

Unveiling the hERG Guardians: Ligand Discovery and Molecular Insights for Cardio-Safe Drug Design

Sarth Diskalkar, Lokesh Joshi, Vikranth Varma Keerthipati, Tarek Hachad, Hargobind Singh

Abstract

Cardiovascular safety is a major concern in drug development, primarily due to the potential for compounds to inadvertently block the hERG channel, a critical regulator of cardiac action potential. This blockage can lead to severe arrhythmias and is a common reason for the withdrawal of drugs from the market. Our project addresses this issue by employing supervised machine learning and deep learning techniques to identify ligands that can safeguard the hERG channel against such adverse effects. Utilizing a comprehensive dataset from a public chemical database, we developed quantitative structure-activity relationship (QSAR) models that predict the protective capabilities of molecular ligands based on their structure. Our approach includes the use of Gradient Boosting Machines, neural network and GNN to assess and predict interactions between these ligands and the hERG channel. Initial results indicate that the models are able to predict with good accuracy and precision which determines that the data is descriptive of the ligands. Further, we have identified key molecular descriptors that are predictive of ligand efficacy, providing insights that are crucial for safer drug design. This study not only contributes to understanding hERG channel interactions but also aids in the repurposing of compounds previously deemed cardiotoxic.

Introduction

The regulation of cardiac action potentials is a critical physiological process, primarily mediated by the human Ether-à-go-go-Related Gene (hERG) encoded potassium channel. Proper functioning of this channel is essential to maintaining the heart's electrical stability. However, inadvertent blockage of the hERG channel by pharmaceutical compounds can lead to severe cardiac adverse effects, such as prolonged QT intervals, leading for example to torsade de pointes, a potentially lethal form of ventricular arrhythmia. This is a major concern for drug safety, often resulting in the withdrawal of drugs from the market.

Our project addresses this critical issue by identifying and characterizing ligands that can protect the hERG channel from potential blockers. We utilize a combination of supervised machine learning and deep learning techniques to predict and evaluate interaction between molecular ligands and the hERG channel. This approach aims to facilitate the identification of the compounds that can mitigate the adverse effects associated with hERG channel blockage.

Employing a comprehensive dataset from a public chemical database, we have developed a quantitative structure-activity relationship (QSAR) models using Gradient Boost Machines (GBM), Neural Networks, and Graph Neural Networks (GNN). These models are designed to uncover key molecular descriptors that significantly influence ligand efficacy, thereby offering crucial insights for the design of safer drugs. Our initial findings indicate that these

models can effectively identify protective ligands with high accuracy and precision. Key results from our experiments show that our QSAR models are not only capable of distinguishing between potential hERG channel blockers and non-blockers but also highlight the molecular features most correlated with protective effects. These outcomes are promising for the future of cardio-safe drug design and offer potential avenues for the repurposing of drugs previously labeled as cardiotoxic.

Related Work

The hERG channel, encoded by the human ether-a-go-go gene, is a voltage-dependent ion channel that is essential in the regulation of action potentials, as mentioned above. The inhibition of the hERG channel is a significant challenge and has become a crucial aspect of drug design and development [1,2]. The International Coordinating Conference on technical requirements for human registration mandates that small molecules be pre clinically evaluated for their effects on hERG channel blockage and QT interval prolongation, a process that is both costly and time-intensive [3]. Consequently, machine learning techniques have been adopted to expedite the evaluation of these small compounds, which can predict early-stage compounds likely to block the hERG channel [3][4].

Recent advancements have been made in the application of machine learning methodologies such as Random Forest, Support Vector Machine, Deep Neural Networks, and Graph Convolution Neural Networks to study hERG channel interactions. These methods, however, are typically restricted to predicting potential blockers or analyzing the structural properties of the hERG channel. Our project aims to identify ligands that can shield the hERG channel from these blockers, potentially enabling the repurposing of drugs that were previously not viable due to cardiotoxicity issues. By identifying effective molecular descriptors, we plan to conduct a structural analysis of the hERG channel, which could illuminate the properties that contribute to its blockage.

In a notable recent study, Zhang et al. (2022) created a consensus model that averages predictions across 30 classifiers including Deep Neural Networks, Random Forest, and XGBoost, developed from each technique. This strategy enhanced detection capabilities by leveraging the distinct characteristics identified by each model, thereby reducing prediction errors and enhancing overall prediction performance. While our project is focused on using supervised learning techniques due to constraints in time and resources, we are open to incorporating deep learning approaches should the need arise for more sophisticated analysis.

Data :

The data that we used for training the models were extracted from the assay “National Center for Biotechnology Information (2024). PubChem Bioassay Record for AID 1511, Source: Johns Hopkins Ion Channel Center. <https://pubchem.ncbi.nlm.nih.gov/bioassay/1511>.” The provided assay description outlines the importance of the hERG channel in cardiac repolarization which we have described in the introduction of this report. This is crucial as it

determines the need to identify compounds that can protect hERG channel from blockage by its known inhibitors. This is crucial as it uncovers a need in drug development to minimize cardiotoxicity. The assay has employed a high throughput TI + flux assay to screen a library of compounds for their effect on the hERG channel's response to dofetilide, a well characterized blocker. The protocol also involves the use of CHO cells expressing the hERG channels. Also, the FluxOR (thallium sensitive fluorescent dye) and measurement of intracellular thallium flux upon compound treatment were used. The assay was conducted in 384 well plates, with cells plated and treated with compounds according to a defined protocol. The fluorescent measurements were taken using a kinetic magnetic imaging plate reader, and data analysis includes calculating fluorescence ratios and percentage inhibition to evaluate compound effects. The data that has been used for the model training is SMILES (Simplified Molecular Input Line Entry System), for calculating the descriptors for the defined ligands used in the study.

Mordred : Descriptor calculator

Mordred calculator is an open source software package designed for the calculation of a wide range of descriptors. It is developed in python and has a very user-friendly interface for generating the molecular descriptors from chemical structures.

Special features for the using mordred:

Mordred offers a lot of descriptors that cover a vast collection of chemical aspects like structure, topology, connectivity, and physicochemical properties. Mordred can calculate descriptors for a wide variety of molecules, ranging from small molecules, polymers, and even large biomolecules as proteins and nucleic acids. We can use mordred to be tailored for the specific research needs and analyses for the specific needs of the researcher. Mordred can also be easily integrated with existing computational workflows. The descriptor type can be 2D, 3D, physiochemical and topological descriptors. The calculator has also been known for its performance and is quite efficient for large datasets. The calculator scales well with the size and complexity of molecular structures making it suitable for high throughput and virtual screening studies like this one. It is a versatile and reliable tool for calculation offering a comprehensive set of descriptors and serves as a valuable resource for researchers especially in the field of drug discovery.

SMILES were used from the assay and the corresponding descriptors were calculated for the ligands provided.

nAcid	nBase	SpAbs_A	SpMax_A	SpDiam_A	SpAD_A	SpMAD_A	LogEE_A	VE1_A	VE2_A	VE3_A
0	0	39.35241079	2.434402071	4.856796945	39.35241079	1.311747026	4.335123212	4.31909628	0.143969876	2.5616584
0	0	24.70760556	2.463300278	4.771223767	24.70760556	1.300400293	3.891809269	3.814599903	0.2007684159	1.9806896
0	0	25.82002363	2.39320448	4.677596862	25.82002363	1.291001182	3.903142331	3.917147167	0.1958573583	2.0585108
0	1	multiple fragmen	multiple fragmen	multiple fragmen	multiple fragmen	multiple fragmen	multiple fragmen	multiple fragmen	multiple fragmen	multiple fragmen
0	0	34.63866037	2.488775595	4.977059323	34.63866037	1.237095013	4.260461092	4.280155748	0.1528627053	2.4836088
2	0	29.87688765	2.453495792	4.822983843	29.87688765	1.358040348	4.050954012	4.248876369	0.1931307441	2.2351119
0	0	26.38429841	2.403397286	4.649341368	26.38429841	1.31921492	3.932694137	3.618978176	0.1809489088	1.9793388
0	0	19.57770881	2.322302805	4.644605609	19.57770881	1.305180587	3.627230862	3.605457326	0.2403638218	1.687913
1	0	multiple fragmen	multiple fragmen	multiple fragmen	multiple fragmen	multiple fragmen	multiple fragmen	multiple fragmen	multiple fragmen	multiple fragmen

Figure 1: Output post descriptor calculation.

As it can be seen in Figure 1, descriptors have varied values and some rows contain error values. Therefore, there was a need for data cleaning, processing and scaling of the data to make the training of the model more comprehensive. Thus, data cleaning, processing and scaling were performed for the datasets. The datasets were divided into active and inactive molecules, active being the molecules that protect the hERG channel and inactive being the molecules that don't protect the hERG channel.

Data preprocessing, cleaning and scaling:

As mordred gave a list of descriptors to select from, we decided to make the selection more comprehensive by choosing biologically important descriptors that were most relevant to the study we conducted

The descriptors selected were based on the following categories, as we are trying to determine the protective nature of the ligands:

1. **Lipophilicity:** This is an important physicochemical property as it influences drug absorption, which is a crucial aspect for ADME (Absorption, Distribution, Metabolism, and Excretion).
2. **Van der Waals surface area descriptors:** This provides information about the molecular size, shape and surface properties, which are important for ligand receptor interactions and binding affinity. They also determine the hydrophobic, hydrophilic and intermediate regions of a molecule's surface, thus contributing to the overall molecular profile.

3. **Log based VSA:** These are important for determining the overall molecular interactions with biological targets. Also, these are important for the drug receptor binding and pharmacological activity.
4. **Kier flexibility indices:** These quantify the molecular flexibility based on the number of rotatable bonds and the connectivity of atoms within a molecule. This provides information regarding the rigidity of a molecule, which can impact its bioactivity.
5. **Spatial arrangements:** Descriptors that determine the spatial arrangements within a molecule, reflecting its 3D structure and electronic properties.

Other descriptors that were included were hydrogen bond donors, acceptors, number of rotatable bonds and rotatable bond ratio. These descriptors were able to capture the structure, size, shape, flexibility, lipophilicity and functional groups.

For descriptor calculation we utilized the descriptor_extraction.ipynb jupyter notebook to extract the descriptors. After that, descriptor_cleaning.ipynb was used for cleaning, processing and scaling, which was performed on the dataset to prepare it for training the models.

The means of the columns for the dataset were computed to check the central tendency of the data.

Mean of each column after replacing non-numeric values:

SLogP	3.125817
SMR_VSA1	12.065745
SMR_VSA2	0.207384
SMR_VSA3	9.218708
SlogP_VSA1	8.610442
SlogP_VSA2	27.229044
SlogP_VSA3	8.045799
Kier1	18.418495
Kier2	8.473385
Kier3	4.822498
Mor01	NaN
Mor02	NaN
Mor03	NaN
GATS1c	1.545269
GATS2c	0.999487
GATS3c	1.083734
MATS1c	-0.526506
MATS2c	0.057616
MATS3c	-0.015988
nHBAcc	4.527062
nHBDOn	1.306057

PNSA1	NaN
PPSA1	NaN
DPSA1	NaN
FPSA1	NaN
WNSA1	NaN
WPSA1	NaN
SpAbs_A	29.800948
SpMax_A	2.419060
SpDiam_A	4.776308
SpAD_A	29.800948
SpMAD_A	1.279744
LogEE_A	4.049045
VE1_A	4.026745
VE2_A	0.178717
VE3_A	2.210938
VR1_A	366.491895
VR2_A	14.218590
VR3_A	6.274785
nRot	4.018041
RotRatio	0.158536

These show a great deal of variance between them, thus there was a necessity for scaling, which was done.

After scaling the data was as follows :

	SLogP	SMR_VSA1	SMR_VSA2	SMR_VSA3	SlogP_VSA1
count	1.552000e+03	1.552000e+03	1.552000e+03	1.552000e+03	1.552000e+03
mean	2.838508e-16	-3.662591e-17	-2.746944e-17	-1.373472e-16	1.465037e-16
std	1.000322e+00	1.000322e+00	1.000322e+00	1.000322e+00	1.000322e+00
min	-4.566919e+00	-1.458897e+00	-1.816042e-01	-1.518147e+00	-1.362892e+00
25%	-5.953180e-01	-8.791785e-01	-1.816042e-01	-6.973800e-01	-5.213315e-01
50%	2.654365e-02	-2.838896e-01	-1.816042e-01	9.251401e-02	-4.553464e-01
75%	6.403266e-01	7.133872e-01	-1.816042e-01	2.689036e-01	3.374827e-01
max	4.584853e+00	6.889638e+00	1.837863e+01	4.349918e+00	4.360024e+00

	SlogP_VSA2	SlogP_VSA3	Kier1	Kier2	Kier3
count	1.552000e+03	1.552000e+03	1.552000e+03	1.552000e+03	1.552000e+03
mean	2.724052e-16	-1.052995e-16	2.952964e-16	1.465037e-16	-3.662591e-17
std	1.000322e+00	1.000322e+00	1.000322e+00	1.000322e+00	1.000322e+00
min	-1.806494e+00	-1.067688e+00	-2.609420e+00	-2.540507e+00	-2.085644e+00
25%	-7.222473e-01	-1.067688e+00	-6.147314e-01	-6.765817e-01	-7.273333e-01

50%	-2.075302e-01	-1.991904e-01	-8.920149e-02	-5.655924e-02	-9.491734e-02
75%	4.738209e-01	6.364150e-01	6.830926e-01	6.409237e-01	5.663240e-01
max	5.485250e+00	4.601999e+00	5.618522e+00	4.173226e+00	8.620205e+00

	GATS1c	GATS2c	GATS3c	MATS1c	MATS2c
count	1.552000e+03	1.552000e+03	1.552000e+03	1.552000e+03	1.552000e+03
mean	-6.775794e-16	-3.113203e-16	-3.983068e-16	-3.868612e-16	3.204767e-17
std	1.000322e+00	1.000322e+00	1.000322e+00	1.000322e+00	1.000322e+00
min	-6.612103e+00	-3.039131e+00	-3.587100e+00	-2.996361e+00	-2.627129e+00
25%	-4.968828e-01	-7.181122e-01	-6.916260e-01	-6.862731e-01	-7.350078e-01
50%	5.775311e-02	-7.585470e-03	-3.939004e-02	-2.815855e-02	-3.126101e-02
75%	6.084317e-01	6.822560e-01	6.589609e-01	6.314019e-01	6.630530e-01
max	2.666798e+00	3.412246e+00	3.558491e+00	4.315540e+00	3.684283e+00

	MATS3c	nHBAcc	nHBDOn	nRot	RotRatio
count	1.552000e+03	1.552000e+03	1.552000e+03	1.552000e+03	1.552000e+03
mean	1.831296e-17	-2.060208e-16	-1.831296e-17	2.861400e-17	2.243337e-16
std	1.000322e+00	1.000322e+00	1.000322e+00	1.000322e+00	1.000322e+00
min	-3.926123e+00	-2.040548e+00	-1.557297e+00	-1.905411e+00	-2.124428e+00
25%	-6.829033e-01	-8.834670e-01	-3.649315e-01	-4.827693e-01	-6.355037e-01
50%	7.011375e-02	-3.049266e-01	-3.649315e-01	-8.555405e-03	-1.143803e-01
75%	6.963929e-01	8.521543e-01	8.274342e-01	4.656585e-01	5.556355e-01
max	2.768235e+00	4.323397e+00	5.596897e+00	5.207797e+00	8.454769e+00

After scaling, the data was more comprehensive and fit for training models.

From the overall cleaning, processing and scaling steps that were done, it resulted in datasets active.csv and inac.csv. The final dataset was made by concatenating these two. As we had imbalance for the active and inactive datasets, we used SMOTE (Synthetic Minority Over-sampling Technique) for that purpose. In the end we had 3552, including both active and inactive molecules with 14 descriptors that were selected for the training of the model.

Methods

When focusing on the development of supervised QSAR models for identifying ligands that protect the hERG channel, we used a variety of machine learning and deep learning techniques. These models work by learning from already known structures of molecules and their effects on the hERG channel in order to predict the activity of new unseen compounds.

Multiple experiments were run that involved simple models such as logistic regression and simple vector machine (SVM). Throughout the rest of the report, the best performing models will be discussed. These models were **XGBoost**, a **feedforward neural network**, and a **graph neural network (GNN)**.

One supervised learning technique we implemented is XGBoost, a gradient boosting algorithm which is an advanced ensemble learning model. XGBoost was chosen as gradient boosting minimizes the overall prediction error while also being resistant to overfitting. Specifically, the XGBoost classifier was utilized from the scikit-learn library. The `scale_pos_weight` of the model was set to be 1552/2000, and the rest of the hyperparameters were set to their default values.

Another method chosen was a feedforward neural network. Feedforward neural networks were selected as they have the ability to capture nuanced relationships within the data and are able to balance variance/bias compared to classical machine learning algorithms. However, compared to XGBoost, there are no explicit training techniques that prevent overfitting onto the dataset other than setting the learning rate and monitoring the loss curve. To reduce the chance of training an overfitted model, a shallow feedforward neural network was used. The model was designed using Tensorflow, and starts with a Dense input layer with 128 units, followed by a rectified linear unit (ReLU) activation function. This was followed with a dropout layer set at 0.5, another Dense layer with 64 units, and a ReLU activation function. Finally, this was followed by another dropout layer set at 0.5, a Dense output layer outputting 1 unit, finally followed by a sigmoid activation function. The purpose of having the last layer of the model be the Sigmoid layer is to output a probability. The optimizer used was Adam with a binary cross entropy loss function. The model was trained for 20 epochs with a batch size of 32 and a learning rate of 0.001.

The last method chosen was utilizing a graph neural network. GNNs are able to capture the structural properties of a molecule, by having node and edge feature vectors. Other classical machine learning models or deep learning models typically aren't able to fully capture the structure relationships that occur within a graph. The data preprocessing step for this method consisted of converting SMILES strings into graph objects, where each node was featurized with the atomic number and the edge was featured with the bond type. The model consists of 2 graph convolutional layers, and 1 linear layer with an input as a graph and the output as a singular, probability value. The model was created using PyTorch Geometric, and trained with a learning rate of 0.01 and the Adam optimizer with a binary cross entropy loss function. The model was trained for 3 epochs.

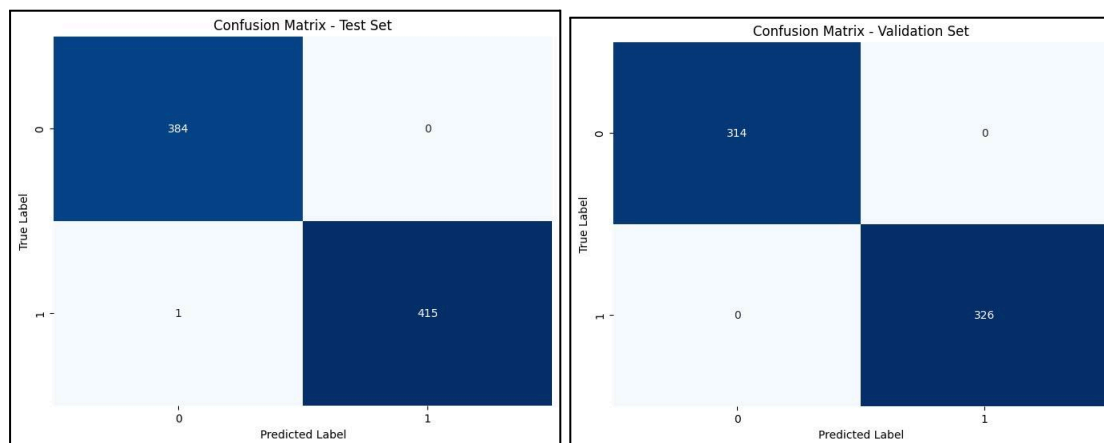
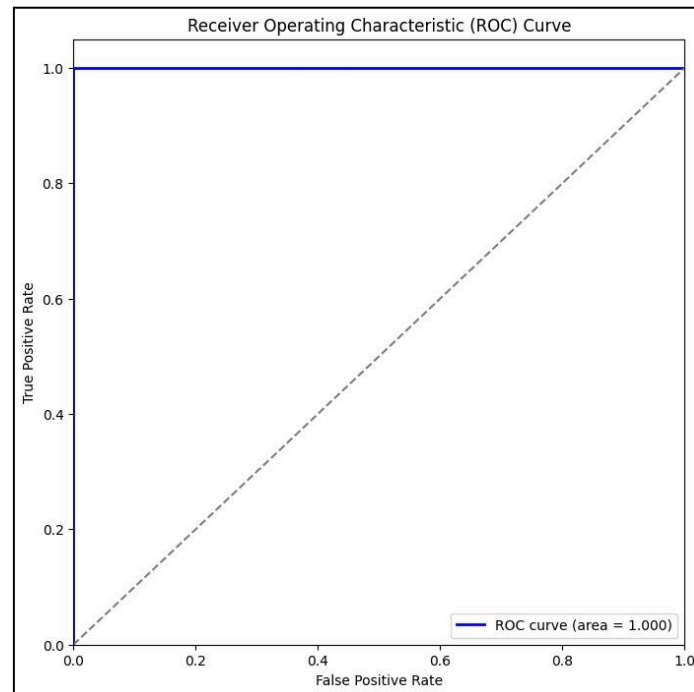
Experiments

Our experimental design was focused on validating the predictive capabilities of our machine learning models, primarily our Gradient Boosting Machine (XGBoost) and neural network models. We aimed to assess how well these models could identify and predict ligands that protect the hERG channel, which involved a series of tests and visual evaluations to understand the models' performance and interpretability.

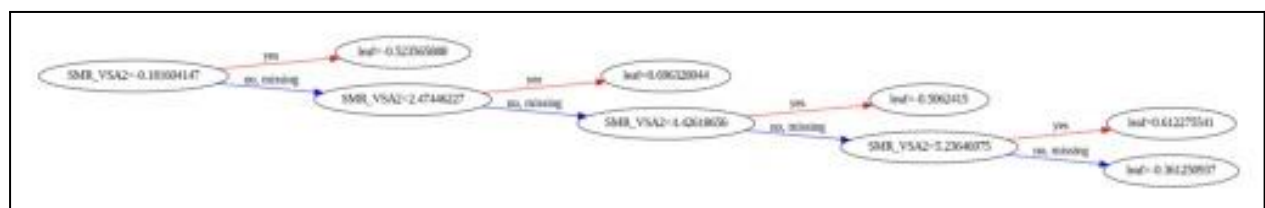
XGBoost Classifier

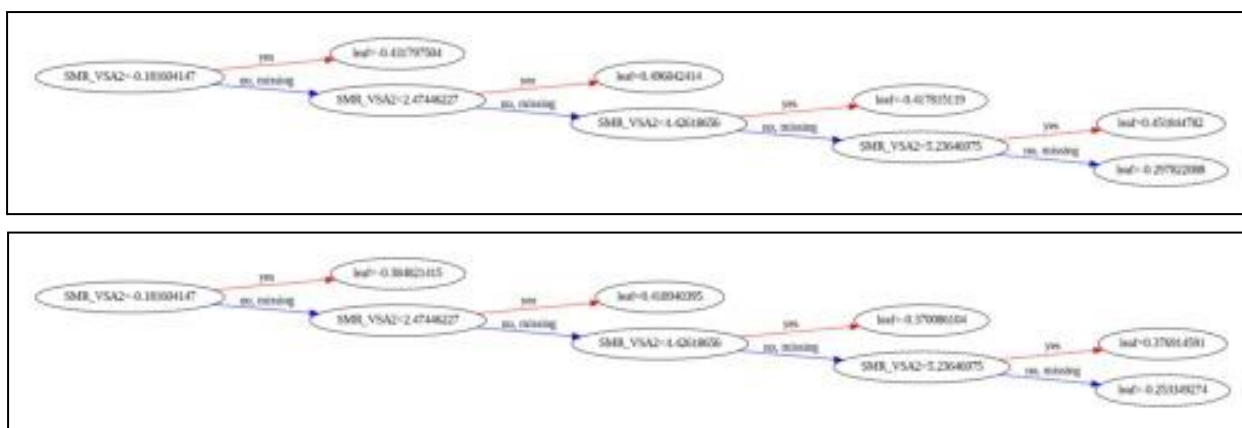
The XGBoost classifier was extensively tested to evaluate its effectiveness in distinguishing between protective and non-protective ligands.

- Performance Metrics: The model's performance was first assessed using a confusion matrix and an ROC curve. The ROC curve for the XGBoost classifier, and the confusion matrices (on the test and validation sets), providing insights into the true positives, false positives, true negatives, and false negatives, are shown below.

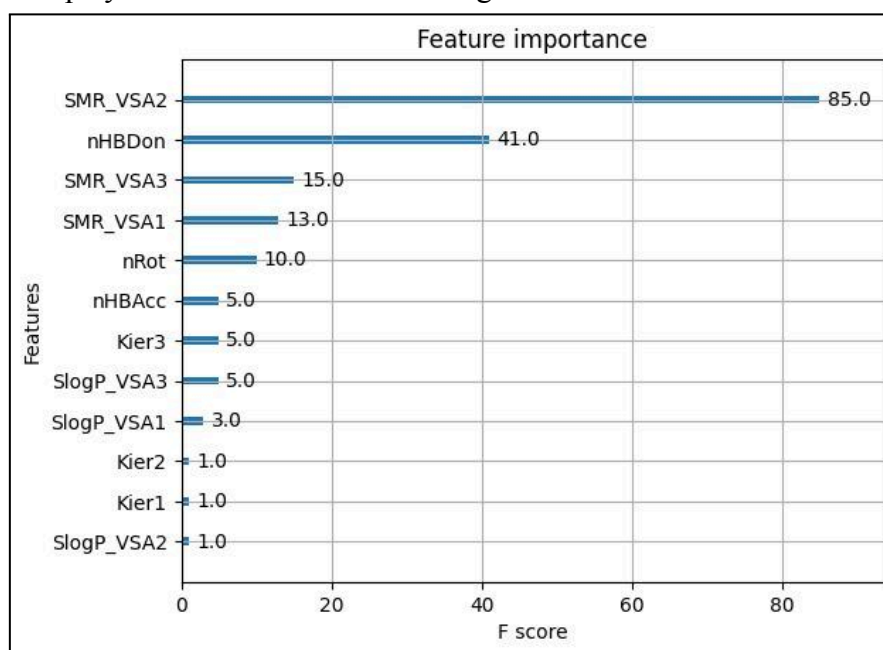


- Decision Trees Visualization: A few representative decision trees from the XGBoost model are presented to illustrate how decisions are being made at various levels of the tree structure. These can be found here:

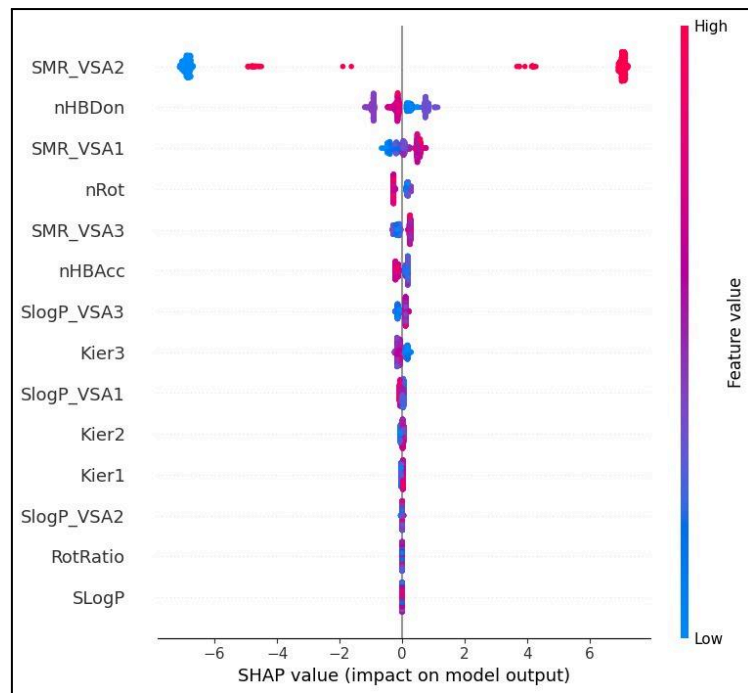




- Feature Importance Analysis: To understand which features significantly impact the model predictions, a feature importance graph was generated. This horizontal bar graph displays the F score on the x-axis against a list of all features on the y-axis:



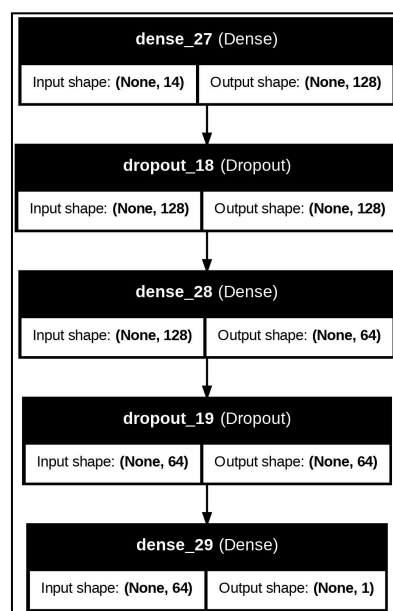
- Model Explanation with SHAP Values: To further dissect the model's decision-making process, SHAP (SHapley Additive exPlanations) values were used to explain the prediction of individual samples. This graph elucidates the contribution of each feature to the model's prediction:



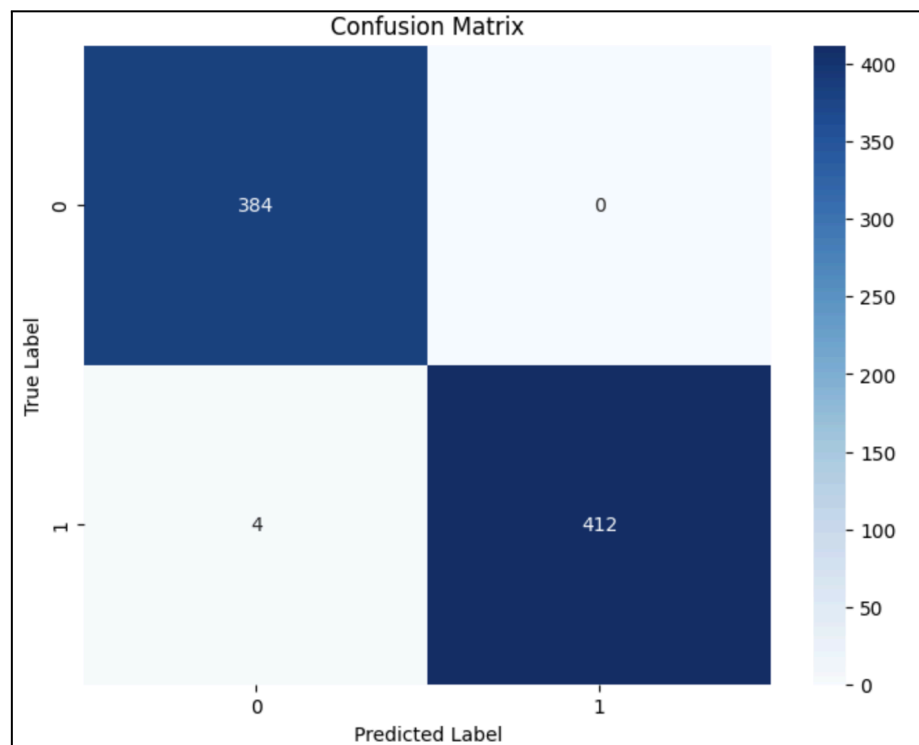
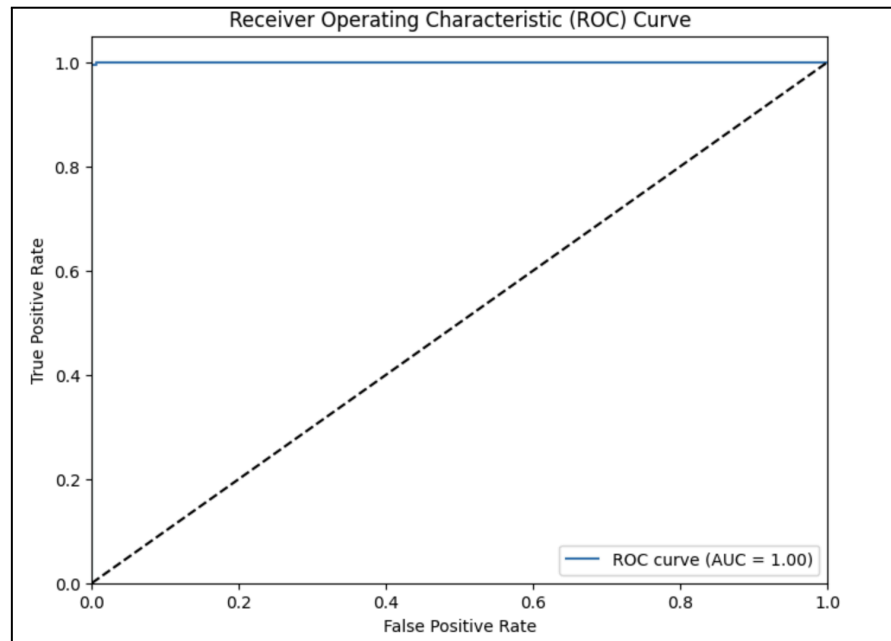
Feedforward Neural Network

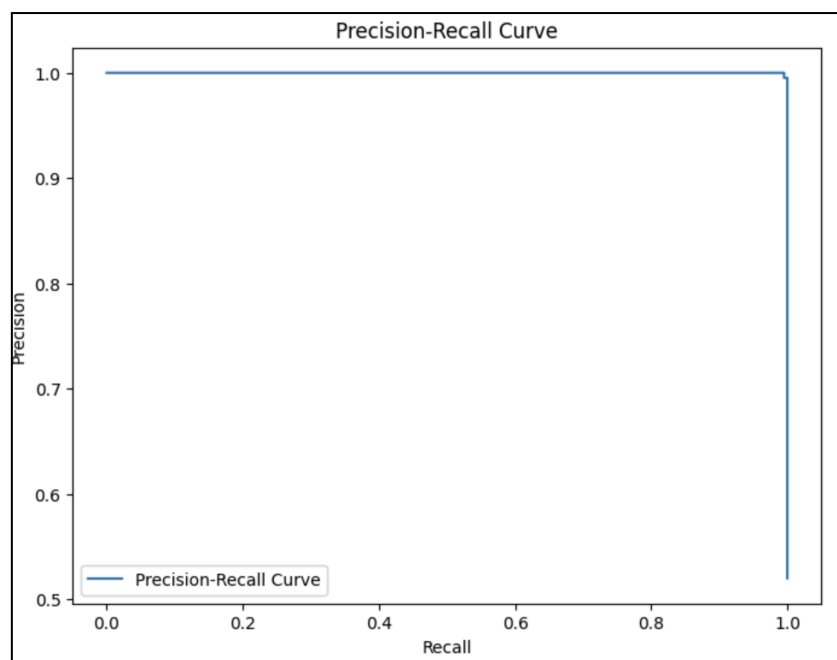
The neural network model was designed to complement the predictions made by the XGBoost, offering a different perspective on data interpretation through its layered architecture.

- **Architecture Visualization:** A diagram of the neural network architecture is provided to showcase the layers and their connections, illustrating the flow of data through the network:



- Performance Metrics: Similar to the XGBoost model, the performance of the neural network was evaluated using a confusion matrix and ROC curve, along with a precision and recall curve. These metrics provide a comprehensive view of model accuracy and its ability to manage class imbalances:





Conclusion and Discussion

In our study, we addressed the critical challenge of drug-induced cardiac arrhythmias caused by blockage of the hERG channel—a well-known risk factor in pharmacology. To mitigate this risk, our team developed a series of sophisticated machine learning models that can predict the protective properties of molecular ligands against such blockages. These models, which include Gradient Boosting Machines (GBM), feedforward neural networks, and Graph Neural Networks (GNN), were trained using a comprehensive dataset sourced from the Johns Hopkins Ion Channel Center. The application of GBM allowed us to leverage an ensemble of decision trees, which worked collectively to refine and improve the prediction accuracy regarding the interaction between ligands and the hERG channel. This method is particularly effective in reducing overfitting and enhancing model generalizability across diverse molecular datasets.

Moreover, our use of feedforward neural networks provided a robust framework for capturing non-linear relationships within the data. By implementing multiple layers with dropout regularization, we managed to reduce the risk of overfitting, making the neural network model both resilient and sensitive to the complex patterns in the data. Perhaps the most innovative aspect of our approach involved the use of Graph Neural Networks. GNNs are uniquely suited to model the molecular structures directly, representing atoms as nodes and bonds as edges, which allows for a nuanced understanding of the molecular architecture and its functional impacts on the hERG channel. This method stood out by effectively capturing the spatial and structural relationships that are crucial for understanding how molecules interact with biological systems.

Through a series of experiments, these models were rigorously tested and validated, demonstrating good precision and accuracy in predicting the protective effects of ligands. The models' ability to identify key molecular descriptors—such as lipophilicity, molecular size, and flexibility—provides crucial insights for drug design, emphasizing properties that

enhance safety without compromising efficacy. This work not only advances our understanding of hERG channel interactions but also sets a new standard for employing advanced machine learning techniques in the safety assessment of pharmaceutical agents. By bringing these innovative tools to bear on drug safety evaluation, we aim to pave the way for developing safer therapeutic solutions that can prevent life-threatening side effects and enhance patient outcomes.

In the discussion of our work, several limitations and future directions emerge, which are crucial for advancing the field and enhancing the practical impact of our findings. One significant limitation is the dependency on the dataset provided by the Johns Hopkins Ion Channel Center. While comprehensive, the dataset may still contain biases or gaps that could skew model predictions. This limitation underscores the importance of continually expanding our data sources to enhance the diversity and accuracy of the training data, which would in turn improve the generalizability of our models across different molecular environments. Another challenge is the computational complexity of the models, particularly the Graph Neural Networks (GNNs). These models, while powerful, require substantial computational resources, which can limit their applicability in less resourced settings. Additionally, the risk of overfitting remains a concern, despite measures such as dropout layers in neural networks. Ensuring that models generalize well to new, unseen data is critical, and thus, extensive external validation is necessary to confirm the models' utility in real-world scenarios.

Looking forward, enhancing data collection is a critical step. By incorporating a broader array of ligands and more varied molecular structures, we can significantly improve the robustness and predictive power of our models. Exploring and integrating novel machine learning techniques could also yield substantial benefits. Emerging algorithms that offer improved efficiency and accuracy could revolutionize our approach to predicting hERG channel interactions. Furthermore, clinical validation of our models represents a crucial frontier. Collaborating with pharmaceutical companies to test predictions in clinical trials would provide invaluable insights into the models' practical utility and could accelerate the adoption of safer drug design practices based on machine learning predictions. An interdisciplinary approach, involving chemists and pharmacologists, would further enrich our understanding of the molecular dynamics at play, leading to more effective and safer pharmaceuticals. This collaborative effort is essential not only for refining the models but also for translating these technological advancements into tangible benefits for drug safety and public health.

Source Code

Below is the link to our GitHub repository, which contains the source code and a README file with detailed information about the code and instructions on how to use it:

<https://github.com/sarth-diskalkar/hERG-Guardians>

(References below)

References

1. Sanguinetti, M. C., Jiang, C., Curran, M. E., & Keating, M. T. (1995). A mechanistic link between an inherited and an acquired cardiac arrhythmia: HERG encodes the IKr potassium channel. *Cell*, 81(2), 299-307.
2. Villoutreix, B. O., & Taboureau, O. (2015). Computational investigations of hERG channel blockers: New insights and current predictive models. *Advanced drug delivery reviews*, 86, 72-82.
3. Liu, M., Zhang, L., Li, S., Yang, T., Liu, L., Zhao, J., & Liu, H. (2020). Prediction of hERG potassium channel blockage using ensemble learning methods and molecular fingerprints. *Toxicology Letters*, 332, 88-96.
4. Sun, F., Sun, J., & Zhao, Q. (2022). A deep learning method for predicting metabolite–disease associations via graph neural network. *Briefings in bioinformatics*, 23(4), bbac266.