molecules on the closely related homologs of the target protein which belong to the same family (Fig. 1). Multi-task learning models can be useful to address the question of selectivity since it can learn from the joint training signals of related tasks (selectivity towards multiple proteins) and generalize better than a single task (selectivity prediction for a single protein) (Caruana 1998). Multi-task learning has been used successfully applied to number of machine learning applications including drug discovery (Ramsundar 2015).

As a test case, we have trained multi-task predictive models on the Janus kinase (JAK) and dopamine receptor (DRD) family of proteins. Various small molecular input representations such as SMILES, ECFP4 fingerprint and molecular graph were tested to identify the most suitable representation for predicting target selectivity.

2 Methods

2.1 Dataset Curation

The dataset for human Janus kinase (JAK1, JAK2, JAK3 and TYK2) and dopamine receptor (DRD1, DRD2, DRD3, DRD4 and DRD5) family of proteins was curated using ExCAPE-DB (Sun et al., 2017) and ChEMBL (Gaulton et al., 2012), respectively. The dataset for each protein was downloaded and canonicalized using RDKit (https://www.rdkit.org) (Table 1). There were 481, 2165, 1674 and 908 molecules in the JAK1, JAK2, JAK3 and TYK2 specific datasets, respectively. For DRD1, DRD2, DRD3, DRD4 and DRD5 receptor there were 1072, 6498, 4385, 2248 and 308 molecules, respectively. The activity of all molecules was reported in pXC50, which is the half-maximal inhibitory concentration of molecules from various comparable methods and converted to negative log scale. Based on the pXC50 values, the molecules were classified as active (pXC50 >= 6) and inactive (pXC50 < 6). The four JAK family datasets were merged to obtain the curated multi-task dataset consisting of 2619 unique molecules, while for DRD family there were 8003 unique molecules.

Table 1. Details of the dataset used for modeling the selectivity of JAK and DRD family of proteins.

Janus Kinase Family			
Protein	Total Molecules	Actives	Inactives
JAK1	481	223	258
JAK2	2537	821	1716
JAK3	1492	226	1266
TYK2	722	68	654

Dopamine receptor family			
Protein	Total Molecules	Actives	Inactives
DRD1	1072	793	279
DRD2	6498	5184	1350
DRD3	4385	3837	548
DRD4	2248	1962	322
DRD5	308	186	122

2.2 Building the multi-task predictive models

The task in this study is to classify the small molecules as active or inactive against a family of homologous proteins. A multi-task predictive model can be ideal for the same, which can simultaneously predict the activity of a small molecule against a family of related proteins (Figure 1). Recent studies have shown that multi-task predictive models can outperform single-task models, as the hidden layers are shared among all tasks and helps the model to learn a task-agnostic representation (Rodríguez-Per´ez et al., 2019). Various machine learning models such as Extreme Gradient Boosting (XGBoost) (Chen et al., 2016), Deep Neural Networks (DNN) (Rodríguez-Per´ez et al., 2019) and graph-based models (GCN and GAT) (Kipf et al., 2017; Veličković et al. 2017) were trained to predict the selectivity of small molecules towards the proteins that belong to the same family. The input representation for the small molecule was chosen according to the algorithm used for the machine learning model, to harness the maximum possible chemical information. For the current study, two different input representations were explored: 1) Extended connectivity fingerprint (ECFP4) (Rogers et al., 2010) and 2) Molecular graph (Duvenaud et al., 2015). Based on the above input representations, five different predictive models were trained.

2.2.1 Extreme Gradient Boosting (XGBoost)

XGBoost (Chen et al., 2016) is an open-source implementation of the gradient boosted tree algorithm and has been widely used for prediction of several molecular properties (Leiet al., 2017; Yang et al., 2019; Jiang et al., 2021). However, there is no direct implementation of XGBoost that can perform multi-task output prediction. To mitigate this issue, a binary bit vector, with length equal to the number of targets considered for multi-task prediction and was concatenated with the ECFP4-based fingerprint of 1024 bits length (li et al.) (Fig. 1). For the current study, the length of input feature vector was considered as 1024+m, where m is the number of proteins in a family against which the selectivity needs to be checked. By appending the m-bit vector, the multi-task model was converted into a binary classification model, where the on bit corresponds to each of the protein/homolog being predicted (Rodríguez-Per´ez et al., 2019) (Fig. 1). The implementation of XGBoost from scikit-learn (Pedregosa et al. 2011) was used and extensive hyperparameter tuning was performed. During hyperparameter tuning, the parameters like learning rate (0.1, 0.01, 0.001), gamma (0.1, 0.2, 0.3, 0.5, 1, 2, 4, 8, 16,

32), max_depth (14-30) and n_estimators (from=5, to=100, step=5) were optimized using grid search.

2.2.2 Deep neural networks (DNN)

DNN algorithms have achieved excellent performance in several drug discovery problems (Goh et al., 2017; Hamadache et al., 2017). In the most simplistic model, a DNN consists of at least two hidden layers of neurons apart from the input and output layers (Rodríguez-Per´ez et al., 2019). The ECFP4-based fingerprint was used as an input to the first layer, and to the subsequent layer, the output from the previous layer was used as input. The final layer consists of m dimensions, where m corresponds to number of proteins in a family, against which selectivity was queried. For each of the intermediate layers the ReLU activation function was used, while the sigmoid activation function was used for the final layer. As the performance of a DNN is sensitive to hyperparameters, a grid search on the layer sizes (32, 64, 128, 256, 512), learning rate (0.01, 0.001, 0.005, 0.0001) and dropout rate (0.25, 0.5) were performed to find the best combination of hyperparameters.

2.2.3 Graph convolution network (GCN)

The GCN, which was originally introduced by Kipf and Welling. (Kipf et al., 2017), have shown promising results for predicting various molecular properties (Weider et al., 2020). A graph is usually defined as G=(V,E), where the atoms are represented as nodes (V) and the bonds between them as edges (E). A GCN with message passing layer transforms the embedding of each node in the following way: 1) aggregates the information from neighbouring nodes (or atoms) where it take help from an adjacency matrix $A \in \{0, 1\}_{n\times n}$ and a node feature matrix $X \in R_{n\times d}$. Here, n represents the number of nodes and d, the dimension of node feature vector (Buffelli et a., 2020). 2) Apply a non-linear activation function on the aggregated embedding (Buffelli et a., 2020). The GCN was implemented using the DeepChem library (Ramsundar et al., 2019). The default node and edge features were used to construct the graph. The learning rate (0.1, 0.01, 0.001), number of layers (1, 2) and size of layers (32, 64, 128, 256) were tuned during the training. The GCN layers were followed by a single dense layer of size 128 before the final output layer with sigmoid activation.

2.2.4 Graph Attention Networks (GAT)

Graph Attention Networks (GAT) use masked self-attention layers to provide improvement over the previous GCN methods (Veličković et al. 2017). The attention mechanism in GAT model can aggregate node information from neighbors effectively by assigning different importance to nodes of the same neighbourhood, enabling a leap in model capacity.

The GAT model consist of four steps: 1) Linear transformation: The input node features are transformed to output features using a learnable weight matrix W (eq. 1); 2) Computing attention coefficients: The pair-wise attention score between all neighbouring nodes in the graph are computed (eq. 2); 3) Normalization: The softmax function is applied over all the neighbouring nodes attention scores to get normalized scores (eq. 3); 4) Aggregation: In this final step, embeddings from the neighbouring nodes are multiplied with their respective attention score followed by aggregation to obtain the new node embedding (eq. 4). Apart from the hyperparameters used for GCN, an additional parameter, the number of attention heads (2, 4 and 8), was tuned while model training.

$$z_i^{(l)} = W^{(l)} h_i^{(l)} \quad (1)$$

$$e_{ij}^{(l)} = LeakyReLU\left(a^{|l|T}\left(z_i^{|l|} \vee ¿\, z_j^{(l)}\right)\right) \quad (2)$$

$$\alpha_{ij}^{(l)} = \frac{\exp\left(e_{ij}^{(l)}\right)}{\sum_{k \in N(i)} \exp\left(e_{ik}^{(l)}\right)} \quad (3)$$

$$h_i^{(l+1)} = \sigma\left(\sum_{j \in N(i)} \alpha_{ij}^{(l)} z_j^{(l)}\right) \quad (4)$$

2.2.5 MolMapNet

MolMapNet uses a convolutional neural network-based approach to incorporate 2D feature maps (MolMaps) based on molecular descriptors and fingerprints (Shen et al., 2021). Recently, this method has been shown to perform well compared to graph-based method against 26 pharmaceutically relevant benchmark datasets (Shen et al., 2021). The pre-trained model provided by Shen et al., was used to train the multi-task model to predict the selectivity of small molecules. The hyperparameters such as learning rate (0.01, 0.001, 0.0001), number of layers (1, 2) and size of layers (32, 64, 128) were tuned during model training.

2.3 Training and evaluation metrics

From the curated dataset, 80% of the data was used for training, while the remaining 20% for testing. For all the models mentioned above the binary cross-entropy loss (eq. 5) between ground truth values (y_{ij}) and predicted values (y_{ij_cap}) is calculated for each task and the combined loss is backpropagated to update the weights of neurons in each layer.

$$Loss = \frac{-1}{N} \sum_{i=1}^{M} \sum_{j=1}^{N} y_{ij} \log\left(\hat{y}_{ij}\right) + \left(1 - y_{ij}\right) \log\left(1 - \hat{y}_{ij}\right) \quad (5)$$

where, N is total number of samples in the dataset and M corresponds to the number of tasks.

The performance of the models was measured using the auROC score. Since the number of actives is less than the number of inactives in the curated dataset, the precision (eq. 6), recall (eq. 7) and f1-score (eq. 8) were also computed for the test and external datasets.

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

$$recall = \frac{TP}{TP + FN} \quad (7)$$

$$F1\,Score = \frac{2 * precision * recall}{precision + recall} \quad (8)$$

2.4 Interpreting the selectivity of JAK inhibitors

Given the state-of-the-art accuracy of the predictive models for the various properties, the need for interpreting such models is essential such that it can provide rationales behind the model decision. To address the issue of interpreting the machine learning models,

Lundberg et al. proposed a unified framework called SHAP (SHapley Additive exPlanations) which assigns every feature an importance score for all the predictions of the model (Lundberg et al., 2017).

In the current study, the predictions of the best performing method, XGBoost was interpreted for inhibitors of JAK and Dopamine Receptor family of proteins. To accomplish this, TreeExplainer (Lundberg et al., 2020) was used, which is a version of SHAP method designed for tree-based algorithms from the SHAP python package. Top few fragments ranked based on SHAP were further analyzed. For each of the top 10 ECFP4 fragments, the ratio of positives (R_p, equation 9) and negatives (R_n, equation 10) were computed to identify the substructures that are prominent in JAK family of proteins (Pope et al., 2018).

$$Ratio\ of\ positives\left(R_p\right)=\frac{N_a}{N_a+N_i} \tag{9}$$

$$Ratio\ of\ negatives\left(R_n\right)=\frac{N_i}{N_a+N_i} \tag{10}$$

Where N_a, N_i corresponds to number of times a particular substructure occurs in actives and inactive.

3 Results and Discussion

3.1 Modeling selectivity of JAK inhibitors

Various multi-task models were trained using different input representations and machine learning algorithms to predict the selectivity of small molecules among closely related homologs. After extensive hyperparameter tuning, the performance of the best models for JAK inhibitors has been summarized in Table 1. The XGBoost, DNN, GCN, GAT and MolMapNet models showed an auROC of 0.9223, 0.8893, 0.8869, 0.8625 and 0.9067 respectively (Table 1). Based on the auROC score, the XGBoost model performed slightly better than the other machine learning models. The DNN and MolMapNet models were close to the best performing XGBoost model. Apart from the auROC, the precision, recall, and f1-score were also calculated for each of the models (Table 1). High precision value indicates that the model is predicting minimum number of false positives. A comparison of precision values across the various models shows that XGBoost and MolMapNet models performed better than the rest of the models. Further, the recall metric indicates the fraction of active samples that are correctly classified. The XGBoost has a better recall with 0.6309 when compared to the other models. Overall, the XGBoost model performed better than the other models based on different metrics used to evaluate the performance of the multi-task predictive models.

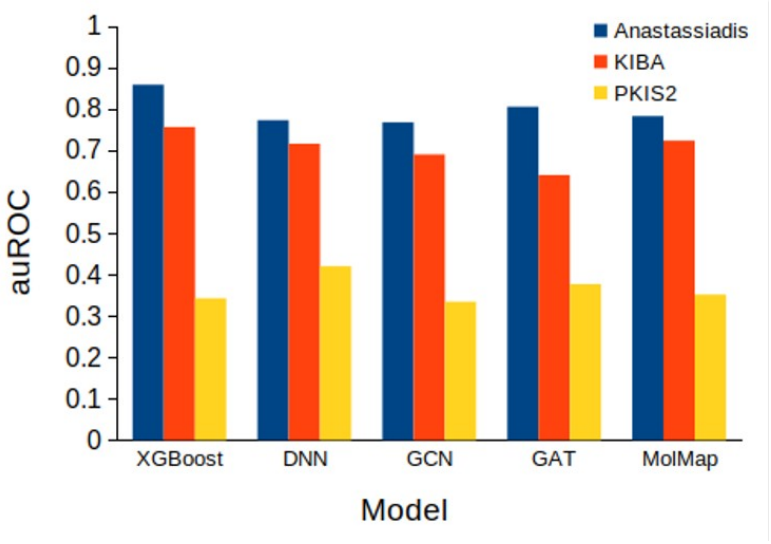
Table 2. Performance metric of five different machine learning algorithms for JAK proteins.

	Train	Test			
Algorithm	auROC	auROC	precision	recall	F1-score
XGBoost	0.9998	0.9233	0.8508	0.6309	0.7143
DNN	0.9798	0.8893	0.8348	0.5934	0.6809

GCN	0.9638	0.8869	0.7649	0.4599	0.5301
GAT	0.8902	0.8625	0.7673	0.4002	0.4671
MolMapNet	0.9298	0.9067	0.8305	0.5675	0.6422

Figure 2. Bar plot showing the performance of different multi-task predictive models on the three external datasets of the Janus kinase family.

3.2 Evaluation of models on external datasets

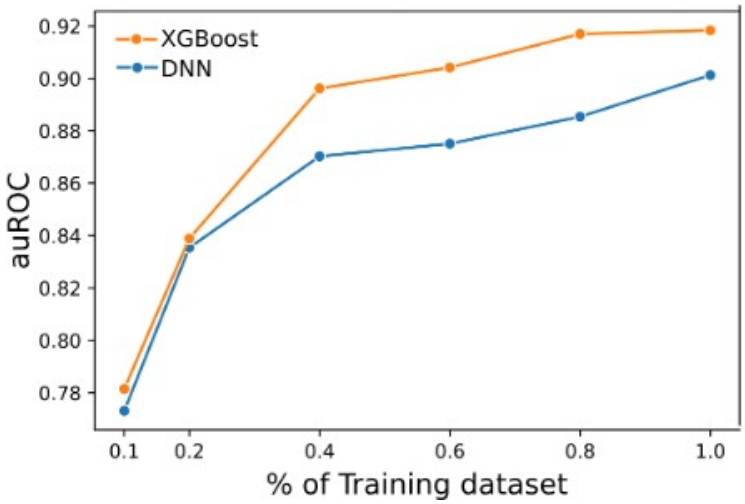


The performances of the predictive models were tested on three different external datasets, Anastassiadis (Anastassiadis et al., 2011), KIBA (Tang et al., 2014) and PKIS2 (Drewry et al., 2017). All the three external datasets consisted of datapoints for various kinases from which, only the datapoints corresponding to the Janus kinase family of proteins were extracted and converted to a classification task. The active and inactive molecules from KIBA dataset were identified through the pX50 values as mentioned in the Methods section. The Anastassiadis and PKIS2 dataset measured the activity of the protein at a fixed concentration of the small molecule. For the Anastassiadis and PKIS2 dataset, a small molecule was considered active, if the reported value of protein activity is less than 50%, else it was considered inactive. For all the three external datasets, any small molecule-protein pair that was present in the training dataset were removed. The final curated dataset consisted of 60, 292 and 599 molecules for Anastassiadis, KIBA and PKIS2 datasets, respectively. Based on the auROC values, the XGBoost model performed better for the Anastassiadis and KIBA datasets (Fig. 2). For the PKIS2 dataset the DNN model performed marginally better when compared to the XGBoost model. However, the performance of all the five multitask models on PIKS2 dataset is low when compared to other external and test dataset due to very less similarity of molecules when compared to training dataset (Li et al., 2019). The high performance of multi-task predictive models on external datasets, Anastassiadis and KIBA, further adds confidence to the predictions made by the machine learning models.

Figure 3. Performance of XGBoost and DNN models by varying the size of training dataset JAK family of proteins.

3.3 Performance of the model depends on the dataset size

From the above two case studies on the Janus kinase and DRD family of proteins, it was

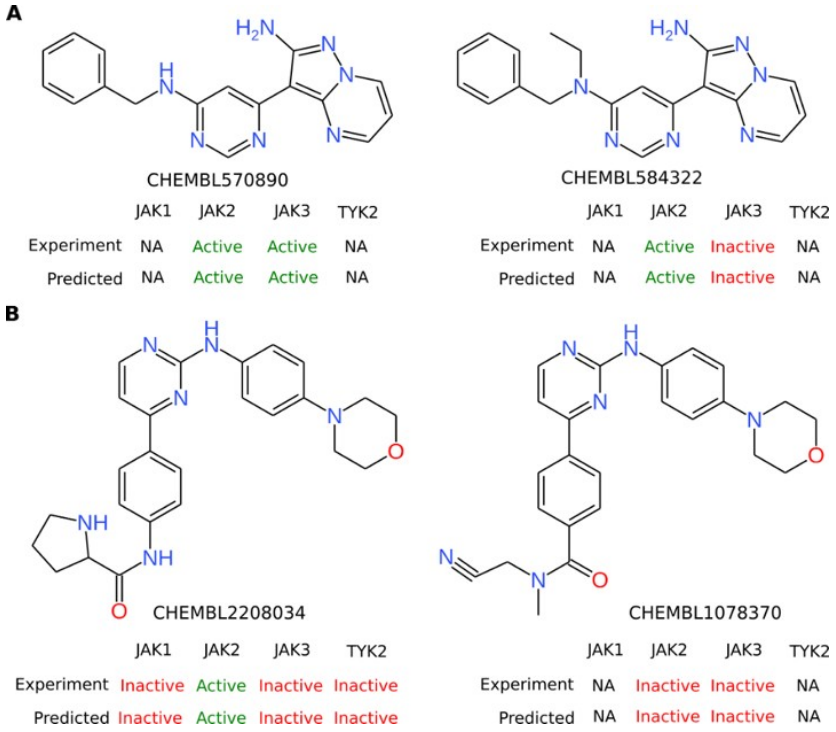


observed that both XGBoost and DNN models perform better when compared to the other deep learning models. Next, we examined the effect of training dataset size on the model performance. To evaluate the effect of training dataset only a fraction of the training dataset was randomly used to train the models. However, the test dataset was uniform across all the evaluations. The test dataset for JAK and DRD consists of 524 and 1600 molecules, respectively. While the size of complete training dataset for JAK and DRD was 2093 and 6402 molecules, respectively. With varying training data size, the auROC of JAK models ranged from 0.78 to 0.92 (Fig. 3). For JAK dataset, the XGBoost model performed better than the DNN model for dataset size larger than 0.4 of the complete training datasets (Fig. 3). Overall, with decreasing sizes of the training dataset, the performance decreased considerably for both the XGBoost and DNN models.

3.4 Model distinguishes structurally similar molecules

The predictions obtained from the XGBoost model were analyzed to check if the model was able to distinguish closely related molecules with common scaffold and correctly classify them into the respective classes. Two such representative pairs are discussed here. The first pair of molecules, CHEMBL584322 and CHEMBL570890 are similar in structure with substitution at one end. The substitution makes the CHEMBL584322 selective towards JAK2, while CHEMBL570890 is active against both JAK2 and JAK3. The current analysis correctly predicted the selectivity of the two molecules in accordance with the observed experimental values (Fig. 4a). Similarly, the model was able to distinguish between the molecules, CHEMBL2208034 and CHEMBL1078370, where a substitution at one site results in the molecule CHEMBL1078370 to be inactive (Fig. 4b). The ability of such selectivity prediction for molecules with common scaffold, but varying substituents (Fig. 4) demonstrates the usefulness of the machine learning models proposed in this work.

Figure 4. Selectivity prediction of structurally similar molecules in the test set and their validation from the experimental results. The active and inactive molecules against a protein are colored in green and red, respectively.



3.5 Modeling selectivity of dopamine receptor inhibitors

The method proposed in the current work was also used to model the selectivity of small molecules against the proteins of the dopamine receptor family. The dopamine receptors are a class of G protein-coupled receptors, mainly present in the central nervous system. They are responsible for various neurological processes such as pleasure, motivation, memory, cognition, learning and also control of fine motor skills (Girault et al. 2004). Each of the five dopamine receptors (DRD1. DRD2, DRD3, DRD4 and DRD5) has different function. Based on the auROC metric the XGBoost model performs better than other deep learning models, followed by DNN model (Table 2). The auROC of XGBoost (auROC – 0.8857) model is slightly better than that of the DNN-based model (auROC – 0.8729). While the XGBoost model performs better in the recall metric, the DNN model performs better in the precision metric for the same test set. However, the f1-score which is the harmonic mean of precision and recall, is similar for both the XGBoost and DNN models. Based on the results it can be inferred that the performance of both the XGBoost and DNN models are marginally better than the MolMapNet and graph-based models for the DRD family.

To our surprise, few of the simplest feature-based XGBoost and DNN models performed better in the classification task when compared to graph-based and other image-based convolution methods. A recent study (Jiang et al., 2021) corroborates well with the findings of the current work, where it was shown that the feature-based methods like XGBoost and random forest (RF) perform better when compared to graph-based methods on classification tasks (Jiang et al., 2021).

Table 3. Performance metric of five different machine learning algorithms for selectivity prediction of small molecules against DRD family of proteins.

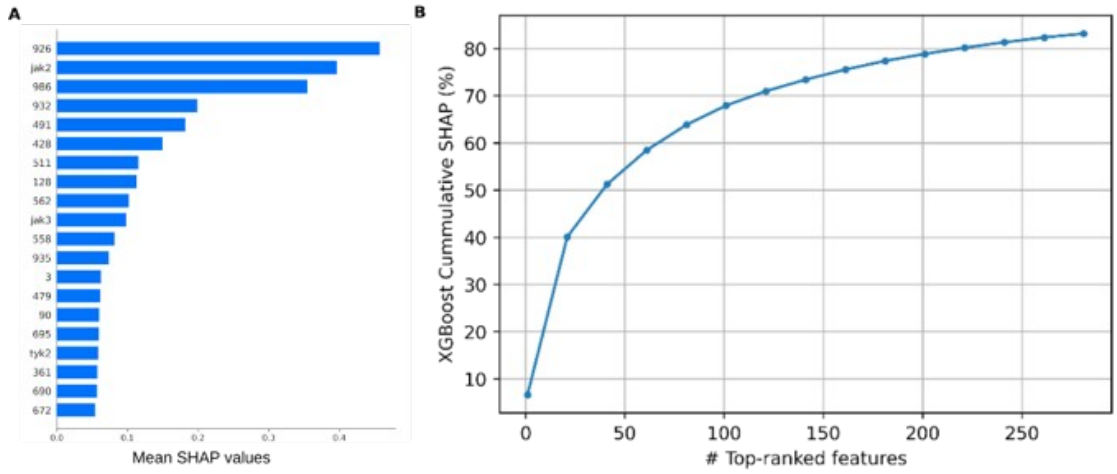
Algorithm	Train	Test			
	auROC	auROC	precision	recall	F1-score
XGBoost	0.9891	0.8857	0.8704	0.9674	0.9160
DNN	0.9692	0.8729	0.8951	0.9448	0.9193
GCN	0.9048	0.8189	0.8371	0.8443	0.8373
GAT	0.7908	0.7361	0.8052	0.9187	0.8558
MolMapNet	0.9624	0.8282	0.8359	0.9409	0.8849

3.6 Explainability of machine learning models

To further understand the features that render selectivity to both the JAK and DRD family of proteins, the SHAP scores were computed and analyzed (see Methods). Figure 5A shows the distribution of mean SHAP values for top 20 ECFP4 bits for JAK inhibitors. The bits corresponding to JAK2, JAK3 and TYK2 were indeed in the top 20 fragments ranked according to the SHAP values (Fig. 5A). These further provides a confidence that the model indeed looks at those bits during classification. To determine the number of features that contribute towards model prediction, the cumulative feature contributions were computed. Figure 5b shows the cumulative SHAP percentage values for top ranked features. While top 220 features contribute to 80% to the overall predictive performance of the model, around 600 features with low SHAP values contributed less than 0.01%. This indicates that the presence of these 600 features do not affect the performance of the model.

Figure 5: **A.** Mean SHAP values for top 20 features corresponding to JAK dataset obtained from XGBoost model. **B.** Distribution of cumulative SHAP percentage with respect to top ranked features.

To further analyse the results from the SHAP method, the substructures of the top 10 ECFP4 bits of all the four targets (JAK1, JAK2, JAK3, and TYK2) were analyzed. The ratio of positives (R_p) was calculated for each protein of the JAK family. A high R_p value for a fragment indicates that the fragment is preferred in the active molecules when compared



to inactive molecules, while a high R_n value would mean otherwise. As these ratios can be

sensitive to substructures whose count is less, a cut-off of 10 was considered (Pope et. al., 2018). The top 10 substructures from each of the target (after removing the redundant & merging the common substructures) are shown in Table 4. If the Rp score of a substructure is significantly high for a particular target protein, then the presence of it makes the small molecule more selective towards that protein. If the score is similar for more than one protein, then presence of it will make the molecule selective towards all those proteins. Few of the fragments like cnc(c(c)F)N(C)C (R_p= 0.947) and cnc(Nc)c(c)F (R_p= 0.906) are mostly found in small molecules against JAK2, while fragment ccc(c(c)n)c(n)[nH] was highly found in small molecules against JAK1 when compared to other homologes (Table 4, Fig. 6). Few fragments were found to be important for more than one protein, such as CC(C)C#N, which was equally observed in the small molecules that are active against JAK1 (R_p= 0.775), JAK2 (R_p= 0.679) and JAK3 (R_p= 0.725) proteins (Table 4). Surprisingly, we could not find any preferred fragment for the actives of TYK2 protein, at least from the top 10 ECFP4 bits. This could be due to poor data size for TYK2 protein (Table 1). A similar analysis was carried out for the homologs of the dopamine receptor family (see Supporting information fig. S1 and Table S2), where active and inactive fragments were identified for all the five homologs.

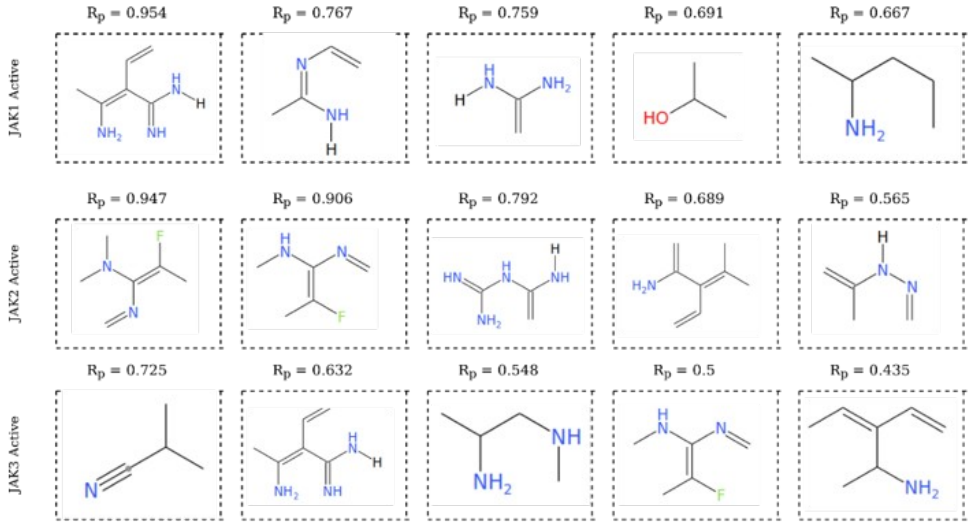


Figure 6. The prevalent fragments of JAK1, JAK2, and JAK3 actives obtained from top 10 ECFP4 bits. The ratio of positives (R_p) value for each of the fragments is also provided.

Table 4. R_p of Fragments that are preferred for active small molecules of JAK1, JAK2, JAK3 and TYK2.

Fragments	JAK1	JAK2	JAK3	TYK2
cnc(c(c)F)N(C)C	-	0.947	-	-
cnc(Nc)c(c)F	-	0.906	0.5	
cc([nH])Nc(n)n	-	0.792	-	-
ccc(c(c)n)c(n)[nH]	0.954	0.719	0.632	-
ccc(c(c)c)c(c)n	-	0.689	-	0.397
CC(C)C#N	0.775	0.679	0.725	-

cn[nH]c(c)C	-	0.565	-	-
cc(c)Nc(n)n	-	0.559	-	0.286
ccc(-c(c)[nH])c(c)N	-	0.55	-	-
cc(n)N(CC)CC	-	0.544	-	-
ccnc(c)[nH]	0.767	-	0.35	-
cc(n)[nH]	0.759	-	0.332	-
cc(n)[nH]	0.759	-	0.332	-
CC(C)O	0.691	-	-	-
cC(c)N	0.687	-	-	0.15
CCCC(C)N	0.667	-	0.375	0.28
ccnn(c)C	0.655	-	-	-
CNCC(C)N	-	-	0.548	-

As mentioned above, a high R_n score would mean that the fragment is dominantly present in the inactives of a given target protein. Figure 7 shows the distribution of fragments in the inactives, which have high SHAP value. The fragments obtained from the SHAP value and further ranked based on Rp score could be used to design selective small molecules against the target. Also, the presence of fragment predominantly in the inactives provide us knowledge on fragments that could be avoided during the design of small molecules.

4 Conclusion

Selectivity of small molecules against homologous protein family remains a challenging problem. This can lead to off-target side effect if the function of the homologous proteins is considerably different. In this work, five machine learning methods were used to identify the selectivity of small molecules. These models, XGBoost, DNN, GCN, GAT and MolMapNet were chosen based on their previous performance on various predictive tasks on biological data. Although, the performance of XGBoost and DNN models were comparable, overall, the XGBoost method performed better in terms of all the metrices. Both these models outperformed other graph-based models. As a case study, we used two well-known family of proteins, JAK and DRD receptors. In both the cases, a similar trend of model performance was observed. The rationales obtained from SHAP values explained the molecular fragments that are responsible for differentiating the affinity towards multiple proteins of the homologous protein family. While the current work can be used to screen molecules for selectivity before experimental testing, it can also be integrated with deep learning-based molecule generation models (Krishnan et al., 2021). The method proposed in this work can be extended to understand the selectivity of existing drug molecules against all druggable protein targets and identify the off-target side effects. Such a model can be potentially used for drug repurposing.

References

Allaway, R. J. *et al.* (2018). Probing the chemical–biological relationship space with the drug target explorer. *J. Cheminform.*, **10**(1), 1–14.

Anastassiadis, T. *et al*. (2011). Comprehensive assay of kinase catalytic activity reveals features of kinase inhibitor selectivity. *Nat. Biotechnol.*, **29**(11), 1039–1045.

Buffelli, D. and Vandin, F. (2020). A meta-learning approach for graph representation learning in multi-task settings. *arXiv preprint arXiv:2012.06755*.
Bung, N. *et al*. (2021). De novo design of new chemical entities for SARS-CoV-2 using artificial intelligence. *Future Med. Chem.*, **13**(06), 575–585.
Caruana, R. (1998). Multitask learning. autonomous agents and multi-agent systems. **27**(1), 95–133.
Chaudhari, R. *et al*. (2020). An up-to-date overview of computational polypharmacology in modern drug discovery. *Expert Opin. Drug Discov.,* **15**(9), 1025–1044.
Chen, C. *et al.* (2017). Mtld, a database of multiple target ligands, the updated version. *Molecules*, **22**(9), 1375.
Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledgediscovery and data mining*, pages 785–794.
Drewry, D. H. *et al*. (2017). Progress towards a public chemogenomic set for protein kinases and a call for contributions. *PloS one*, **12**(8), e0181585.
Duvenaud, D. *et al*. (2015). Convolutional networks on graphs for learning molecular fingerprints. *arXiv preprint arXiv:1509.09292*.
Dymock, B. W. *et al.* (2014). Selective jak inhibitors. *Future Med. Chem*., **6**(12), 1439–1471.
Dymock, B. W. and See, C. S. (2013). Inhibitors of jak2 and jak3: an update on the patent literature 2010–2012. *Expert Opin. Ther. Pat.*, **23**(4), 449–501.
Gaulton, A. *et al*. (2012). Chembl: a large-scale bioactivity database for drug discovery. *Nucleic acids research*, **40**(D1), D1100–D1107.
Girault, J. A. and Greengard, P. (2004). The neurobiology of dopamine signaling. *Arch Neurol*., **61**(5), 641–644.
Goh, G. B. *et al*. (2017). Chemception: a deep neural network with minimal chemistry knowledge matches the performance of expert-developed qsar/qspr models. *arXivpreprint arXiv:1706.06689.*
Hamadache, M. *et al.* (2017). Application of multilayer perceptron for prediction of the rat acute toxicity of insecticides. *Energy Procedia*, **139**, 37–42.
Huggins, D. J. *et al*. (2012). Rational approaches to improving selectivity in drug design. *J. Med. Chem*., **55**(4), 1424–1444.
Jiang, D. *et al*. (2021). Could graph neural networks learn better molecular representation for drug discovery? a comparison study of descriptor-based and graph-based models. *J. Cheminform*., **13**(1), 1–23.
Keck, T. M. *et al*. (2019). Dopamine d4 receptor-selective compounds reveal structure–activity relationships that engender agonist efficacy. *Journal of medicinal chemistry*, **62**(7), 3722–3740.
Kipf, T. N. and Welling, M. (2016). Semi-supervised classification with graphconvolutional networks. a*rXiv preprint arXiv:1609.02907*.
Krishnan, S. R. *et al*. (2021a). Accelerating de novo drug design against novel proteins using deep learning. *J. Chem. Inf. Model.*, **61**(2), 621–630.
Krishnan, S. R. *et al.* (2021b). De novo structure-based drug design using deep learning. *Journal of Chemical Information and Modeling*.

Lei, T. *et al*. (2017). Admet evaluation in drug discovery. 18. reliable prediction of chemical-induced urinary tract toxicity by boosting machine learning approaches. *Mol. Pharm*., **14**(11), 3935–3953.
Li, X. *et al*. (2019). Deep learning enhancing kinome-wide polypharmacology profiling: model construction and experiment validation. *J. Med. Chem.*, **63**(16), 8723–8737.
Lundberg, S. M. *et al*. (2020). From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, **2**(1), 56–67.
Lundberg, S. M. and Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, pages 4768–4777.
Mishra, A. *et al*. (2018). Physiological and functional basis of dopamine receptors and their role in neurogenesis: possible implication for parkinson's disease. *Journal of experimental neuroscience*, **12**, 1179069518779829.
Moya-García, A. A. and Ranea, J. A. (2013). Insights into polypharmacology from drug-domain associations. *Bioinformatics*, **29**(16), 1934–1937.
Norman, P. (2012). Selective jak1 inhibitor and selective tyk2 inhibitor patents. *Expert opinion on therapeutic patents*, **22**(10), 1233–1249.
Olivecrona, M. *et al*. (2017). Molecular de-novo design through deep reinforcement learning. *J. Cheminform*., **9**(1), 1–14.
Pedregosa, F. *et al.* (2011). Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, **12**, 2825–2830.
Pei, H. *et al.* (2018). Discovery of a highly selective jak3 inhibitor for the treatment of rheumatoid arthritis. *Scientific reports*, **8**(1), 1–11.
Peón, A. *et al*. (2019). Moltarpred: a web tool for comprehensive target prediction with reliability estimation. *Chem. Biol. Drug. Des.*, **94**(1), 1390–1401.
Pope, P. *et al*. (2018). Discovering molecular functional groups using graph convolutional neural networks. *arXiv preprint arXiv:1812.00265*.
Popova, M. *et al*. (2018). Deep reinforcement learning for de novo drug design. *SciAdv.*, **4**(7), eaap7885.
Ramsundar, B., Eastman, P., Walters, P., *et al.* (2019). Deep learning for the lifesciences: applying deep learning to genomics, microscopy, drug discovery, and more. " *O'Reilly Media, Inc.*".
Rodriguez-Perez, R. and Bajorath, J. (2019). Multitask machine learning for classifying highly and weakly potent kinase inhibitors. *ACS Omega*, **4**(2), 4367–4375.
Rogers, D. and Hahn, M. (2010). Extended-connectivity fingerprints. *J. Chem. Inf. Model*., **50**(5), 742–754.
Santos, R. *et al.* (2017). A comprehensive map of molecular drug targets. *Nat. Rev. Drug Discov*., **16**(1), 19–34.
Schneider, P. *et al*. (2020). Rethinking drug design in the artificial intelligence era. *Nat. Rev. Drug Discov.*, **19**(5), 353–364.
Segler, M. H. *et al*. (2018). Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Cent Sci.*, **4**(1), 120–131.
Shen, W. X. *et al*. (2021). Out-of-the-box deep learning prediction of pharmaceutical properties by broadly learned knowledge-based molecular representations. *Nat. Mach. Intell.,* **3**(4), 334–343.

Sun, J. *et al*. (2017). Excape-db: an integrated large scale dataset facilitating big data analysis in chemogenomics. *J. Cheminform*., **9**(1), 1–9.
Tang, J. *et al.* (2014). Making sense of large-scale kinase inhibitor bioactivity datasets: a comparative and integrative analysis. *J. Chem. Inf. Model.*, **54**(3), 735–743.
Veli ̌ckovi ́c, P. *et al.* (2017). Graph attention networks. *arXiv preprint arXiv:1710.10903.*
Wieder, O. *et al*. (2020). A compact review of molecular property prediction with graph neural networks. *Drug Discov. Today Technol*.
Yang, Z. Y. *et al*. (2019). Structural analysis and identification of colloidal aggregators in drug discovery. *J. Chem. Inf. Model*., **59**(9), 3714–3726.
Zhavoronkov, A. *et al.* (2019). Deep learning enables rapid identification of potent ddr1 kinase inhibitors. *Nat. Biotechnol*., **37**(9), 1038–1040.