




For reprint orders, please contact: reprints@future-science.com

An interpretable machine learning model for selectivity of small molecules against homologous protein family

Sarveswara Rao Vangala¹ , Navneet Bung¹ , Sowmya Ramaswamy Krishnan¹  & Arijit Roy*,¹ 

¹TCS Research (Life Sciences Division), Tata Consultancy Services Ltd, Hyderabad, 500081, India

*Author for correspondence: roy.arijit3@tcs.com

Aim: In the early stages of drug discovery, various experimental and computational methods are used to measure the specificity of small molecules against a target protein. The selectivity of small molecules remains a challenge leading to off-target side effects. **Methods:** We have developed a multitask deep learning model for predicting the selectivity on closely related homologs of the target protein. The model has been tested on the Janus-activated kinase and dopamine receptor families of proteins. **Results & conclusion:** The feature-based representation (extended connectivity fingerprint 4) with Extreme Gradient Boosting performed better when compared with deep neural network models in most of the evaluation metrics. Both the Extreme Gradient Boosting and deep neural network models outperformed the graph-based models. Furthermore, to decipher the model decision on selectivity, the important fragments associated with each homologous protein were identified.

First draft submitted: 17 April 2022; Accepted for publication: 26 August 2022; Published online: 28 September 2022

Keywords: explainable models • machine learning • multitask models • selectivity • SHAP values

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 [1]. Examples of common druggable homologous protein families include protein kinases, ion channels, rhodopsin-like G protein-coupled receptors and nuclear hormone receptors [1]. A more specific case is of the four kinases, Janus-activated kinase (JAK)1, JAK2, JAK3 and tyrosine kinase 2 (TYK2), which form the JAK 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 [2,3] and therefore, selective inhibitors against individual proteins are now a key goal [3]. Several selective inhibitors against JAK1 [4], JAK2 [2,3], JAK3 [5] and TYK2 [4] have been identified and used for treating specific diseases. For example, the JAK2-specific inhibitors which are used to treat myeloproliferative neoplasms are now being extended to treat leukemia, lymphoma and solid tumors [2]. The JAK3-specific inhibitors are used in immune-inflammatory diseases, such as rheumatoid arthritis and psoriasis [5]. Similarly, the proteins of the dopamine receptor (DRD) family have different functions and there are ongoing attempts to prepare selective inhibitors against the individual proteins [6,7].

Various *in silico* methods have been developed for improving the selectivity and have been extensively discussed in few review articles [8,9]. 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 [10,11]. Peón *et al.* [12] have developed a webserver, MolTarPred, to predict the targets of a molecule. Similarly, structure-based approaches like docking or 3D-Quantitative Structure-Activity Relationships methods have also been found to be useful for improving

newlands
press

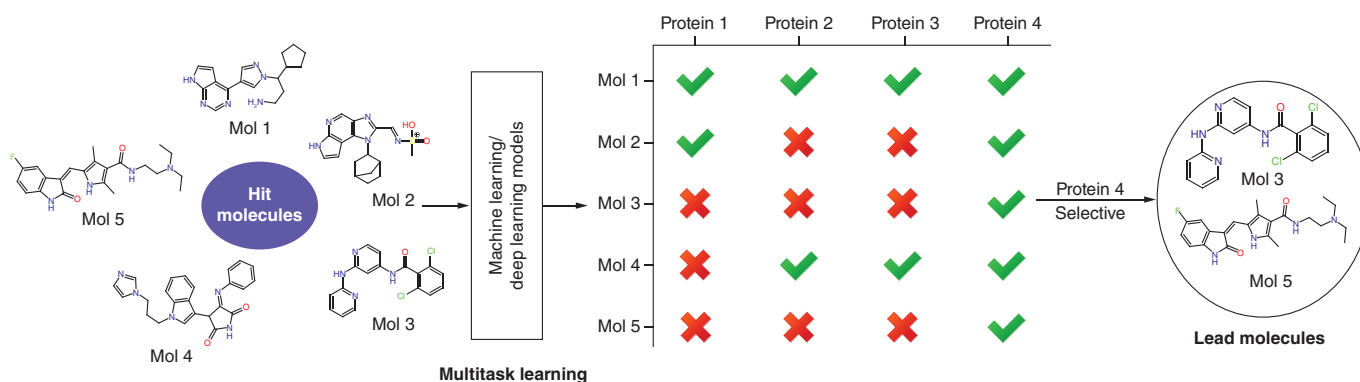


Figure 1. Modeling selectivity using multitask learning. The toy example shows that during experiments all the molecules were found to inhibit protein 4, but only Mol 3 and Mol 5 were found to be selective. A multitask machine learning model can be trained to screen the hit molecules to check the selectivity against the target protein from the structurally related family of proteins. By monitoring the selectivity at an early stage, only selective molecules can be considered for further drug development.

selectivity [8]. There are also attempts to develop network-based approaches which can identify off-target effects [13]. 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 [14]. Recent deep learning-based generative models [15–21] have helped explore the vast chemical space while optimizing various physicochemical properties. These methods have helped to drastically reduce the time required for hit identification [22]. However, none of the above approaches address 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 [1,23]. 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. Instead of building individual selectivity models against each homolog, the advantage of the similarities between individual tasks (selectivity prediction toward individual protein) is exploited using a multitask learning model (selectivity toward multiple proteins). Multitask learning models are useful to address the question of selectivity, since they can learn efficiently from the joint training signals of related tasks and increase the performance compared with individual single-task models [24]. Multitask learning has been successfully applied to a number of machine learning applications: for example, in the case of autonomous driving, a multitask model can detect stop signs, pedestrians and other obstacles together [25].

In the current study, we have trained multitask models to predict the selectivity of molecules against a homologous family of proteins (Figure 1). As a test case, we have trained multitask predictive models on the JAK and DRD families of proteins. Various small molecular input representations such as extended connectivity fingerprint 4 (ECFP4), molecular graph and a combination of molecular features and fingerprints were tested to identify the most suitable representation for predicting target selectivity.

Materials & methods

Dataset curation

The dataset for human JAK (JAK1, JAK2, JAK3 and TYK2) and DRD (DRD1, DRD2, DRD3, DRD4 and DRD5) families of proteins was curated using ExCAPE-DB [26] and ChEMBL [27], respectively. The dataset for each protein was canonicalized using RDKit (www.rdkit.org) (Table 1). There were 481, 2537, 1492 and 722 molecules in the JAK1-, JAK2-, JAK3- and TYK2-specific datasets, respectively. For DRD1, DRD2, DRD3, DRD4 and DRD5, there were 1072, 6498, 4385, 2248 and 308 molecules, respectively. The activity of all molecules was reported in the half-maximal inhibitory concentration of molecules from various comparable methods and converted to negative log scale (pXC_{50}). Based on the pXC_{50} values, the molecules were classified as active ($\text{pXC}_{50} \geq 7$) and inactive ($\text{pXC}_{50} < 7$). The four JAK family datasets were merged to obtain the curated multitask dataset consisting of 2619 unique molecules, while for DRD family there were 8003 unique molecules.

Table 1. Dataset used for modeling the selectivity of Janus-activated kinase and dopamine receptor families of proteins.

Janus-activated kinase family			
Protein	Total molecules	Actives	Inactives
JAK1	481	223	258
JAK2	2537	821	1716
JAK3	1492	226	1266
Tyrosine kinase 2	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

DRD: Dopamine receptor; JAK: Janus-activated kinase.

Building the multitask predictive models

The task in this study is to classify the small molecules as active or inactive against a family of homologous proteins. A multitask predictive model can be ideal for this and can simultaneously predict the activity of a small molecule against a family of related proteins (Figure 1). Recent studies have shown that multitask predictive models can outperform single-task models, as the hidden layers are shared among all tasks and help the model to learn a task-agnostic representation [28]. Various machine learning models such as Extreme Gradient Boosting (XGBoost) [29], deep neural networks (DNN) [28] and graph-based models such as graph convolution networks (GCNs) and graph attention networks (GATs) [30,31] were trained to predict the selectivity of small molecules toward 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: ECFP4 [32] and molecular graphs [33]. Based on the above input representations, five different predictive models were trained.

Extreme gradient boosting

XGBoost [29] is an open-source implementation of the gradient-boosted tree algorithm and has been widely used for prediction of several molecular properties [34–36]. However, there is no direct implementation of XGBoost that can perform multitask output prediction. To mitigate this issue, a binary bit vector, with length equal to the number of targets considered for multitask prediction and was concatenated with the ECFP4-based fingerprint of 1024 bits length (Supplementary Figure 1A) [23]. This approach was successfully applied to identify highly potent and weak compounds against kinase proteins [28]. 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 multitask model was converted into a multiclass predictive model, where the on bit corresponds to each of the proteins/homologs being predicted (Supplementary Figure 1A) [28]. The implementation of XGBoost from scikit-learn [37] was used and extensive hyperparameter tuning was performed. During hyperparameter tuning, the model parameters including 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.

Deep neural networks

DNN algorithms have achieved excellent performance in several drug discovery problems [38,39]. In the most simplistic model, a DNN consists of at least two hidden layers of neurons apart from the input and output layers [28]. 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 (Supplementary Figure 1B). For each of the intermediate layers the rectified linear activation unit 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. The multitasking described above (Supplementary Figure 1B) was used for subsequent deep learning methods.

Graph convolution network

The GCN, which was originally introduced by Kipf and Welling [30], has shown promising results for predicting various molecular properties [40]. 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: firstly, it aggregates the information from neighboring nodes (or atoms) where it takes help from an adjacency matrix $A \in \{0, 1\}_{n \times n}$ and a node feature matrix $X \in \mathbb{R}_{n \times d}$. Here, n represents the number of nodes, and d the dimension of node feature vector [41]. Secondly, a nonlinear activation function on the aggregated embedding is applied [41]. The GCN was implemented using the DeepChem library [42]. 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.

Graph attention networks

GATs use masked self-attention layers to provide improvement over the previous GCN methods [31]. The attention mechanism in a GAT model can aggregate node information from neighbors effectively by assigning different importance to nodes of the same neighborhood, enabling a leap in model capacity.

The GAT model consists of four steps: 1) linear transformation: the input node features are transformed to output features using a learnable weight matrix $W^{(l)}$ (Equation 1 & 2) computing attention coefficients: the pair-wise attention score between all neighboring nodes in the graph are computed (Equation 2 & 3) normalization: the softmax function is applied over all the neighboring nodes' attention scores to get normalized scores (Equation 3 & 4) aggregation: in this final step, embeddings from the neighboring nodes are multiplied with their respective attention score followed by aggregation to obtain the new node embedding (Equation 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.

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

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

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

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

In above the Equations 1–4, $h_i^{(l)} \in \mathbb{R}^F$, where $h_i^{(l)}$ represents hidden node features of i^{th} node in layer L , and F is the number of features in node; $z_i^{(l)}$ is the embedding of the i^{th} node in layer l ; $a^{(l)}$ is the learnable weight vector; $e_{ij}^{(l)}$ is the un-normalized attention score between i^{th} and j^{th} node; and $\alpha_{ij}^{(l)}$ is the normalized attention.

MolMapNet

MolMapNet uses a convolutional neural network-based approach to incorporate 2D feature maps (MolMaps) based on molecular descriptors and fingerprints [43]. It uses 1456 molecular descriptors and 16,204 fingerprint features of 8,506,205 molecules to learn the molecular representations [43]. Recently, this method was shown to perform well compared with graph-based methods against 26 pharmaceutically relevant benchmark datasets [43]. The pretrained model provided by Shen *et al.* [43] was used to train the multitask 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.

Training & evaluation metrics

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

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

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 area under the receiver operating characteristics score. Since the number of actives is less than the number of inactives in the curated dataset, the precision (Equation 6), recall (Equation 7), f1-score (Equation 8) and Matthews correlation coefficient (Equation 9) were also computed for the test and external datasets. The above metrics provided better understanding about the performance of the models.

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

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

$$F1\text{Score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (\text{Equation 8})$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP}) \times (\text{TP} + \text{FN}) \times (\text{TN} + \text{FP}) \times (\text{TN} + \text{FN})}} \quad (\text{Equation 9})$$

FN: False negatives; FP: False positives; MCC: Matthews correlation coefficient; TN: True negatives; TP: True positives.

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 and Lee proposed a unified framework called Shapley additive explanations (SHAP) which assigns an importance score to every feature for all the predictions of the model [44].

In the current study, the results were interpreted using the TreeExplainer [45], a version of the SHAP method designed for tree-based algorithms. The top few fragments ranked based on SHAP were further analyzed. For each of the top ten ECFP4 fragments, the ratio of positives (R_p , Equation 10) and negatives (R_n , Equation 11) were computed to identify the substructures that are prominent in the JAK family of proteins [46].

$$\text{Ratio of positives } (R_p) = \frac{N_a}{N_a + N_i} \quad (\text{Equation 10})$$

$$\text{Ratio of negatives } (R_n) = \frac{N_i}{N_a + N_i} \quad (\text{Equation 11})$$

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

Results & discussion

Modeling selectivity of JAK inhibitors

Various multitask 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 are summarized in Table 2. The hyperparameter values for

Table 2. Performance metric of five different machine learning algorithms for Janus-activated kinase proteins.

Algorithm	Train			Test		
	auROC	auROC	Precision	Recall	F1-score	Matthews correlation coefficient
Extreme Gradient Boosting	1.0	0.92	0.85	0.63	0.71	0.64
Deep neural network	0.98	0.89	0.83	0.59	0.68	0.67
Graph convolution network	0.96	0.89	0.76	0.46	0.53	0.58
Graph attention network	0.89	0.86	0.77	0.40	0.47	0.43
MolMapNet	0.93	0.91	0.83	0.57	0.64	0.43

auROC: Area under the receiver operating characteristics.

the best model are provided in supplementary data (Supplementary Table 1). The results with ECFP4 fingerprint (XGBoost and DNN), molecular graph (GCN and GAT) and a combination of molecular features and fingerprints (MolMapNet) are shown in Table 2. The XGBoost, DNN, GCN, GAT and MolMapNet models showed an auROC of 0.92, 0.89, 0.89, 0.86 and 0.91 respectively (Table 2). 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 2). High precision value indicates that the model is predicting the 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. The recall metric indicates the fraction of active samples that are correctly classified. The XGBoost model has a better recall with 0.63 when compared with the other models. Furthermore, a Matthews correlation coefficient value of 0.64 for the XGBoost model represents that both the active and inactive classes are predicted with reasonable accuracy. Overall, the XGBoost model performed better than the other models based on different metrics used to evaluate the performance of the multitask predictive models.

Evaluation of models on external datasets

The performance of predictive models were tested on three different external datasets, Anastassiadis [47], Kinase Inhibitor BioActivity (KIBA) [48] and PKIS2 [49]. All three external datasets consisted of datapoints for various kinases, from which only the datapoints corresponding to the JAK family of proteins were extracted and converted to a classification task. The active and inactive molecules from the KIBA dataset were identified through the pXC₅₀ values as mentioned in the *Materials & methods* section. The Anastassiadis and PKIS2 datasets measured the activity of the protein at a fixed concentration of the small molecule. For the Anastassiadis and PKIS2 datasets, a small molecule was considered active if the reported value of protein activity was 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 was 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 (auROC: 0.86) and KIBA (auROC: 0.76) datasets (Figure 2A). For the PKIS2 dataset, the DNN model (auROC: 0.42) performed better when compared with the XGBoost model (auROC: 0.34). However, the performance of all the five multitask models on the PKIS2 dataset is low when compared with other external and test datasets due to much less similarity of molecules when compared with the training dataset [23]. The high performance of multitask predictive models on external datasets, Anastassiadis and KIBA, further adds confidence to the predictions made by the machine learning models.

Performance of the model depends on the dataset size

From the above case study on the JAK family of proteins, it was observed that both XGBoost and DNN models perform better when compared with 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 consisted 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 (Figure 2B). For the JAK family of proteins, the XGBoost model performed better than the DNN model for dataset size larger than 0.4 of the complete training dataset (Figure 2B).

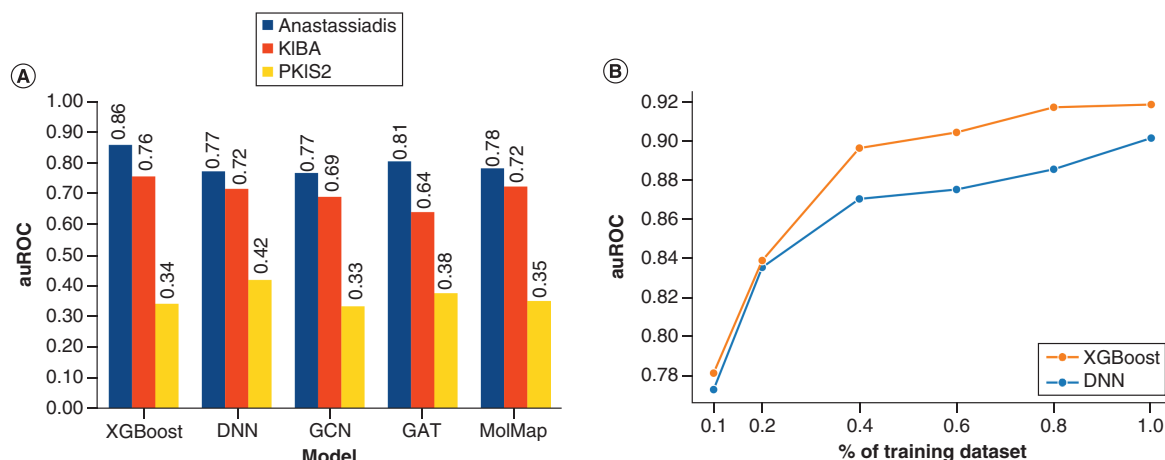


Figure 2. Performance on external dataset and effect of training dataset size. (A) Bar plot showing the performance of different multitask predictive models on three external datasets of the Janus kinase family. **(B)** Performance of XGBoost and DNN models by varying the size of training dataset Janus-activated kinase family of proteins. auROC: Area under the receiver operating characteristics; DNN: Deep neural network; GAT: Graph attention network; GCN: Graph convolution network; KIBA: Kinase Inhibitor BioActivity; XGBoost: Extreme Gradient Boosting.

Overall, with decreasing sizes of training dataset, the performance decreased considerably for both the XGBoost and DNN models.

Model distinguishes structurally similar molecules

The predictions obtained by 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 toward JAK2, while CHEMBL570890 is active against both JAK2 and JAK3. The current model correctly predicted the selectivity of the two molecules in accordance with the observed experimental values (Figure 3A). Similarly, the model was able to distinguish between the molecules CHEMBL2208034 and CHEMBL1078370, where a substitution at one site results in the molecule CHEMBL1078370 being inactive (Figure 3B). The ability of such accurate selectivity prediction for molecules with a common scaffold, but varying substituents (Figure 3), demonstrates the usefulness of the machine learning models proposed in this work.

Modeling selectivity of DRD inhibitors

The method proposed in the current work was also used to model the selectivity of small molecules against the proteins of the DRD family. The DRDs are a class of rhodopsin-like 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 [50]. Each of the five DRDs (DRD1, DRD2, DRD3, DRD4 and DRD5) has a different function. Based on the auROC metric, the XGBoost model performs better than the other deep learning models, followed by the DNN model (Table 3). The hyperparameter values for the best models are provided in supplementary data (Supplementary Table 2). The auROC of the XGBoost (auROC: 0.89) model is slightly better than that of the DNN-based model (auROC: 0.87). 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 with graph-based and other image-based convolution methods. A recent study [36] corroborates well with the findings of the current work, where it was shown that the feature-based methods like XGBoost and random forest perform better when compared with graph-based methods on classification tasks [36].

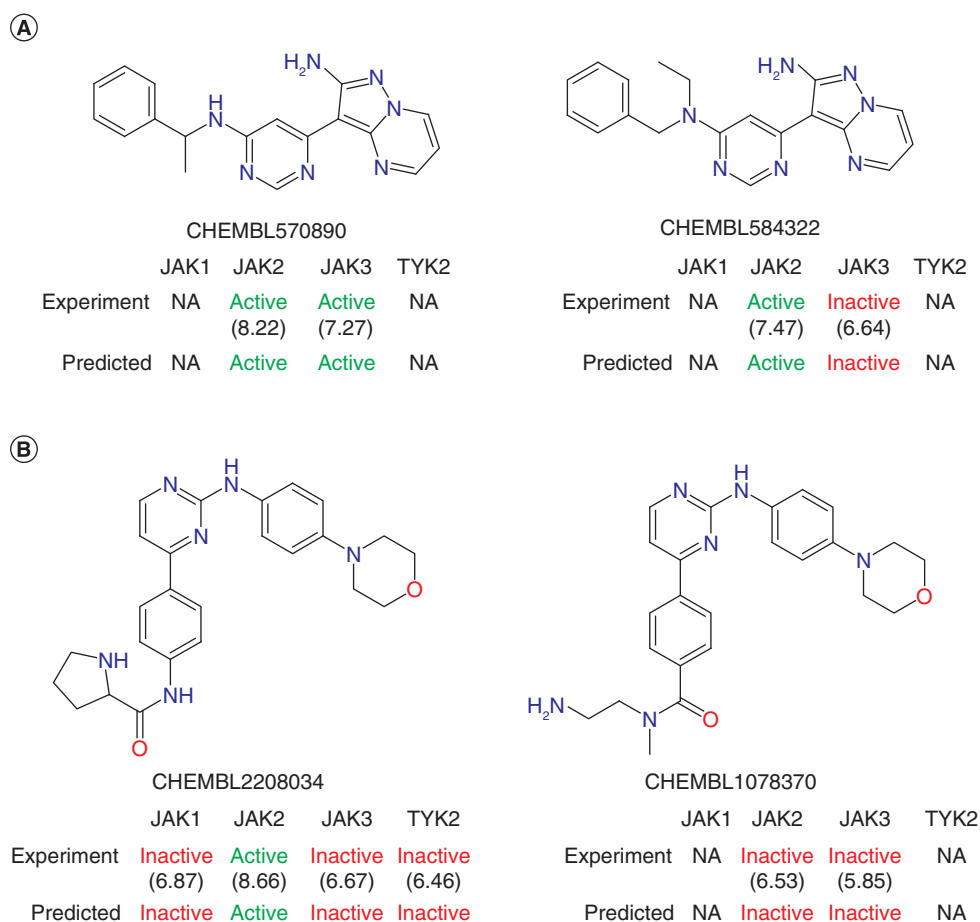


Figure 3. Validation of selectivity prediction on structurally similar molecules. (A) and (B) are examples of structurally similar molecules in the test set and their validation from the experimental results. The active and inactive molecules against a protein are colored in green and red, respectively. The experimental bioactivity data (pXC₅₀, half-maximal inhibitory concentration of molecules from various comparable methods and converted to negative log scale) is provided in brackets below each protein class (active/inactive). JAK: Janus-activated kinase; NA: Not applicable; TYK2: Tyrosine kinase 2.

Table 3. Performance metrics of five different machine learning algorithms for selectivity prediction of small molecules against dopamine receptor family of proteins.

Algorithm	Train				Test	
	auROC	auROC	Precision	Recall	F1-score	Matthews correlation coefficient
Extreme Gradient Boosting	0.99	0.89	0.87	0.97	0.92	0.51
Deep neural network	0.97	0.87	0.90	0.94	0.92	0.52
Graph convolution network	0.90	0.82	0.84	0.84	0.84	0.37
Graph attention network	0.79	0.74	0.81	0.91	0.86	0.27
MolMapNet	0.96	0.83	0.84	0.94	0.88	0.37

auROC: Area under the receiver operating characteristics.

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 Materials & methods section). Figure 4A shows the distribution of mean SHAP values of the 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 (Figure 4A). These further provide confidence that the model indeed looks at those bits during classification. To determine the number of features that maximally

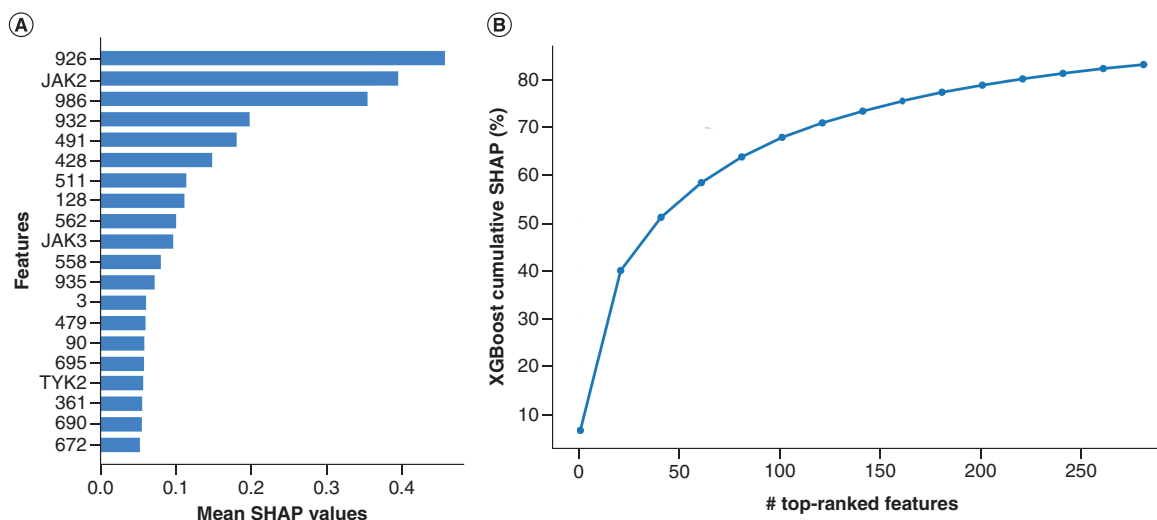


Figure 4. Top features and their contribution from SHAP analysis. (A) Mean SHAP values for top 20 features corresponding to JAK dataset obtained from XGBoost model. **(B)** Distribution of cumulative SHAP percentage with respect to top-ranked features.

JAK: Janus-activated kinase; SHAP: Shapley additive explanations; TYK2: Tyrosine kinase 2; XGBoost: Extreme Gradient Boosting.

contribute toward model prediction, the cumulative feature contributions were computed. Figure 4B shows the cumulative SHAP percentage values for top-ranked features. While the 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 does not affect the performance of the model.

To further analyze the results from the SHAP method, the substructures of the top ten 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 with inactive molecules (Supplementary Table 3), while a high R_n value would mean otherwise (Supplementary Table 4). As these ratios can be sensitive to substructures whose count is low, a minimum substructure incidence cutoff of 10 was considered [46]. The top ten substructures from each target (after removing redundancy and merging the common substructures) are shown in Supplementary Table 3. If the R_p score of a substructure is significantly high for a particular target protein, then the presence of the substructure makes the small molecule more selective toward that protein. If the score is similar for more than one protein, then presence of the substructure will make the molecule selective toward 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 with other homologs (Figure 5 & Supplementary Table 3). A comparison of predicted fragments and fragments present in highly selective JAK inhibitors is discussed in Supplementary Section 1. 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 (Supplementary Table 3). Surprisingly, we could not find any preferred fragment for the actives of TYK2 protein, at least from the top ten ECFP4 bits. This could be due to the low dataset size for TYK2 protein (Table 1). A similar analysis was carried out for the homologs of the DRD family (Supplementary Figure 3 & Supplementary Tables 5 & 6), where active and inactive fragments were identified for all the five homologs.

As mentioned above, a high R_n score would mean that the fragment is dominantly present in the inactives of a given target protein. Supplementary Figure 1 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 their R_p score could be used to design selective small molecules against the target. In addition, the presence of a fragment predominantly in the inactives provides knowledge on fragments that could be avoided during the design of small molecules.

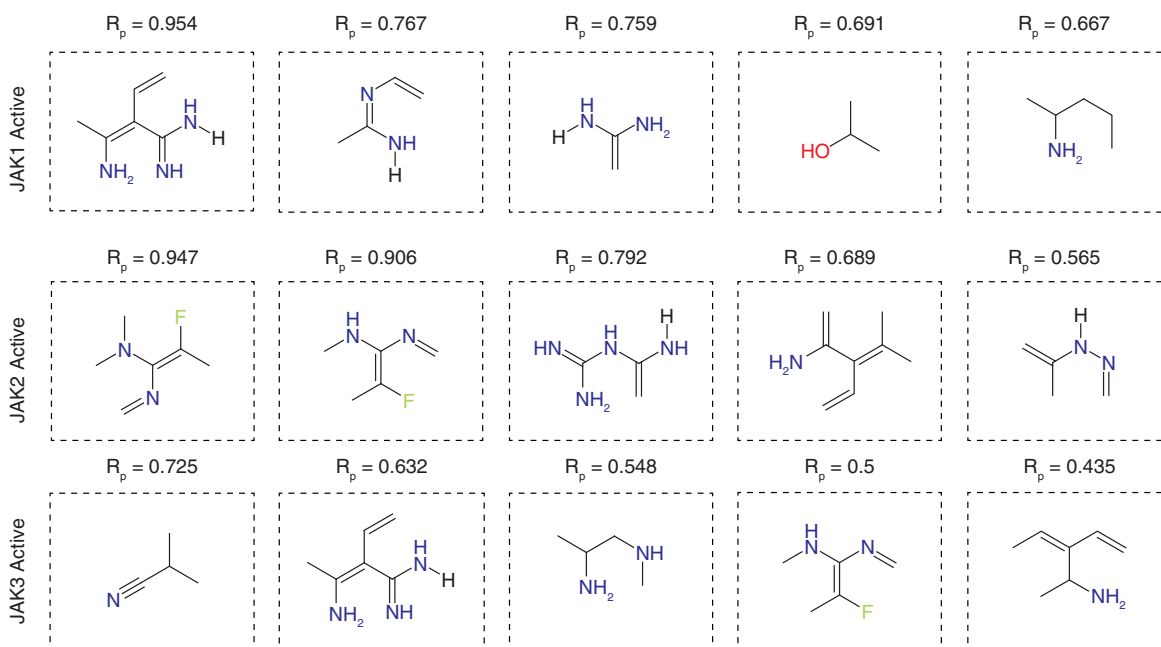


Figure 5. Top fragments responsible for selectivity. The prevalent fragments of JAK1, JAK2 and JAK3 actives obtained from top ten extended connectivity fingerprint 4 bits. The R_p value for each of the fragments is also provided.

JAK: Janus-activated kinase; R_p : Ratio of positives.

Conclusion

Selectivity of small molecules against a homologous protein family remains a challenging problem. This can lead to off-target side effects 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 metrics. Both models outperformed other graph-based models. As a case study, we used two well-known families of proteins, JAK and DRD. In both 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 toward 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. 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.

Summary points

- Selectivity of small molecules is a challenging problem which can often lead to off-target side effects.
- Extreme Gradient Boosting, deep neural network, graph convolution network, graph attention network and MolMapNet were chosen to predict selectivity of small molecules for Janus-activated kinase and dopamine receptors.
- Performance of Extreme Gradient Boosting and deep neural network models were comparable and both the models outperformed other graph-based models.
- The explainability of the models in terms of the molecular fragments that are responsible for differentiating the affinity toward multiple proteins of the homologous protein family was obtained using Shapley additive explanations.

Supplementary data

To view the supplementary data that accompany this paper please visit the journal website at: www.future-science.com/doi/suppl/10.4155/fmc-2022-0075

Financial & competing interests disclosure

All the authors are employed by Tata Consultancy Services Ltd. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Data & software availability

All input data are publicly available and a detailed description for the same is mentioned in the Materials & methods section. The code used to generate results shown in this study is available from the corresponding author upon request for academic use only.

References

Papers of special note have been highlighted as: • of interest; •• of considerable interest

- Santos R, Ursu O, Gaulton A *et al.* A comprehensive map of molecular drug targets. *Nat. Rev. Drug Discov.* 16(1), 19–34 (2017).
- Dymock BW, See CS. Inhibitors of JAK2 and JAK3: an update on the patent literature 2010–2012. *Expert Opin. Ther. Pat.* 23(4), 449–501 (2013).
- Dymock BW, Yang EG, Chu-Farseeva Y, Yao L. Selective JAK inhibitors. *Future Med. Chem.* 6(12), 1439–1471 (2014).
- Norman P. Selective JAK1 inhibitor and selective Tyk2 inhibitor patents. *Expert Opin. Ther. Pat.* 22(10), 1233–1249 (2012).
- Pei H, He L, Shao M *et al.* Discovery of a highly selective JAK3 inhibitor for the treatment of rheumatoid arthritis. *Sci. Rep.* 8(1), 1–11 (2018).
- Keck TM, Free RB, Day MM *et al.* Dopamine D4 receptor-selective compounds reveal structure–activity relationships that engender agonist efficacy. *J. Med. Chem.* 62(7), 3722–3740 (2019).
- Mishra A, Singh S, Shukla S. Physiological and functional basis of dopamine receptors and their role in neurogenesis: possible implication for Parkinson's disease. *J. Exp. Neurosci.* 12, 1179069518779829 (2018).
- Huggins DJ, Sherman W, Tidor B. Rational approaches to improving selectivity in drug design. *J. Med. Chem.* 55(4), 1424–1444 (2012).
- Chaudhari R, Fong LW, Tan Z, Huang B, Zhang S. An up-to-date overview of computational polypharmacology in modern drug discovery. *Expert Opin. Drug. Discov.* 15(9), 1025–1044 (2020).
- **Review article on polypharmacology.**
- Allaway RJ, la Rosa S, Guinney J, Gosline SJC. Probing the chemical–biological relationship space with the Drug Target Explorer. *J. Cheminform.* 10(1), 1–14 (2018).
- Chen C, Wu M, Cen S, Wu J, Zhou J. MTLTD, a database of multiple target ligands, the updated version. *Molecules* 22(9), 1375 (2017).
- Peón A, Li H, Ghislat G *et al.* MolTarPred: a web tool for comprehensive target prediction with reliability estimation. *Chem. Biol. Drug. Des.* 94(1), 1390–1401 (2019).
- Moya-García AA, Ranea JAG. Insights into polypharmacology from drug-domain associations. *Bioinformatics* 29(16), 1934–1937 (2013).
- Schneider P, Walters WP, Plowright AT *et al.* Rethinking drug design in the artificial intelligence era. *Nat. Rev. Drug Discov.* 19(5), 353–364 (2020).
- **Advancements in application of artificial intelligence-based methods in drug discovery.**
- Krishnan SR, Bung N, Bulusu G, Roy A. Accelerating *de novo* drug design against novel proteins using deep learning. *J. Chem. Inf. Model.* 61(2), 621–630 (2021).
- Krishnan SR, Bung N, Vangala SR, Srinivasan R, Bulusu G, Roy A. *De novo* structure-based drug design using deep learning. *J. Chem. Inf. Model.* doi: 10.1021/acs.jcim.1c01319 (2021) (Epub ahead of print).
- Bung N, Krishnan SR, Bulusu G, Roy A. *De novo* design of new chemical entities for SARS-CoV-2 using artificial intelligence. *Future Med. Chem.* 13(06), 575–585 (2021).
- Bung N, Krishnan SR, Roy A. An *in silico* explainable multiparameter optimization approach for *de novo* drug design against proteins from the central nervous system. *J. Chem. Inf. Model.* 62(11), 2685–2695 (2022).
- **Recent development in artificial intelligence-based drug design.**
- Olivecrona M, Blaschke T, Engkvist O, Chen H. Molecular *de-novo* design through deep reinforcement learning. *J. Cheminform.* 9(1), 1–14 (2017).

20. Segler MHS, Kogej T, Tyrchan C, Waller MP. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Cent. Sci.* 4(1), 120–131 (2018).
21. Popova M, Isayev O, Tropsha A. Deep reinforcement learning for *de novo* drug design. *Sci. Adv.* 4(7), eaap7885 (2018).
22. Zhavoronkov A, Ivanenkov YA, Aliper A *et al.* Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat. Biotechnol.* 37(9), 1038–1040 (2019).
23. Li X, Li Z, Wu X *et al.* Deep learning enhancing kinome-wide polypharmacology profiling: model construction and experiment validation. *J. Med. Chem.* 63(16), 8723–8737 (2019).
- **Selectivity model for kinase inhibitors.**
24. Caruana R. Multitask learning. *Auton. Agent Multi-Agent Syst* 27(1), 95–133 (1998).
- **Multitask learning model.**
25. Chowdhuri S, Pankaj T, Zipser K. Multinet: multi-modal multi-task learning for autonomous driving. *Presented at: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. US, 1496–1504 Hawaii, USA, 1496–1504 (2019).
26. Sun J, Jeliaskova N, Chupakhin V *et al.* ExCAPE-DB: an integrated large scale dataset facilitating big data analysis in chemogenomics. *J. Cheminform.* 9(1), 1–9 (2017).
27. Gaulton A, Bellis LJ, Bento AP *et al.* ChEMBL: a large-scale bioactivity data-base for drug discovery. *Nucleic Acids Res.* 40(D1), D1100–D1107 (2012).
28. Rodriguez-Perez R, Bajorath J. Multitask machine learning for classifying highly and weakly potent kinase inhibitors. *ACS Omega* 4(2), 4367–4375 (2019).
- **Method to perform multitask learning using Extreme Gradient Boosting.**
29. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. CA, USA, 785–794 (2016).
30. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. *arXiv* arXiv:1609.02907 (2016). (Preprint).
31. Veličković P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y. Graph attention networks. *arXiv* arXiv:1710.10903 (2017). (Preprint).
32. Rogers D, Hahn M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* 50(5), 742–754 (2010).
33. Duvenaud DK, Maclaurin D, Iparraguirre J *et al.* Convolutional networks on graphs for learning molecular fingerprints. *Adv. Neural Inf. Process Syst.* 28 (2015).
34. Lei T, Sun H, Kang Y *et al.* 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 (2017).
35. Yang Z-Y, Yang Z-J, Dong J *et al.* Structural analysis and identification of colloidal aggregators in drug discovery. *J. Chem. Inf. Model.* 59(9), 3714–3726 (2019).
36. Jiang D, Wu Z, Hsieh C-Y *et al.* 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 (2021).
37. Pedregosa F, Varoquaux G, Gramfort A *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830 (2011).
38. Goh GB, Siegel C, Vishnu A, Hodas NO, Baker N. Chemception: a deep neural network with minimal chemistry knowledge matches the performance of expert-developed QSAR/QSPR models. *arXiv* arXiv:1706.06689 (2017). (Preprint).
39. Hamadache M, Benkortbi O, Hanini S, Amrane A. Application of multilayer perceptron for prediction of the rat acute toxicity of insecticides. *Energy Procedia.* 139, 37–42 (2017).
40. Wieder O, Kohlbacher S, Kuenemann M *et al.* A compact review of molecular property prediction with graph neural networks. *Drug Discov. Today Technol.* 37, 1–12 (2020).
41. Buffelli D, Vandin F. A meta-learning approach for graph representation learning in multi-task settings. *arXiv* arXiv:2012.06755 (2020). (Preprint).
42. Ramsundar B, Eastman P, Walters P, Pande V. *Deep Learning for the Life Sciences: Applying Deep Learning to Genomics, Microscopy, Drug Discovery, and More*. O'Reilly Media, CA, USA (2019).
43. Shen WX, Zeng X, Zhu F *et al.* Out-of-the-box deep learning prediction of pharmaceutical properties by broadly learned knowledge-based molecular representations. *Nat. Mach. Intell.* 3(4), 334–343 (2021).
44. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process Syst.* 30 (2017).
- **Interpretation of deep learning model.**
45. Lundberg SM, Erion G, Chen H *et al.* From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* 2(1), 56–67 (2020).
46. Pope P, Kolouri S, Rostrami M, Martin C, Hoffmann H. Discovering molecular functional groups using graph convolutional neural networks. *arXiv* arXiv:1812.00265 (2018). (Preprint).
- **Describes method to evaluate explainable models.**

47. Anastassiadis T, Deacon SW, Devarajan K, Ma H, Peterson JR. Comprehensive assay of kinase catalytic activity reveals features of kinase inhibitor selectivity. *Nat. Biotechnol.* 29(11), 1039–1045 (2011).
48. Tang J, Szwajda A, Shakyawar S *et al.* Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. *J. Chem. Inf. Model.* 54(3), 735–743 (2014).
49. Drewry DH, Wells CI, Andrews DM *et al.* Progress towards a public chemogenomic set for protein kinases and a call for contributions. *PLOS ONE* 12(8), e0181585 (2017).
50. Girault JA, Greengard P. The neurobiology of dopamine signaling. *Arch. Neurol.* 8, 641–644 (2004).