Prediction of hERG potassium channel PAS-domain variants trafficking via machine learning

Kalyan Immadisetty[1] and Peter Kekenes-Huskey[1,*]

[1]Stritch School of Medicine, Maywood IL 60153,
Email: pkekeneshuskey@luc.edu

November 4, 2021

1 Abstract

Long QT syndrome (LQTS) is a cardiac condition characterized by a prolonged QT-interval. This prolongation can contribute to fatal arrhythmias despite otherwise normal metrics of cardiac function. Hence, genetic screening of individuals remains a standard for initial identification of individuals susceptible to LQTS. Several genes, including KCNH2 that encodes for the Kv11.1 channel are known to cause LQTS2, however, only a small subset of variants found in the human population are established as pathogenic. Hence, the majority of its missense mutations are known to be benign or are variant of unknown significance (VUS). Here we use molecular dynamics (MD) simulations and machine learning (ML) to determine the propensity for Kv11.1 channel variants to present loss-of-function behavior. Specifically, we use these computational techniques to correlate structural and dynamic changes in an important Kv11.1 subdomain, the PAS-domain (PASD), with the channel's ability to traffic normally to the plasma membrane. With these techniques, we have identified several molecular features, namely, number of hydrating waters and intra-PASD hydrogen bonds, as moderately predictive of trafficking. Together with bioinformatics data including sequence conservation and folding energies, we are able to predict with reasonable accuracy ($\approx$75%) the ability of VUS' to traffic. Additionally, we compared two ML algorithms i.e., Decision tree (DT)s and Random forest (RF) for their robustness, and we report that RF performed moderately better than DTs'. Features derived from MD trajectories particularly help improve the prediction of trafficking deficient variants by both ML techniques.

2 Introduction

Long QT syndrome is a condition that predisposes individuals to potentially fatal arrhythmias[1]. Several genes have been identified that are implicated in LQTS. These include KCNQ1, KCNH2, and SCN5A, which encode the ion channels KvLQT1, hERG and NaV1.5[1]. Screening of these genes for mutations has become a de facto standard for identifying patients with the potential for LQTS. While variants have been identified in the human population, only a small subset yield disease phenotypes. This motivates the need for improved strategies for identifying potentially pathogenic variants from a large set of variants of unknown significance (VUS).

The recent availability of moderate and high resolution structural data for the hERG channel raises the possibility that information beyond its primary sequence can be used to infer pathogenicity.

A recent study indicates that PAS domain variants that give rise to a loss-of-function phenotype can be determined using a solubility assay [2]. Namely, variants with reduced solubility correlated with the loss-of-function phenotype. In that study, the authors also used bioinformatics analyses (FoldX scores) to show that loss-of-function variants tended to exhibit thermodynamically unfavorable folding scores. This finding motivated our hypothesis that LOF variants perturb the native structure of PASD.

To investigate this hypothesis, we used the structure of the PAS domain that was obtained from X-ray crystallography at resolution. molecular dynamics (MD) simulations were used to predict the impact of a given variant on the protein's native structure and dynamics. MD is a powerful computational tool that offers atomic level insights[3], and we have successfully used this tool to study the structural, dynamic, and functional aspects of various therapeutically important proteins such as troponin C[4], parvalbumins[5] and purinergic receptors[6]. From these simulations, we evaluated several features to determine their capacity to discriminate benign from loss-of-function variants. These features, along with additional features predicted from bioinformatics, were used with statistical and machine learning techniques to classify novel variants as benign versus LOS. Despite the strong bias of the training data toward LOF variants, we found that the machine learning approaches, and in particular, the random forest approach, improved upon predictions based on bioinformatics alone. Using the trained ML approaches, we proposed ten novel variants that we estimated would yield normal and LOF phenotypes to validate our predictions.

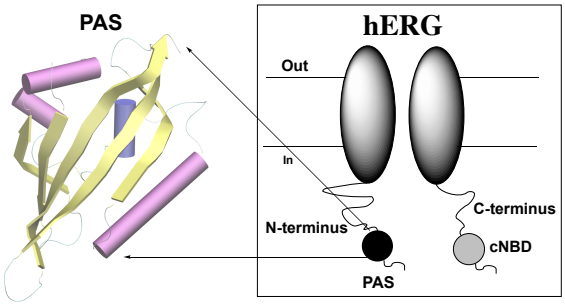


Figure 1: Schematic of hERG potassium channel is shown on the right. Structure of PAS domain on the left generated using the crystal structure (PDB: 4HQA [7]). β-strands are colored yellow and helices are colored purple.

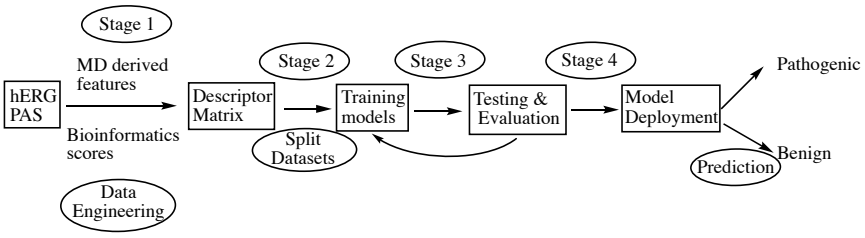


Figure 2: The supervised ML protocol adopted in this study for classifying the variants of hERG K^+ channel.

We utilized supervised machine learning (SML) [8] in this work. SML is a ML approach in which the target labels are already known. The target labels are pathogenic/benign. The two most popular SML

techniques are classification and regression [8]. This is a classification problem, since we are classifying the variants of this channel into pathogenic or benign. Among all classification methods, decision trees (DT) are always the first choice, especially for small/medium data sets and particularly for biological/medical data sets [9]. The reason being that DTs are easily interpretable, classification attributes at each stage can be easily understood by a nontechnical person such as physicians, efforts that are required to manipulate the data such as data cleaning and transformation/normalization are very minimal, and DTs are non-parametric because they make no distributional assumptions on the data[10, 11]. In some scenarios DTs are error prone and biased, particularly in situations where the data is too noisy, hence the outcome might not be always reliable[11]. Therefore, it is always advisable to try a second method as a backup strategy, which also helps to verify the results of the first method.

The second method we have used is random forest (RF). It is an ensemble method and is supposed to be better than the DT, given the hyperparameters are appropriately tuned to fit the data during training [12, 13]. The basic approach of RF is similar to DT, however, instead of fitting data to a single DT, models are trained on multiples of DTs, each on an independent subset of data, that is drawn randomly and repeatedly (also called as bootstrapping) and the aggregate values of all DTs are estimated (called as bagging), which then will be used to classify data [12, 13]. In this approach, not only the data, even the features are drawn randomly to fit the DTs and not all features are used by a single DT. Therefore, the chance of error is lower in this approach. The other advantages of RF are that it can handle high dimensionality data efficiently, handle missing values, and identifying outliers [12, 13].

3 Results and discussion

3.1 WT vs pathogenic/benign structural comparisons from MD

Molecular dynamics simulations were performed on PAS domain variants. Among the variants chosen for the study, we present representative structural data and dynamic data for the WT, the loss of function variant, and a normally trafficking, . In Fig. 3 we superimpose three representative structures obtained from these simulations, which show relatively similar folds. We therefore examined other structural and dynamic features as a potential basis for loss-of-function versus normal phenotypes. We first examined changes in the hydrogen binding networks intrinsic to the native structure relative to the site-directed variants. The WT variant features 66 hydrogen bonds, relative to about 65 for V113E and 62 for T74P. . As anticipated, the hydrogen bond loss occurs. We also observe that the root mean square fluctuations are generally low, with notable increases at residues 85–95 and 115–125. In contrast, for the T74P variant, we observe a substantial increase in RMSF localized to the site 74 region (Fig. 4). This coincides with appreciable increases across all residues. This is not unexpected given that the proline substitution is expected to perturb the region given that its backbone torsion angles compared to most other amino acids.

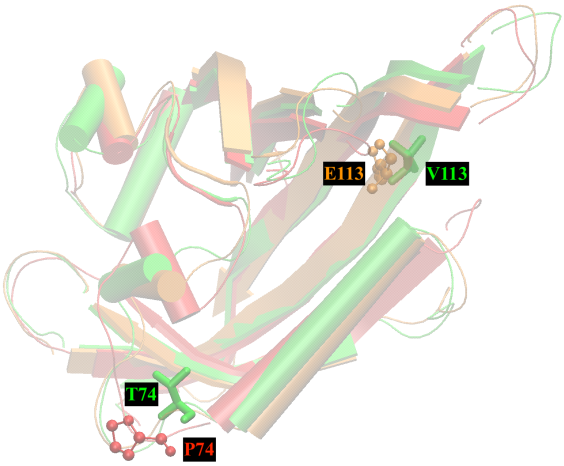


Figure 3: WT, V113E, and T74P variants are colored green, orange, and red, respectively. Residues of WT are shown in the licorice representation and of the others as ball and stick models. T74/P74 is in the loop and E113/V113 is stacked between β-strands.

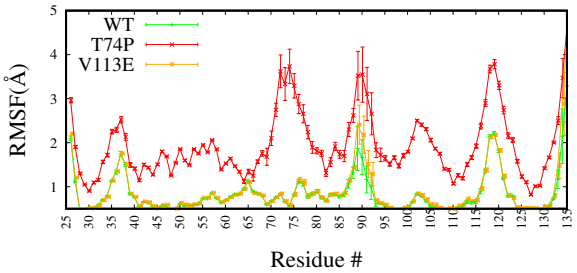


Figure 4: Comparison of RMSFs of WT, V113E, and T74P. WT, V113E, and T74P are colored green, orange, and red, respectively. T74P is pathogenic and V113E is benign [2].

3.2 Prediction of PAS domain solubility

We next evaluated these features in addition to other structural and dynamic features across the broader set of variants considered in the study. These features include solvent-accessible surface area, number of hydrating waters, , which are quantified by $\lambda_i = value_i - value_{WT}$ In Fig. 5 we present rank ordered feature data for each variant relative to the WT. We highlight normally trafficking versus loss of function variants with blue and red bars, respectively. Among the features we considered, the number of waters appeared to most strongly discriminate normal from loss of function trafficking variants. This is evidenced by the tendency of the blue (normal) variants to distribute to lower λ values relative to the LOF variants. We anticipate that the increased number of waters in the LOF variants is attributed to the partial unfolding of the PAS structure. However, this structural change does not appear to impact SASA in an obvious manner. Similar to the data for the V113E and T74P variants assessed in the previous section, we observe that LOF variants tend to have a lower number of hydrogen bonds compared to WT.

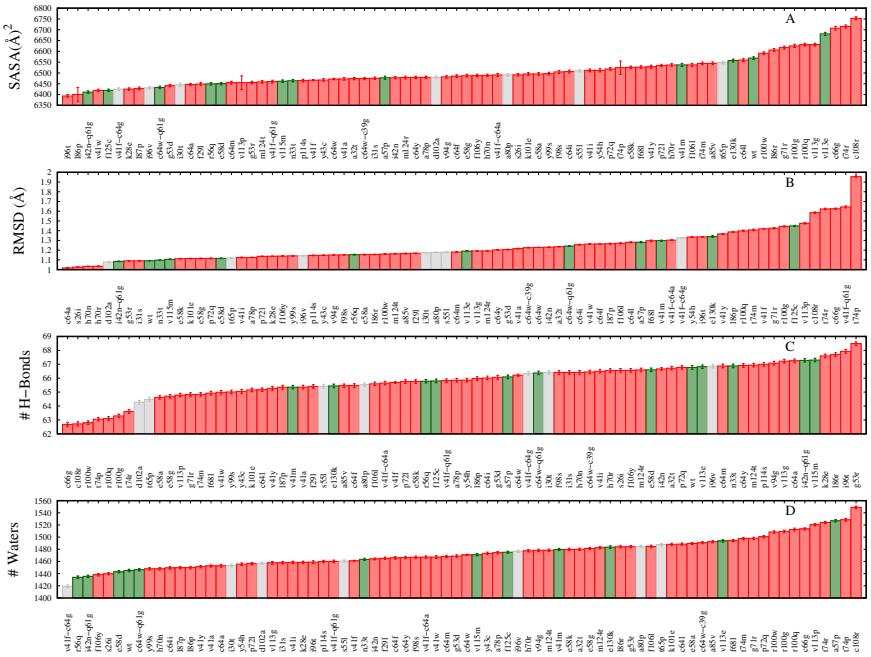


Figure 5: SASA, protein backbone RMSD, # hydrogen bonds, and # waters are shown in panels A, B, C and D, respectively. Waters within 4Å of the protein were calculated. Average of three MD trials and the standard error of mean are reported for each variant. Red, green, and white represent pathogenic, benign and variants of unknown significance , respectively.

To assess the predictive power of these data, we computed the conditional probabilities for LOF and benign variants with respect to lambda in the second column of Fig. 5. For the LOF conditional probability, we report

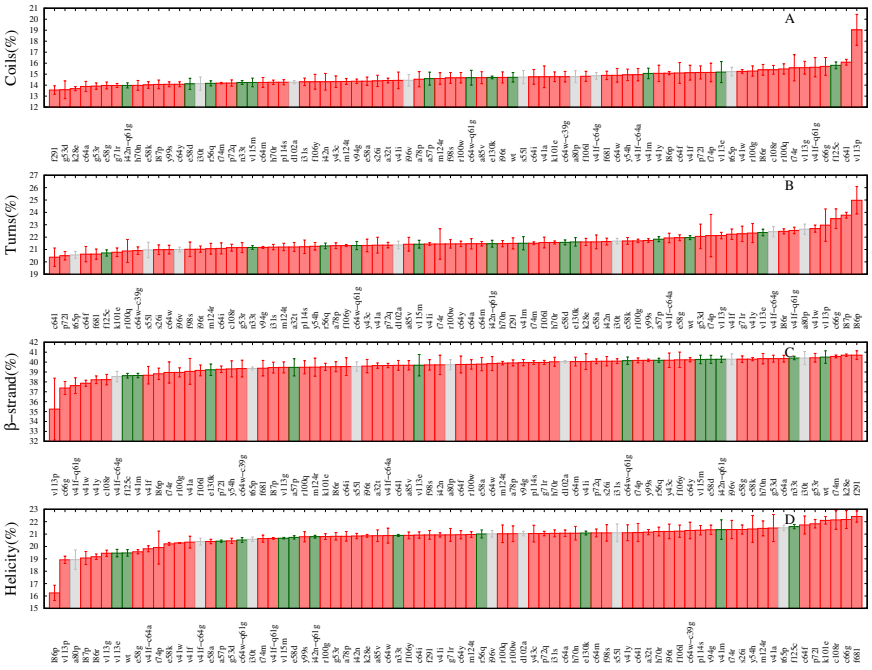


Figure 6: Protein secondary structural components compared. Average of three MD trials and standard deviation are reported for each variant.

the probability that a variant with $\lambda > value$ is a LOF phenotype, e.g. $P(LOF|\lambda > value)$. Conversely, the benign variant conditional probability refers to variants with $\lambda < value$, e.g. $P(LOF|\lambda > value)$ With these data, we observe that the LOF conditional probability for the water feature is monotonic, whereby increasing threshold parameters yields LOF probabilities approaching 100%. Conversely, the benign conditional probability monotonically decreases from 100% to as the threshold is increased. In contrast, less predictive metrics such as the RMSF exhibited nonmonotonic trends that complicate the classification of variants.

We complement these data with receiver operator characteristic (ROC) curves for each of these features in Fig. 8. These ROC curves present the true positives and false positive rate for each feature as a function of the threshold value. Random features that are unable to discriminate true and false positives, that is, yield equal rates for both terms as a function of the threshold, are designated by the dashed black line. Metrics that perform better than random exhibit ROC curves above this line, which is quantified by reporting the area under the curve (AUC) values. AUC values approaching 1.0 represent metrics that maximize the TPR and minimize the FPR.

We define the false positive rate by

$$\text{FPR} = \frac{\text{false positive}}{\text{false positive} + \text{true positive}}$$

which we later use to evaluate the performance of the classification approaches we consider.

3.3 WT vs pathogenic/benign structural comparisons from bioinformatics

3.4 Machine learning

It is clear from our data that the LOF and benign conditional probabilities monotonic for both the bioinformatics-based and molecular dynamics based features. This indicates that the probabilities are not adequate classifiers for partitioning the variants into benign or LOF classifications. Hence, we consider machine learning approaches that rely on nonprobabilistic means of classifying the data. Here we considered ML approaches including DTs and RFs, which are well suited for data sets that have a small number of training data points. In Fig. 2 we present an example decision tree for the bioinformatics analyses summarized in the previous section.

To robustly assess the efficacy of the decision tree and other ML approaches, we present additional metrics to gauge the performance of this machine learning algorithm. These include the F1 score Our data therefore

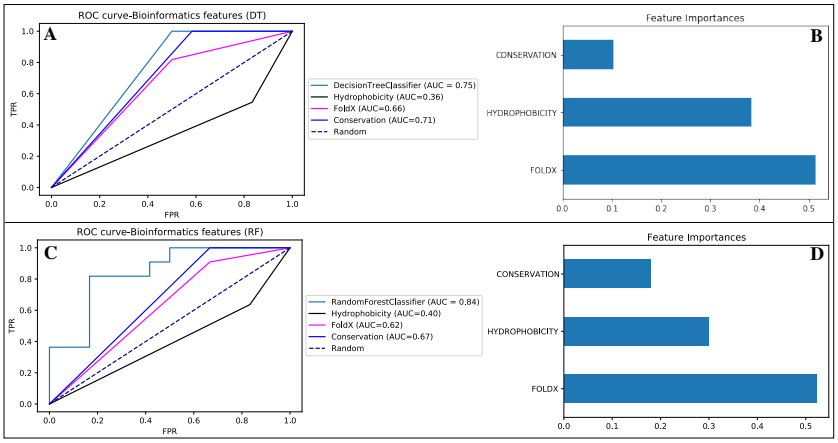


Figure 7: Application of DT and RF to bioinformatics features. Top panels are DTs (A,B) and the bottom panels (C, D) are RF. The target class is trafficking. ROC indicates the probability with which the model could predict the a variant correctly. Threshold ranges from 0 to 1. AUC indicates that RF is better than DT. 70% of the data was used for training. Hyperaparameters for DT are: criterion=entropy, random state=50, max. depth=None, min. samples leaf=1. Hyperaparameters for RF model are: bootstrap=True, # estimators=10,000, criterion=entropy, max. depth=None, max. features=2, max. leaf nodes=None, min. samples leaf=1, min. samples split=2, random state=50. Maximum two features are drawn randomly for splitting a node in RF model. ROC curves for individual features were generated by fitting them individually against the target data.

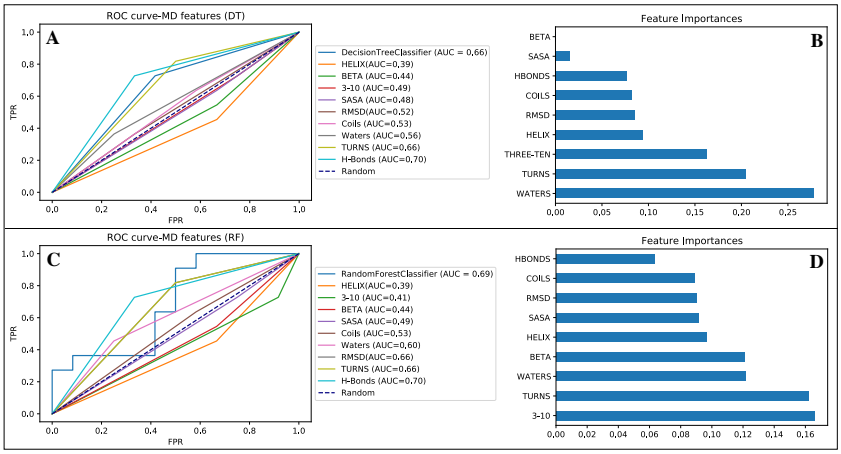


Figure 8: Application of DT and RF to MD features. Top panels are DTs (A, B) and the bottom panels (C, D) are RFs. The target class is trafficking. 70% of data was used for training in both models. Random seed was used to split the data splitting for training/testing and for fitting the models, which enables reproducibility . For DT the hyperparmeters are: criterion=entropy, max depth=None, max. leaf nodes=None, min. samples leaf=1, min. samples split=2, max. features=3, and random state=50. When all features were used for splitting at a node, DT are heavily biased towards certain features and ignoring the rest (data not shown). Therefore, we have used a maximum of three features for splitting the data at a given node. For RF method the hyperparameters are: bootstrap=True, # estimators=10,000, criterion=entropy, max depth=None, max. features=3, max. leaf nodes=None, min. samples leaf=1, min. samples split=2, random state=50. Any tree can only use the maximum three features in RF model while splitting at a node, and these three features are randomly drawn.

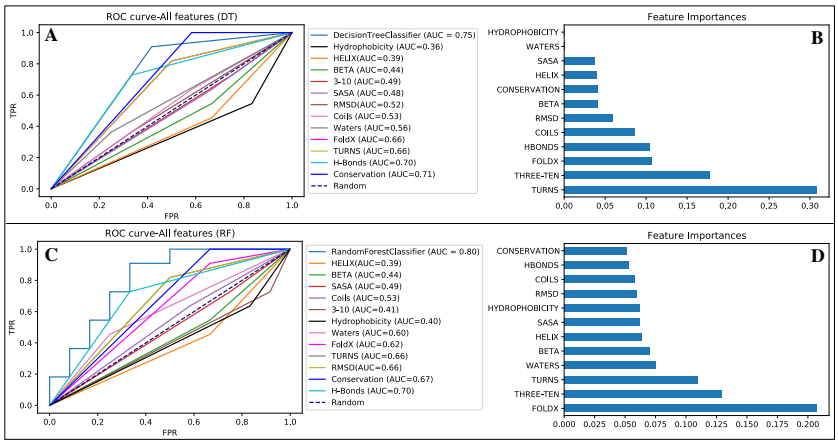


Figure 9: Application of DT and RF to all (MD + bioinformatics) features. Top panels are decision trees (A, B) and the bottom panels (C, D) are random forest. The target class is trafficking/non-trafficking. 70% of data was used for training the data in both models. Random seed was used to increase reproduciblity for both data splitting for training/testing and for fitting the models. For DT method the hyperparmeters are: criterion=entropy, max. depth=None, max. leaf nodes=None, min. samples leaf=1, min. samples split=2, max. features=4 and random state=50. When all features are used to train a model while splitting at a node, DT is only using certain features and ignoring the rest. Therefore, here we have used the maximum features of four while splitting at a node. For RF method, the hyperparmeters are: bootstrap=True, # estimators=10,000, criterion=entropy, max depth=None, max. features=4, max. leaf nodes=None, min. samples leaf=1, min. samples split=2, random state=50. Any given tree can only use the maximum four features in RF model while splitting at a node, and these four features are randomly drawn.

suggest that among the algorithms considered, the random forest presents the most robust results.

We next determined which data sets and methods could provide the most accurate classification. We applied the ML approaches using bioinformatics data alone, molecular dynamics data alone, and both sets of data and report the corresponding F1 scores in Fig. 10. These data show that molecular dynamics alone provides rates of XYZ and Z using the decision tree, compared to for the random forest and SBM respectively. The accuracy of the DT model on the test data when applied to the MD features, bioinformatics scores, and the combination of scores were 67% (Fig. S1), 83% (Fig. S1), 97% (Fig. S1). In contrast, consideration of only the bioinformatics data improves the F1 score to . Combining both the MD and bioinformatics data decreases the F1 score to . This suggests that some features are unreliable for classification.

To ascertain which features afford the greatest prediction accuracy, we report in Fig. 7, Fig. 8, and Fig. 9 the relative contribution of each feature to the F1 score. Here we find that the conservation score, foldX energies, total number of waters around the protein, and the total number of intra-domain hydrogen bonds are the most predictive. We then utilize only those features with the DT and RF approaches to test the predictive accuracy. ...

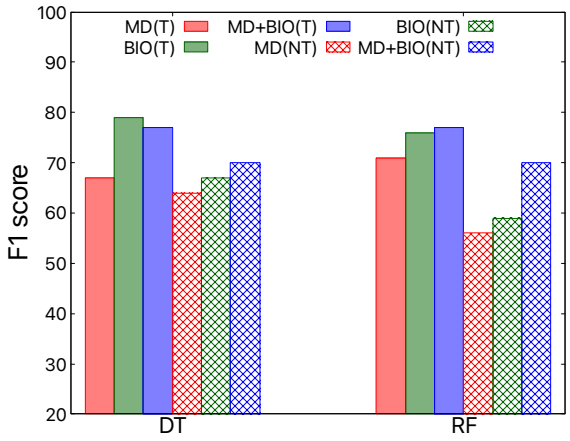


Figure 10: Comparing performance of DTs and RFs on the three data sets by F1 score. MD, bioinformatics, and a combination of both features are colored purple, green, and blue, respectively. F1 score is a weighted average of recall and precision. T and NT refer to trafficking and non-trafficking, respectively. Target class is trafficking. MD improved the performance of both the ML models for predicting the non-trafficking variants .

4 Conclusions

In this study, we discuss the implementation of a computational protocol comprising of MD simulations and machine learning techniques to predict the trafficking deficient variants of hERG PAS domain. We also tested if the addition of structural features obtained from MD simulations to traditionally used bioinformatics scores/features enhances the predictive power of the most commonly used ML techniques such as DT and RF. We have simulated 75 variants of PAS-domain, for which the experimental data (solubilities and trafficking) are available via all atom MD. Overall, we have generated 100 μs of MD data and extracted nine different features from these trajectories. Additionally, we have collected three other features such as conservation scores, change in hydrophobicity and foldX energies. We have trained both the ML algorithms on three different sets of data: (1) MD data set, (2) bioinformatics scores, and (3) the combination of MD and bioinformatics scores. Furthermore, we evaluated their predictive power on the remaining test data and we report that both the ML algorthims trained on the MD data set alone offered ≈65% accuracy. Particularly, MD features when used along with the bioinformatics scores help improve the identification of non-trafficking class of variants (55 vs. 70% accuracy). The features that proved to be most critical for the classification of PAS variants for their trafficking defects are foldX energies, conservation scores, number of intra-domain hydrogen bonds, and number of waters. To our knowledge this is the first computational study incorporating MD derived features into a ML protocol, particularly in this class of proteins, with a goal to estimate the predictive power of structural factors.

Table 1: Comparison of DT and RF performance for classifying PAS variants

Metric	DT		RF	
	NT	T	NT	T
Precision	88	67	88	67
Recall/sensitivity/TPR	58	91	58	91
F1 score	70	77	70	77
AUC	75		80	
Train accuracy	100		100	
Test accuracy	74		74	

All values are in percentages. DT and RF refers to decision tress and random forest, respectively. Target class is trafficking. NT and T refer to non-trafficking and trafficking, respectively. The data set includes all features.

5 Methodology

5.1 MD simulations

We have used the crystal structure of the isolated PAS domain (PDB:4HQA [7]) to set up the MD simulations. CHARMM-GUI web server [14] solution builder was used to build the simulation system. Hydrogens were added and then the protein was placed in a TIP3P [15] rectangular water box size size of 69 Å × 69 Å × 69 Å with an edge distance of 15 Å on all sides. 0.15 M kcl was added to the water box to neutralize the system. The total # of ions in the system were 27 potassium and 27 chlorides. The total number of atoms in the system was 30,495. Total 80 variants were simulated via all atom MD.

Each system was energy minimized for 100,000 steps using a conjugate gradient algorithm [16], and further equilibrated in the NVT ensemble for 5 ns. Production simulations were conducted under periodic boundary conditions in NPT ensemble without restraints. Each system on average was simulated for 1.5 μs (3 replicas for each system, 500 ns/replica). All systems were simulated with amber16 [17] and amber force field parameters (ff12SB) [18] were used for all components. A 310 K temperature was maintained using a Langevin thermostat and a 1 atm pressure was maintained by Nose-Hoover Langevin piston method [19, 20]. The cut-off for nonbonded interactions was 12 Å and the long-range electrostatics were treated using the particle mesh Ewald (PME) method [21]. Trajectories were saved every 20 ps. Hydrogen bonds were constrained with the SHAKE algorithm [22]. All simulations were carried on a local GPU cluster and XSEDE resources.

5.2 Feature matrix generation

5.2.1 MD Features

Data analysis was conducted using VMD plugins [23] two data points per ns were used for analysis.

5.2.2 Bioinformatics scores

We used several techniques to quickly approximate the features available for machine learning algorithms.

- FoldX scores: The impact of mutation on domain stability was estimated via FoldX webserver[24]. Change in free energy relative to the wild type ($\Delta\Delta G$) was extracted from the FoldX webserver.
- Conservation scores: Conservation scores were estimated via ConSurf Server[25] to assess the evolutionary rate (conservation) of each residue. Positive and negative values represent high and low conservation, respectively. Normalized scores are considered. Conservation score for a residue at a particular position is similar irrespective of the mutation. For example, S26I and S26Y will have similar conservation scores.
- Hydrophobicity scores: Hydrophobicity scores of residues were estimated via the eisenberg and weiss scale [26, 27]. Change in hydrophobicity upon mutation of a residue were estimated compared to the WT. Positive and negative values represent the increase and decrease of hydrophobicity, respectively.

5.3 Machine learning methodology

Our ML approach has four stages (Fig. 2) and are explained in detail below.

Step 1: In this stage, we developed a descriptor matrix by collecting all meaningful features. First, we have simulated all variants via all atom MD Next, we have extracted all possible features from the MD simulations. Additionally, bioinformatics scores were estimated from the web servers. This is a SML approach, hence it is mandatory that we know the target label. The target labels are pathogenic or benign. No data transformation was done, which is the added advantage of the two methods I am proposing for this study. There were no missing values in the data set. We do not have to worry about the data distribution due to the nonparametric nature of these methods. Categorical variables were converted to numerical values. Data was balanced by upsampling[28] the benign variants.

Step 2: In this stage, models were trained on the collected features. For this purpose, the data set was split into training and test sets. 70% of the data was used for training and 30% for testing[29]. Starting with the base model, the hyperparameters were optimized in an iterative manner to optimize the models. Entropy will be used as a selection measure[30]. The training models were assessed by its performance on the training data.

Step 3: The models that were trained in stage 2 were evaluated for their quality on the test data. Their performances were evaluated via various metrics such as confusion matrix, accuracy, precision, recall, and F1 scores. Please see references[31, 32] for a more detailed explanation of these terms. The best metric for a problem will depend on the nature of the problem. For instance, if we are concerned about false negatives, then precision is the appropriate metric, whereas if false positives are a concern, recall scores will be used to evaluate the model performance. If one cannot decide between precision and recall, then F1 scores can be

used[32]. While dealing with medical issues, specificity and sensitivity are the best measureS[31]. Specificity and sensitivity of models can also be compared by area under ROC curve (AUROC) estimations[33]. Stages 2 & 3 will be repeated in an iterative manner. Feature importance or weightage of each feature was estimated as well to get an understanding of their significance or predicting power in this classification. This helps to eliminate nonsignificant or redundant data from the dataset and helps to reduce the model complexity.

Step 4: In this stage, the optimized ML models from stages 2 & 3 will be deployed to predict the outcome of a VUS.

Scikit-learn ML library[34], numpy and pandas were used.

References

[1] M. Ono et al. "Long QT Syndrome type 2: emerging strategies for correcting class 2 KCNH2 (hERG) mutations and identifying new patients". In: *Biomolecules* 10.8 (2020), p. 1144.

[2] C. L. Anderson et al. "A rapid solubility assay of protein domain misfolding for pathogenicity assessment of rare DNA sequence variants". In: *Genetics in Medicine* 22.10 (2020), pp. 1642–1652.

[3] K. Immadisetty and J. D Madura. "A review of monoamine transporter-ligand interactions". In: *Current computer-aided drug design* 9.4 (2013), pp. 556–568.

[4] M. van de Locht et al. "Pathogenic variants in TNNC2 cause congenital myopathy due to an impaired force response to calcium". In: *The Journal of Clinical Investigation* (2021).

[5] K. Immadisetty, B. Sun, and P. M. Kekenes-Huskey. "Structural Changes beyond the EF-Hand Contribute to Apparent Calcium Binding Affinities: Insights from Parvalbumins". In: *The Journal of Physical Chemistry B* 125.24 (2021), pp. 6390–6405.

[6] K. Immadisetty, J. Alencicks, and P. Kekenes-Huskey. "Modulation of P2X4 pore closure by magnesium, potassium, and ATP". In: *bioRxiv* (2021), pp. 2021–05.

[7] X. Tang, J. Shao, and X. Qin. "Crystal structure of the PAS domain of the hEAG potassium channel". In: *Acta Crystallographica Section F: Structural Biology Communications* 72.8 (2016), pp. 578–585.

[8] D. Soni. *Supervised vs. Unsupervised Learning*. July 2020.

[9] P. M. do Nascimento et al. "A decision tree to improve identification of pathogenic mutations in clinical practice". In: *BMC medical informatics and decision making* 20.1 (2020), pp. 1–11.

[10] J. Tang et al. "Application of machine-learning models to predict tacrolimus stable dose in renal transplant recipients". In: *Scientific reports* 7.1 (2017), pp. 1–8.

[11] D. L. Yse. *The Complete Guide to Decision Trees*. en. Apr. 2019.

[12] T. Yiu. *Understanding Random Forest*. en. Sept. 2021.

[13] *Machine Learning Random Forest Algorithm - Javatpoint*. en.

[14] S. Jo, T. Kim, and W. Im. "Automated builder and database of protein/membrane complexes for molecular dynamics simulations". In: *PLoS One* 2 (2007), e880.

[15] W. L. Jorgensen et al. "Comparison of Simple Potential Functions for Simulating Liquid Water". In: *J. Chem. Phys.* 79 (1983), pp. 926–935.

[16] J. K. Reid. "On the Method of Conjugate Gradients for the Solution of Large Sparse Systems of Linear Equations". In: *Large Sparse Sets of Linear Equations*. Ed. by J. K. Reid. London: Academic Press, 1971, pp. 231–254.

[17] D. Case et al. "Amber 16". In: *University of California: San Francisco, CA* (Apr. 2016).

[18] D. Case et al. "Amber 12". In: *University of California: San Francisco, CA* (2012).

[19] G. J. Martyna, D. J. Tobias, and M. L. Klein. "Constant pressure molecular dynamics algorithms". In: *J. Chem. Phys.* 101.5 (1994), pp. 4177–4189.

[20] S. E. Feller et al. "Constant pressure molecular dynamics simulation: The Langevin piston method". In: *J. Chem. Phys.* 103.11 (1995), pp. 4613–4621.

[21] T. Darden, D. York, and L. Pedersen. "Particle mesh Ewald: An Nfffdfffdfffdlog(N) method for Ewald sums in large systems". In: *The Journal of chemical physics* 98.12 (1993), pp. 10089–10092.

[22] J.-P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen. "Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes". In: *Journal of Computational Physics* 23.3 (1977), pp. 327–341.

[23] W. Humphrey, A. Dalke, and K. Schulten. "VMD: visual molecular dynamics". In: *J. Mol. Graphics* 14 (1996), pp. 33–38.

[24] J. Schymkowitz et al. "The FoldX web server: an online force field". In: *Nucleic acids research* 33.suppl_2 (2005), W382–W388.

[25] F. Glaser et al. "ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information". In: *Bioinformatics* 19.1 (2003), pp. 163–164.

[26] D. Eisenberg et al. "Hydrophobic moments and protein structure". In: *Faraday Symposia of the Chemical Society*. Vol. 17. Royal Society of Chemistry. 1982, pp. 109–120.

[27] D. Eisenberg, R. M. Weiss, and T. C. Terwilliger. "The hydrophobic moment detects periodicity in protein hydrophobicity". In: *Proceedings of the National Academy of Sciences* 81.1 (1984), pp. 140–144.

[28] T. Boyle. *Methods for Dealing with Imbalanced Data*. en. Feb. 2019.

[29] A. Roche-Lima et al. "Machine learning algorithm for predicting warfarin dose in Caribbean hispanics using pharmacogenetic data". In: *Frontiers in pharmacology* 10 (2020), p. 1550.

[30] A. Hershy. *Gini Index vs Information Entropy*. en. Oct. 2020.

[31] A. Lekhtman. *Should I Look at Precision & Recall OR Specificity & Sensitivity?* en. Jan. 2021.

[32] *Confusion Matrix for Machine Learning*. en. Apr. 2020.

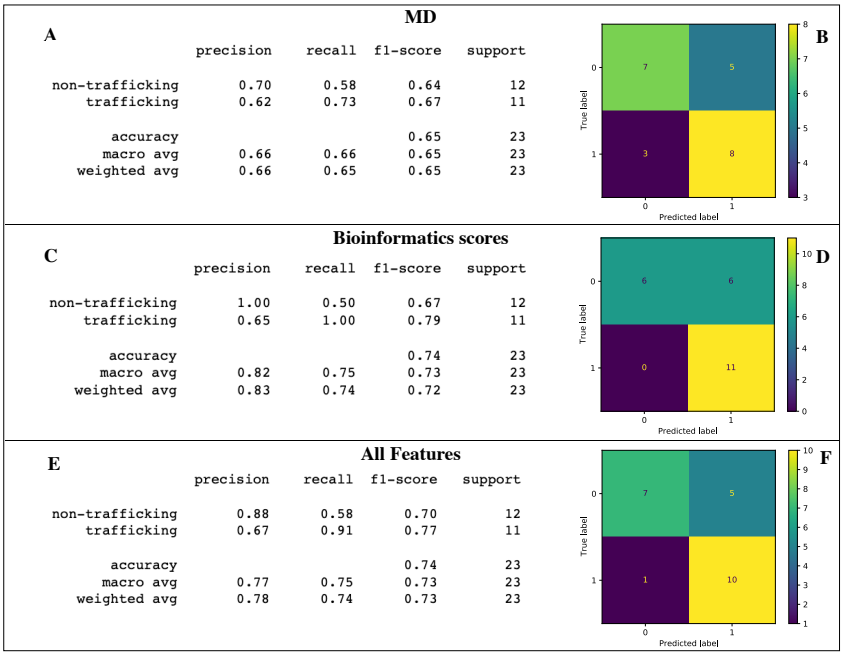
[33] M. M. Kalscheur et al. "Machine learning algorithm predicts cardiac resynchronization therapy outcomes: lessons from the COMPANION trial". In: *Circulation: Arrhythmia and Electrophysiology* 11.1 (2018), e005499.

[34] F. Pedregosa et al. "Scikit-learn: Machine learning in Python". In: *the Journal of machine Learning research* 12 (2011), pp. 2825–2830.

S1 Supplementary Information

Prediction of hERG potassium channel PAS-domain variants trafficking via machine learning

Kalyan Immadisetty and Peter M. Kekenes-Huskey*
Stritch school of medicine, Loyola University Chicago, Maywood, IL, USA 60153
E-mail: pkekeneshuskey@luc.edu



MD

A

	precision	recall	f1-score	support
non-trafficking	0.70	0.58	0.64	12
trafficking	0.62	0.73	0.67	11
accuracy			0.65	23
macro avg	0.66	0.66	0.65	23
weighted avg	0.66	0.65	0.65	23

Bioinformatics scores

C

	precision	recall	f1-score	support
non-trafficking	1.00	0.50	0.67	12
trafficking	0.65	1.00	0.79	11
accuracy			0.74	23
macro avg	0.82	0.75	0.73	23
weighted avg	0.83	0.74	0.72	23

All Features

E

	precision	recall	f1-score	support
non-trafficking	0.88	0.58	0.70	12
trafficking	0.67	0.91	0.77	11
accuracy			0.74	23
macro avg	0.77	0.75	0.73	23
weighted avg	0.78	0.74	0.73	23

Figure S1: Evaluating the DT model performance on three data sets (MD, bioinformatics and combination) of PAS domain. Shown on right are the confusion matrices that help evaluate the models performance when the true values of the test data are known. The target classes in the confusion matrices are true positive, true negative, false positive and false negative.

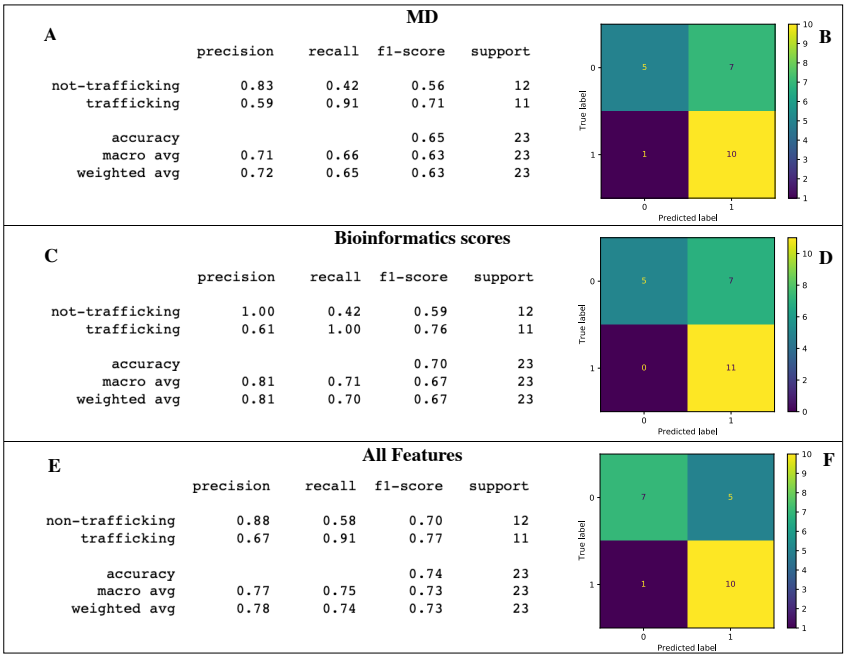


Figure S2: Evaluation of RF performance on three PAS data sets.