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

Kalyan Immadisetty[1], Xuan Fang[1], and Peter M. Kekenes-Huskey[1,*]

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

December 16, 2021

1 Abstract

Long QT syndrome (LQTS) is a cardiac condition characterized by a prolonged QT-interval. This prolongation can precipitate fatal arrhythmias despite otherwise normal metrics of cardiac function. Hence, genetic screening of individuals remains a standard for the 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 percentage of variants found in the human population are established as pathogenic. As an example, the majority of KCNH2 missense mutations are either classified as benign or are a variant of unknown significance (VUS). Here, we evaluate whether structure-based molecular dynamics (MD) simulations and machine learning (ML) can improve the accuracy of classifying Kv11.1 channel variants with normal or loss-of-function phenotypes. We focus on defective Kv11.1 trafficking, as it is the most common cause of KCHN2 loss-of-function phenotypes. Specifically, we use these computational techniques to correlate structural and dynamic changes in a Kv11.1 subdomain, the PAS-domain (PASD), with the channel's ability to traffic normally to the plasma membrane. These simulations unveiled several molecular features, including the number of hydrating waters and H-Bonds, as well as FoldX scores, that are predictive of trafficking. We then used statistical and machine learning (ML) (Decision tree (DT), Random forest (RF), and Support vector machine (SVM)) techniques to classify KNCH2 variants using these simulation-derived features. Together with bioinformatics data, such as sequence conservation and folding energies, we are able to predict with reasonable accuracy (≈75%) the KCNH2 variants that do not traffic. Therefore, structure-based simulations of KCNH2 variants localized to the Kv11.1 PAS domain lead to a significant improvement (≈10%) in classification accuracy and may ultimately complement standard genetic testing.

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 Kv7.1, Kv11.1 (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 1000[2] variants have been identified in the human population, only a small subset yielding disease phenotypes. This motivates the need for improved strategies for identifying potentially pathogenic variants from a large set of variants of unknown significance (VUS).

KCNH2 variants that encode dysfunctional Kv11.1 are attributed to a long QT syndrome subtype called LQTS. Kv11.1 is a homo-tetrameric K^+- channel comprising 155 kDa monomers[3] that contain transmembrane, Per-Arnt-Sim (PAS), and cyclic nucleotide binding (CNB) domains (see Fig. 1). Gain- and loss-of-function (LOF) variants have been characterized in each of these domains, of which the vast majority of LOF missense mutations present a defective membrane trafficking phenotype[3]. Prior studies have further classified non-trafficking variants into drug-correctable, temperature-correctable, and non-correctable trafficking subtypes. Intriguingly, a medium-throughput immunoblot assay developed in [3] revealed that LOF variants of the Kv11.1 PAS domain exhibited reduced solubility relative to wild-type and variants that trafficked normally. In that study, the authors also used bioinformatics analyses (FoldX scores[4]) to show that LOF PAS domain variants tended to exhibit unfavorable scores indicative of impaired folding thermodynamics. These findings raise the possibility that LOF PAS domain variants perturb its native structure.

To investigate this hypothesis, we performed molecular dynamics (MD) simulations of the Kv11.1 PAS domain (PASD). These simulations used the human Kv11.1 PASD structure that was solved at 1.96 Å resolution by Tang *et al* using X-ray crystallography [5]. With this approach, we considered wild type (WT) and 74 PASD-localized KCNH2 variants that were characterized by the immunoblot solubility assay developed by Anderson *et al*[3]. From these simulations, we assessed the PASD's native structure, conformational dynamics, and several physiochemical features including hydrophobicity. These features were evaluated to determine their capacity to discriminate trafficking from non-trafficking KCNH2 variants.

These molecular dynamics (MD)-derived features were used with statistical and machine learning techniques to classify known and variants of unknown significance as soluble versus insoluble as a proxy for normal and LOF trafficking phenotypes. We further complemented the MD data with bioinformatics-based features including FoldX folding scores and sequence conservation. Despite the strong bias of the training data toward variants with impaired trafficking, we found that the machine learning approaches, and in particular, the random forest approach, achieved reasonable accuracy using bioinformatics-based features alone. The accuracy was further improved by including a subset of MD-derived features that we identified. Using the trained ML models, we propose ten novel variants that we predict would yield normal and LOF phenotypes, to motivate future studies using functional assays to validate our approach.

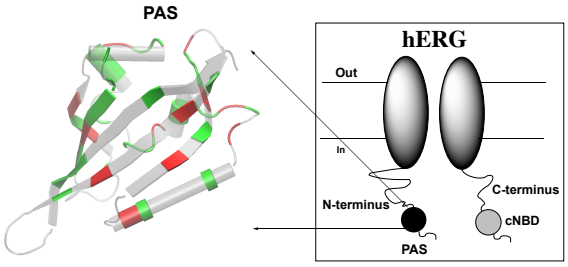

Figure 1: Schematic of hERG potassium channel is shown on the right. Structure of PAS domain on the left generated using the crystal structure (PDB: 4HQA [5]). Normal and LOF variants are shown in green and red, respectively .

3 Results

3.1 Predictive features of PAS domain solubility

3.1.1 Bioinformatics analyses of the entire variant set

We first considered 'bioinformatics-based' criteria such as the amino acid side chain hydrophobicity and volume, conservation scores for each sequence position, and FoldX scores for a given missense mutation (45 trafficking and 30 non-trafficking variants in total). The variants' raw values for each criterion were offset by the corresponding value for the WT, such that the wild-type sequence yielded a score of zero. The first two criteria were computed using the KCNH2 primary sequence. The hydrophobicity changes generally assign more negative values for apolar to charged amino acids, less negative values for mutations from apolar to polar amino acids, and positive values for charged to charged to apolar substitutions. Similarly, we include changes to reflect mutations from large (e.g., W, Y, and F) side groups into smaller groups (e.g. A). These features are intended to identify mutations that we postulated could disrupt the native packing of the PASD globular structure and promote unfolding. The conservation score reflects the degree to which a given amino acid site is evolutionarily conserved; conservation scores 1 and 9 are considered highly variable and conserved, respectively. The Foldx scores, meanwhile, utilize an all-atom force field-based empirical scoring function [4] to estimate the impact of a given variant on the free energy, G, of the mutated protein relative to the WT.

Hence, $\Delta G > 0$ signifies a destabilizing mutation. We rank order the variants according to these features in Fig. 2. The wt-shifted raw feature values are presented on the left axes, while the right axes report wt-shifted z-scores as described in Section 6.3.1 The scores corresponding to the WT are 0 in each panel. All variants are highlighted as normally trafficking versus loss of function variants with green and red bars, respectively. In the event that each classifier were perfect, loss-of-function variants would have scores that are distinct (e.g. positive) from those of the WT and normal traffickers (e.g. $z_j \leq 0$). As an example, the FoldX scores for non-traffickers tend to have z-scores above 0.5, with a few traffickers adopting similar values. In general, the Foldx scores tend to better partition the trafficking from nontrafficking, with the latter yielding more positive z-scores; however, the features were imperfect.

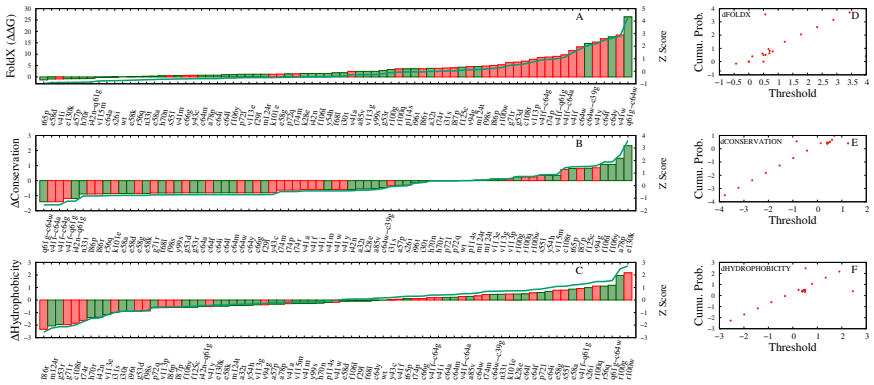

Figure 2: Bioinformatics scores compared. Y-axes reflect the change of the respective bioinformatics scores of the variants with respect to the WT. Y2-axes reflect the Z-score, which describes the distribution of the bioinformatics score of a given variant compared to the population mean. Green and red represent trafficking and non-trafficking, respectively Right panels represent cumulative probabilities as a function of threshold values.

To quantitatively assess the ability to differentiate normal and LOF (non-trafficking), we compute the conditional probability for LOF variants to adopt a z-score above some threshold value λ. These data are summarized in the second column of Fig. 6 as the nontrafficking (NT) conditional probability, $p_{ij}(NT|z_{ij} > \lambda_j)$ (see Eq. 1 in Section 6.3.1) that we abbreviate as p_j for convenience. With these data, we observe that p_j for the FoldX mostly monotonically approaches an 80% classification rate until $z_i \approx 0$, after which the probability approaches 0. In other words, variants with Foldx z-scores near 0 have an 80% true positive rate of classification as a LOF variant. Meanwhile, the p_j values based on the conservation scores also monotonically increase, but are maximized at 65%. In contrast, less predictive metrics such as the hydrophobicity score yielded p_j curves that were nonmonotonic and roughly 50% in scale. Interestingly, the amino acid volume score also exhibited a nonmonotonic trend. This contrasts with what has been reported in several studies, where amino acid volume is correlated with LOF. The features we considered are generally uncorrelated (see

Fig. S1). Hence, the P(LOF) suggest Foldx has the most predictive power for differentiating the nontraffickers from traffickers. Later we will use these probabilities to classify variants.

3.1.2 Molecular dynamics simulations of variants

We next performed molecular dynamics simulations on the 75 PAS domain variants. This was done to determine if the missense variants impacted the structure, dynamics, and energetics of the hERG PASD domain to yield additional features that could be predictive of LOF phenotypes. The features we considered include solvent accessible surface area (SASA), number of hydrating waters, number of hydrogen bonds, helicity, beta character, turns, coils, 3-10 helix, root mean squared deviations (RMSD) , root mean squared fluctuations (RMSF) and solvation energy computed by APBS. The SASA , hydrating waters and solvation energy criteria are intended to assess changes in protein solubility. The number of hydrogen bonds and secondary structure metrics (helicity, 3-10 helix, etc), reflect changes in protein folding. RMSD and RMSF assess changes in a proteins dynamic motions. In Fig. 3 we present rank ordered data for each variant with the most predictive features.

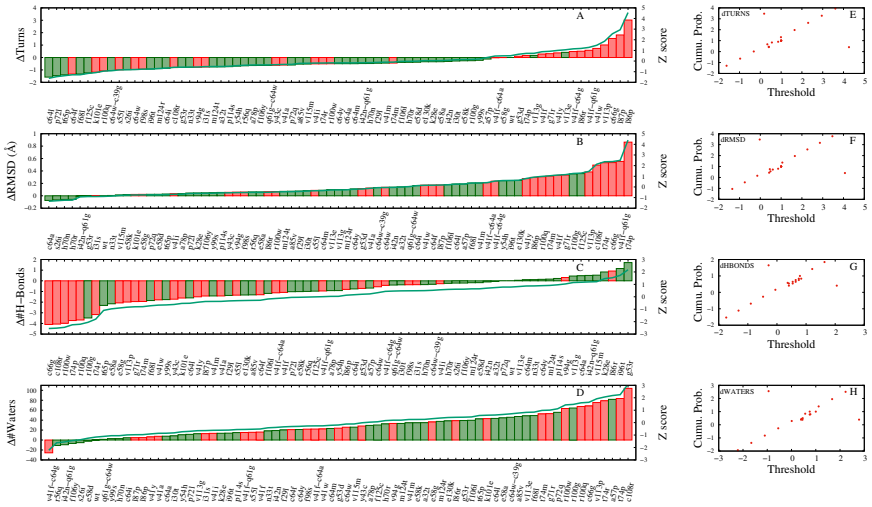

Figure 3: Turns, protein backbone RMSD, # H-Bonds, and # waters are shown in panels A, B, C and D, respectively. Waters within 4Å of the protein were calculated. See Fig. 6 for conditional probabilities .

As an improvement over the bioinformatics-based features, the structure-based features yielded $p_j \to 100\%$ for almost all features we considered except for the solvation energy and single site $SASA$ scores. We observed that p_j for the hydrating water feature was mostly monotonic, e.g. LOF variants tend to have more hydrating waters than WT reference. We additionally observed that LOF variants tend to have a lower number of hydrogen bonds compared to WT ($-\Delta Waters$) and are correlated (see Fig. 3) with nearly all traffickers restricted to the low water count, high hydrogen bond values. We anticipate that the increased number of waters in the LOF variants is attributed to the partial unfolding of the PAS structure. This apparent unfolding coincides with the variant structures having fewer hydrogen bonds than the WT. Hence, these features may improve classification beyond bioinformatics-based scores alone. Other metrics exhibited significant nonmonotonic behavior, such as for the $SASA$ and *coils* features. This nonmonotonic behavior is expected to complicate the classification of variants.

To provide context for the MD features we assessed in the previous section,we present representative structural and dynamic data for the WT, the T74P loss-of-function variant, and a normally trafficking, V113E. In Fig. 4 we superimpose three representative structures obtained from these simulations. Each variant presents a similar backbone conformation relative to the WT, which suggests all adopt roughly the same global fold. We next evaluate the features identified in the previous section 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. In Fig. 5 we highlight where the loss of hydrogen bonding occurs for the T74P variant. As anticipated, the hydrogen bond loss occurs near the site of mutation. The hydrogen bond interactions are found between T74 and other residues such as H70, G71, and I96 in the WT (Fig. 5A) are lost in the T74P variant (Fig. 5B)

We also observe that the root mean square fluctuations are generally low for the WT and V113E, with notable increases at residues 35–40, 87–93, and 117–123, which comprise loops spanning folded beta sheets

and helices. In contrast, for the T74P variant, we observe both a substantial increase in RMSF localized to the site 74 region (Fig. 4) as well as a general increase in the baseline RMSF across all residues. We were unable to directly relate these structural and dynamic changes to differences in each variant's hydrating water shell. As is made clear in the conditional probabilities plots shown in Fig. 3, the variants have diverse effects on the MD-derived features that are challenging to interpret. Namely, missense variants can have both local and global impacts on protein structure and dynamics.

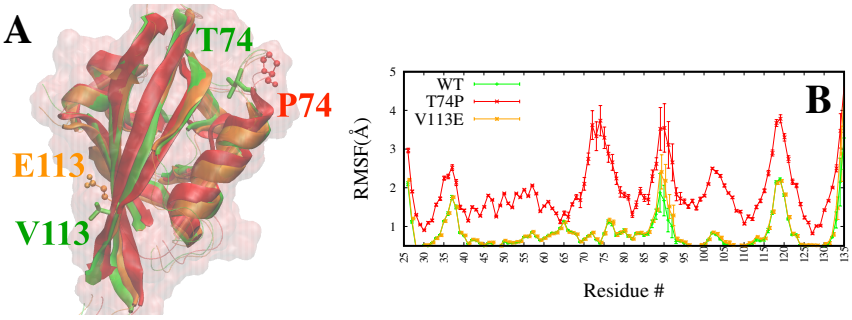

Figure 4: (A) Superposition of representative MD snapshots of WT (green), V113E (orange), and T74P (red) variants. Residues of WT are shown in the licorice representation and of the others as ball and stick models. T74/P74 is in the loop and E113/V113 is stacked between β-strands. (B) Comparison of RMSFs of WT, V113E, and T74P. T74P is pathogenic (non-trafficking) and V113E is benign (trafficking) [3].

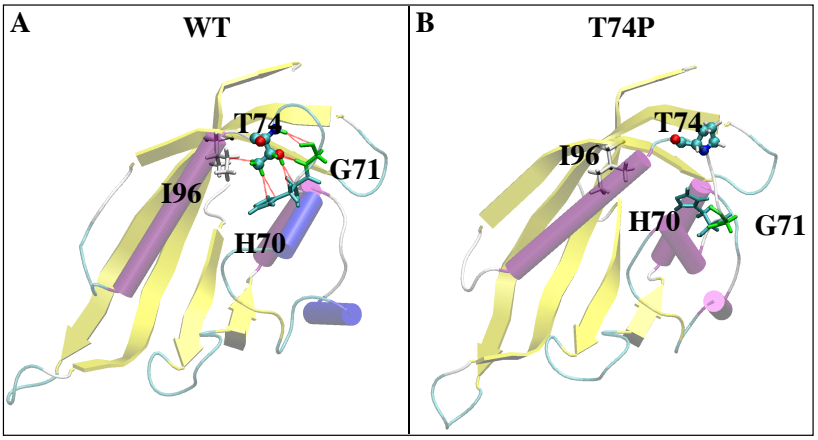

Figure 5: Comparing the hydrogen bond network near the site of mutation in WT (A) and T74P (B) variant. There are at least three hydrogen bonds that T74 forms with H70 (cyan), G71 (green), and I96 (silver/white), respectively, in the WT. Hydrogen bonds are shown as red lines in the panel A. All these hydrogen bonds are lost in the T74P variant (B). Shown are representative snapshots from the MD simulation trajectories.

3.2 Classifiers

In the previous section, we identified sequence- and structure-based features that are moderately predictive of normal- versus LOF phenotypes. However, the significant overlap for the non-trafficking and trafficking in these features limits a straightforward classification scheme. In this section, we leverage these data in aggregate to predict a variant's phenotype. We approach this by training and evaluating several classifiers that use these features to predict the phenotype in a test data set. Per standard practice guidelines in [6], we partition our data into a training set (70%, 52 entries) and a test set (30%, 23 entries). These classifiers include a probabilistic classifier, decision tree, random forests, and support vector machine, which

are generally well-suited for small data sets[7]. Performance metrics defined in Section 6.4 are included in these evaluations.

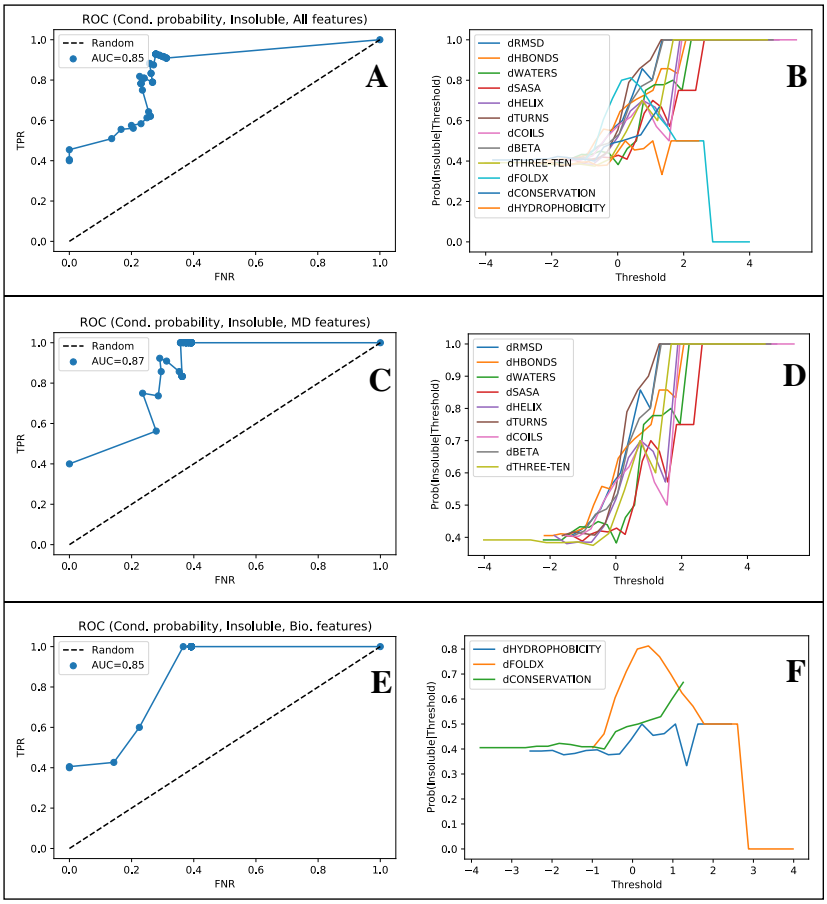

Figure 6: ROC curve displaying the probability of predicting the insoluble/non-trafficking variants using features from MD alone, bioinformatics alone, and all features .

3.2.1 Probabilistic classifier

We first construct a classifier based on the product of the conditional probabilities estimated above. This classifier assumes that each feature is statistically independent, although we recognize that some features are correlated (see Fig. S2) Variants for which $P_i(NT|\{z_{i1}, \ldots z_{in}\})$, the left-hand side of Eq. 2, exceeds a user-defined Λ are classified as non-trafficking. This probability is annotated as P_i for clarity. With this approach, we present a receiver operator characteristic (ROC) curve in Fig. 6. These ROC curves enumerate the true positives, e.g., the number of NT variants versus the total number of variants with $P_i > \Lambda$, as well as the false positive rate, e.g. the number of NT variants with $P_i \leq \Lambda$. Random features that are unable to discriminate true and false positives, i.e., 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. These data indicate that the classifier achieves very high true positive rates relative to false positive rates, as evidenced by an AUC of 0.85. As an example, many of the threshold values (Γ) in Fig. 6A achieve TPRs within 0.6 and 0.8 and have FPRs of less than 0.4. This indicates that the probabilities are reasonable classifiers for partitioning the variants into trafficking or non-trafficking classifications based on threshold alone.

We additionally provide a summary of performance metrics in Table 1. These metrics include accuracy, precision, recall, specificity, and F1-score. Of these, we focus on the precision, recall, and F1-score metrics in our analyses. The precision reflects the probability of correctly classifying positive hits, while recall assesses the completeness of the classification, e.g., classifying all positives in the data set. Our analyses indicate that

the probabilistic classifier yields precisions of 67 and 70 for NT and T, while the recall is 46 and 83. The F1 score assesses the trade-off in recall and precision performance; our strategy therefore yielded F1-scores of 50 and 74 for the NT and T classifications, respectively. Overall, the probabilistic classifier performed well with classifying variants as nontrafficking and trafficking.

Table 1: Comparison of DT and RF performance for classifying PAS variants using all features.

Metric	PC*		DT		RF		SVM	
	NT	T	NT	T	NT	T	NT	T
Precision	67	70	88	67	88	67	89	71
Recall	46	83	58	91	58	91	67	91
F1 score	50	74	70	77	70	77	76	80
AUC	89		75		80		85	
Train accuracy	73		100		100		78	
Test accuracy	67		74		74		78	

All values are in percentages. PC, DT, RF, and SVM refers to the probabilistic, decision tree, random forest, and support vector machine classifiers, respectively. Target class is trafficking. NT and T refer to non-trafficking and trafficking, respectively. The data set includes all features. *Statistics were computed using 500 bootstrapped samples.

3.2.2 Decision tree classifier

We next considered ML approaches that rely on means of classifying the data that are not reliant on probabilities. This is motivated by our observation that several of the classifiers in Fig. 6A are non-monotonic with respect to threshold. We first evaluated DTs, which are well suited for data sets that have a small number of training data points [7]. DTs hierarchically partitions the data by dividing a dataset within a given node according to a threshold value of a feature, such that the resulting partition minimizes the entropy in each 'leaf'. As an example, we provide in Fig. S8 a DT that partitions the data based on FoldX scores first, followed by hydrophobicity and conservation scores. The first node in this tree divides 52 samples based on whether a variant has a FOLDX score above or below 3.77, yielding two leaves with 5 and 13 nontrafficking variants, respectively, as well as entropies of 0.6 and 0.85, respectively. A perfect classifier would partition the variants into two leaves, each with either all traffickers or nontraffickers, yielding an entropy of zero. Based on this partitioning, the decision tree yields a TPR of 1.0 for the training set versus 0.74 for the test set.

We report in Table 1 F1 scores, recall, precision, accuracy, and AUC (i.e., derived from the ROC curve) assess the DT classification performance. The decision tree using all features yielded a precision of 88% and 67% for the non-traffickers and traffickers, respectively. Meanwhile, recalls of 58% and 91% were estimated for the two classes. Overall, this suggests that the decision tree true positive prediction rate is high and reliable for variants with a potential LOF phenotype. The recall is reduced somewhat for non-traffickers, which suggests a misclassification of positive hits. To balance precision and recall, F1 score is reported. For the DT we observed F1 scores of 70 and 77 for non-traffickers and traffickers, respectively, which suggests that the classifier is both accurate and comprehensive. Overall, these two factors yield F1 scores of 70% and 77% respectively, which represent harmonic averages of precision and recall. This suggests that the DT performance is superior to that of the probabilistic classifier.

We also computed Receiver operating characteristic (ROC) curves for decision trees trained from three feature sets: bioinformatics alone, MD alone, and all. All data sets yield AUCs greater than 0.5, which suggests the classifiers perform better than random. The classifier with the entire feature set yields an AUC of 0.75, which is comparable to that of the bioinformatics-only trained classifier. The classifier trained with MD data was less accurate (AUC=0.66). This would seem to suggest that the bioinformatics criteria are the most informative for classification. However, when we evaluate which features contribute most to the DT classification scheme (feature importances), Turns and 3-10 appeared to contribute most significantly, followed by FOLDX scores.

3.2.3 Random forest classifier

We additionally considered an 'ensembling method',RFs, which bootstrap and aggregate thousands of randomized decision trees. This bootstrapping and aggregation approach is generally found to improve classification performance and accuracy[8, 9]. This method yielded similar results as the DT as reported Table 1.. Namely, 100% training and 74% testing accuracy were achieved, with nearly identical precision, recall, and F1 scores for the nontrafficking and trafficking variants. ROC curve AUCs were modestly improved for the RF relative to the DT, with values of 0.69, 0.84, and 0.8 for the MD-features alone, bioinformatics

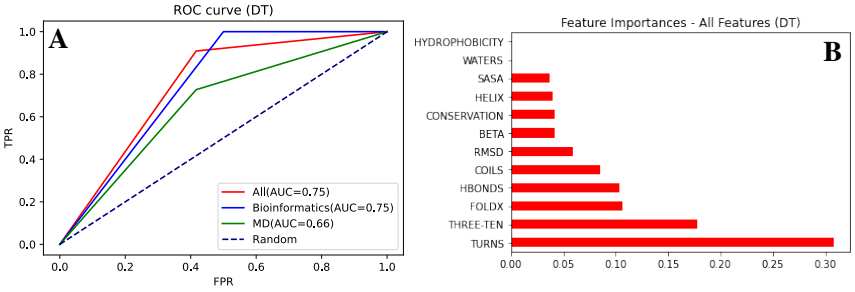

Figure 7: Comparing DT performance on MD, bioinformatics and combination of both feature sets .

features alone and the MD+bioinformatics feature set, respectively (see Fig. 8. Similar feature importances were identified, although the FoldX scores predominated overall all features (0.2), followed by the MD-based features including 3-10, Turns, Waters (0.13,0.11,0.075).

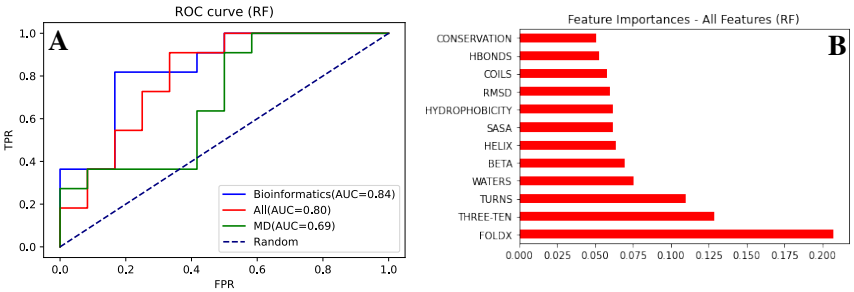

Figure 8: Comparing RF performance on MD, bioinformatics and combination of both feature sets.

3.2.4 Support vector machine classifier

Lastly, we evaluated SVMs, which map the input data into high-dimensional feature spaces that are successively partitioned into classes by hyperplanes. These partitions are typically based on polynomial functions [10]. After training SVM models with variant feature sets, we report a superior test accuracy of 78% relative to the DTs and RF approaches, despite a considerably reduced training accuracy (78%) (Table 1). Importantly, both precision and recall were also improved for the SVM relative to the tree-based models, which resulted in F1 scores of 76 and 80 for the non-trafficking and trafficking variants, respectively. In sharp contrast to the DT and RF approaches, the robust SVM performance is mostly attributed to the MD-based features. This is shown by the ROC curves in Fig. 9 that were computed for the SVMs trained from MD-only, bioinformatics-only, and combined feature sets, which yielded AUCs of 0.81, 0.11 and 0.85, respectively. Accordingly, we found that the HBOND feature importance was the highest among the features used for training, followed by hydrophobicity, 3-10 and other MD-related features.

3.2.5 Classifier performance with feature sets

Our performance analyses suggest that SVMs yield the best balance between precision and recall when both MD- and bioinformatics-derived features are used. This is demonstrated by the F1 scores shown in Fig. 10 that are marginally superior for the SVM non-trafficking and trafficking variants relative to the other models we considered (see solid blue and hashed blue bars). Additionally, regardless of the model, the F1 scores for non-traffickers are modestly lower than those for traffickers, albeit the difference is smallest for the SVM(80 versus 75). These scores also demonstrate that a combined feature set comprising MD- and bioinformatics-based metrics yields the best performance for all models and classifications, with the exception of the DT. Interestingly, models trained with the MD-only feature sets performed most poorly, especially for the SVM models applied to non-traffickers. Altogether, these data suggest the ML algorithms have better

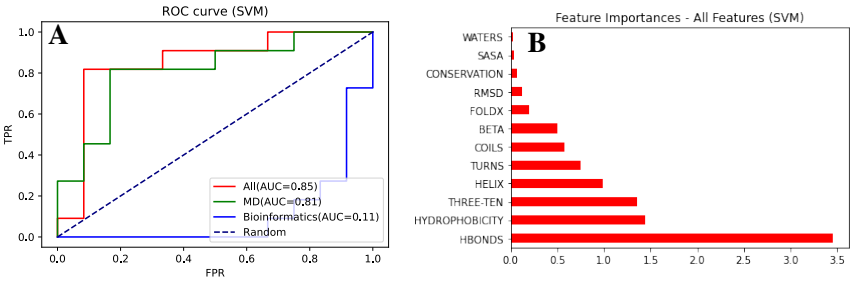

Figure 9: Comparing SVM performance on MD, bioinformatics and combination of both feature sets.

performance relative to a probabilistic classifier and that SVMs trained from the MD+bioinformatics features yield the highest F1 scores.

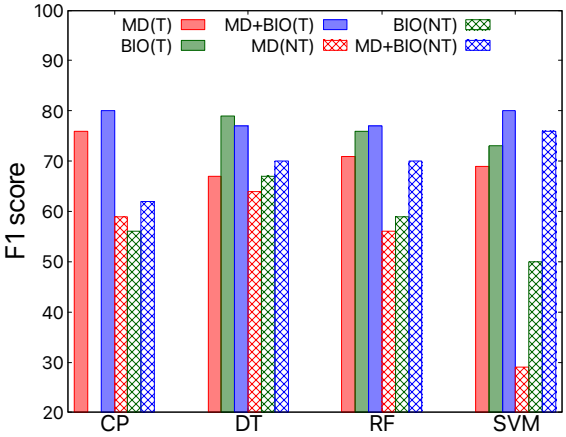

Figure 10: Comparing performance of DTs and RFs on the three data sets by F1 score. MD, bioinformatics, and a combination of both features are colored purple, green, and blue, respectively. F1 score is a harmonic mean of recall and precision. T and NT refer to trafficking and non-trafficking, respectively.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

Our analyses suggest that the ML classifiers demonstrate the strongest performance (by F1 score) when a full MD and bioinformatics feature set was used. Based on our review of the feature importances we reported for each model (for example, panel B in Fig. 7), conservation score, FoldX energies, total number of waters around the protein, and the total number of intra-domain hydrogen bonds and turns were consistently identified. We therefore retrained and tested the classifiers using those features alone to determine if a reduced feature set could improve the models' performance. Indeed, we found that the F1 scores for non-trafficking and trafficking variants generally improved . Additionally, the ROC performance for each model was also found to improve (DT: 0.83 vs. 0.80, RF: 0.91 vs 0.84, SVM: 0.85 vs. 0.89)(see Fig. S6). Overall, our analyses suggest that by using a pared feature set, the classifier performance is further improved and results in comparable performance by the RF and SVM models.

4 Discussion

4.1 Summary of Results

Machine learning approaches have been used to probe small molecule binding to the Kv11.1 channel encoded by the KCNH2 gene. One of the first studies included using machine learning approaches like DTs to predict Torsade de Pointes (TdP)-causing drugs that block Kv11.1 current [11]. More recently, approaches including deep neural network models have been used to predict abrogation of K^+ current [12] and ensemble methods for predicting cardiotoxicity of molecular compounds [13]. In a related study, six ML techniques such as linear regression, logistic regression, naive bayes, ridge regression, random forest, and neural networks were used to predict Kv11.1 related cardiotoxicity [14]. Fewer ML-based studies aiming to link KCNH2 variants to LQTS have been published. In a recent study, nearly 15 variants from the Kv11.1 PAS, CNBD, or transmembrane domains were classified by their potential to influence trafficking [15]. In addition to their in vivo approaches, the study incorporated structural information from the Wang *et al* Kv11.1a cryo-EM structure (PDB: 5VA1 [16]). To our knowledge, our study is the first to leverage an extensive set of KCNH2 variants localized to the PAS domain to train classifiers. By using an extensive feature set for tens of PAS domain variants, the predictive power of the ML models described herein is expected to improve.

ML has been extensively used to mine and interpret MD simulation data. Recent examples of these include using supervised ML techniques such as logistic regression, random forest, and multilayer perceptron (MLP) to predict interactions between the SARS-CoV-2 Spike Protein and ACE2 [17], as well as determine physiochemical factors that promote amyloid beta aggregation [18]. Refinement of computational protein structure prediction using ML+MD has also been reported [19]. With respect to variants that could impact protein function, several studies have used MD and ML to predict drug-resistant mutations in Mycobacterium tuberculosis that is involved in tuberculosis[20, 21], mutations that promote CTX-M9 mediated antibiotic resistance[22], mutations in epistatic enzyme that impact its stability and aggregation [23], and impacts of variants on protein binding dynamics [24] or inducing the phenotypic alterations [25].

We applied several methods to classify the PASD variants considered in this study. There was a general concurrence for the features deemed important for classification, as was shown in Fig. 10. The decision tree method had an advantage in yielding intuitive partitions for classifying the variants. The RF and SVM approaches were less intuitive, but were better performing. The probabilistic method performed poorly compared to the nonparametric method, which could be attributed to the nonmonotonic nature of the features.

Among the features we considered, those based on sequence alone had mixed performance, dependent on the classification model used. This may in part be due to the relatively low sequence conservation among PAS domains[3]. The strong correlation of trafficking with foldx suggests that the stability of the folded protein is of paramount importance. The apparent relationship between PASD folding stability and trafficking is supported by several features determined from MD simulation. These include the scores for $RMSF$, $HBond$, $SASA$, and $H2O$. An increase in the protein RMSF is commonly interpreted as a destabilization of the native folded structure, although it should be recognized that unfolded proteins can be entropically favorable. The decrease in hydrogen bonds, however, is likely to both decrease the favorable enthalpy of the protein and contribute to its partial unfolding. The loss of hydrogen bonds accounts for the reduction in helical character while increasing turns and 3-10 character, as well as the decrease in beta sheet character that was compensated for by random coil. These relationships are demonstrated in the scatter plots shown in Fig. S2. Commensurate with this unfolding are increases in the protein's solvent accessible surface area and thereby the number of hydrating waters.

While it is clear that our simulations of some pathogenic (non-trafficking) variants present substantial decreases in the number of hydrogen bonds relative to WT, other non-trafficking variants have intermediate reductions that are comparable to the benign trafficking variants. When other MD properties are considered, including hydration, the ability to differentiate trafficking from non-trafficking is improved. It is apparent therefore that several molecular features will need to be determined in order to have sufficient data to discriminate between phenotypes, which was the goal of our simulations. It is also important to note that solubility data by immunoblot versus trafficking status from Anderson *et al* was continuously distributed, with trafficking-deficient variants being more common at lower solubilities. Therefore, there was no clear threshold between folded and unfolded states that could be used for straightforward classification, which advocates for the machine learning approaches we considered here.

We also found that our MD-based features were mostly consistent with the interpretations outlined in Anderson *et al*. For instance, they reported lesser solvent accessibility of LQTS2-associated variants relative to other positions. Additionally, they indicated that increases in the volume of hydrophobic amino acid mutations correlated with decreased solubility, although we did not find a strong association . Our simulations, however, indicated that mutations involving proline were generally well tolerated in the folded structure, which conflicts with their suggested that prolines impose backbone strain upon folding. We also found that a limited number of variants located at the N- and C-termini are generally trafficked, which is likely because these termini are unfolded in the native structure. Hence, the MD simulations provide structure-based rationale for the potential of non-trafficking variants to misfold, as evidenced in particular by changes in

secondary structure and hydration.

Despite the apparent correlation between FoldX scores and trafficking, there are likely variants that may not negatively impact folding, but rather impair the KCNH2 PAS domain's interaction with other proteins involved in its trafficking. As an example, exposure of hydrophobic, solvent-exposed regions of the protein may constitute a recognition signal for chaperones of the endoplasmic reticulum [26]. Similarly, disruption of the protein's native folds may impact different mechanisms of trafficking quality control. These mechanisms could include auxiliary proteins involved in Kv11.1a's degradation by the proteasome, those recognizing permissive conformations of Kv11.1 necessary for its trafficking from the ER, or those mediating retrotranslocation of ER-localized Kv11.1 to or from the cytosol [27]. To our knowledge, there is a lack of structural data for evaluating this hypothesis, with exception to a PASD structure complexed with the KCNH2 cyclic nucleotide binding domain that was determined from x-ray crystallography [28].

4.2 Limitations

We discuss several limitations in our current study that could be addressed to improve the classification accuracy. The Anderson *et al*study utilized a solubility assay as a surrogate for channel trafficking. Ostensibly, misfolded Kv11.1 variants predisposed the proteins to assemble into insoluble aggregates, as has been observed in other misfolded proteins like amyloid beta [29]. Our study identified molecular features that were likely to predict misfolding, but not the aggregation or solubility of the assembled aggregates owing to the difficulty in modeling the latter processes [30]. Along these lines, experimental verification of misfolded proteins could provide a more direct means to link the MD-based features, such as hydrogen bonding, to the Anderson assay. For instance, melting temperatures and circular dichroism have recently been used to probe the proper assembly of PAS domain variants [31]; however, these data were only available for a small subset of the variants we considered in our study. Similarly, our analyses focused on the isolated PAS domain, whereas the determinants of efficacious trafficking may actually reside in the PAS domain's impact on the assembled Kv11.1a channel conformation. Such molecular studies are unfortunately not possible with current modeling approaches. Nonetheless, we believe our classification strategy has identified a common set of molecular features in the PAS domain that correlate with impaired trafficking of the intact Kv11.1 channel. We were also limited by the predominance of non-trafficking versus trafficking variants in our dataset. To better balance the data set, we considered both benign variants and KCNH2 variants with correctable trafficking defects as 'trafficking' (45), whereas only those with uncorrectable defects were labeled as nontrafficking (30) in our model building. This grouping yielded a more even partitioning of the data set, but likely masked more subtle features in the correctable traffickers that would be predictive of the phenotype.

5 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 Kv11.1 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[32], change in hydrophobicity, and foldX energies[4]. 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 Kv11.1 based computational study incorporating MD derived features into a ML protocol with a goal to estimate the predictive power of structural factors.

At a minimum, our strategy for classifying KCNH2 variants of unknown significance comprises a first step toward reducing the number of VUS subject to more exhaustive functional testing. Given the generality of the approach, the accommodation of new functional testing results is straightforward and in principle will improve in accuracy as more data are collected. For instance, current state-of-the-art medium-throughput electrophysiology methods are being employed by various groups to study and characterize the variants, which could accelerate the expansion of data sets for model training. In addition, structural data for the Kv11.1 protein and its intracellular domains are becoming increasingly available, which could unveil new, undiscovered molecular features that are predictive of protein function and ultimately cellular phenotype. Similarly, future studies may adopt data from proteins unrelated to Kv11.1, given the structural similarity of

PAS domains in other proteins such as EAG (ether-a-go-go) and ELK (EAG-like K+) channels[33], as well as the channels like hyperpolarization-activated, cyclic nucleotide-gated channels [34].

6 Computational Methodology

6.1 Molecular dynamics simulations

We have used the crystal structure of the isolated PAS domain (PDB:4HQA [5]) to set up the MD simulations. CHARMM-GUI web server [35] solution builder was used to build the simulation system. Hydrogens were added and then the protein was placed in a TIP3P [36] 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 [37], 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 [38]. A 310 K temperature was maintained using a Langevin thermostat and a 1 atm pressure was maintained by Nose-Hoover Langevin piston method [39, 40]. The cut-off for nonbonded interactions was 12 Å and the long-range electrostatics were treated using the particle mesh Ewald (PME) method [41]. Trajectories were saved every 20 ps. Hydrogen bonds were constrained with the SHAKE algorithm [42]. All simulations were carried out on a local GPU cluster and XSEDE resources. CHARMM36 force field was used for the entire system.

6.2 Machine learning methodology

Our ML approach comprises four stages (Fig. 11).

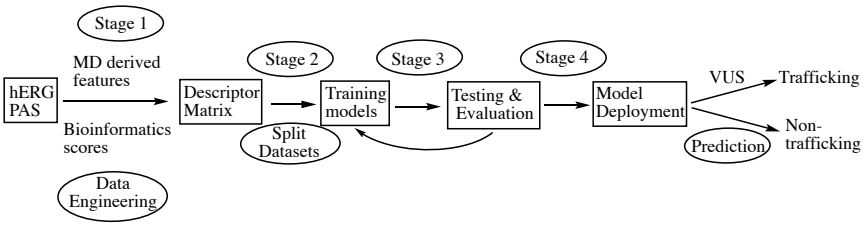

Figure 11: The supervised ML protocol adopted in this study for classifying the variants of Kv11.1 K^+ channel. In step 1, the feature set is generated, in step 2, the data will be split into training and testing sets, and the models will be trained on the training dataset, in step 3 the models will be evaluated on the test data. The trained modeled is intended to be deployed for the evaluation of novel VUS to determine their propensity for trafficking.

- **Step 1:** A descriptor matrix is assembled by collecting features for each variant. An excerpt of these data are summarized in Table S1.
- **Step 2:** In this stage, models were trained on the collected features. For this purpose, the data set was split into training (70%) and test (30%) sets, per recommendations in [43]. Model training, e.g., optimization of a given model's hyperparameters, was conducted using the native scikit-learn library.
- **Step 3:** Model performance was assessed on the test data via metrics such as confusion matrix, accuracy, precision, recall, and F1 scores. We emphasized the F1 score because it is a better metric for the imbalanced dataset compared to accuracy[44]. Steps 2 & 3 were carried out in an iterative manner to optimize the performance of the model on the test data.
- **Step 4:** Optimized ML models are deployed to predict the status of novel VUS.

Scikit-learn ML library[45], numpy and pandas were used for these steps.

6.2.1 Descriptor matrix definition from features

Bioinformatics-based features include FoldX, conservation, and hydrophobicity. The FoldX scores indicate the impact of mutation on domain stability and were computed via FoldX webserver[4]. The Change in free

energy relative to the wild type ($\Delta\Delta G$) was extracted from the FoldX webserver. Conservation scores were estimated via ConSurf Server[32] to assess the evolutionary rate (conservation) of each residue. Positive and negative values represent high and low conservation, respectively. 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 of residues were estimated via the Eisenberg and Weiss scale [46, 47]. Changes 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. Categorical variables (e.g. non-numeric) were converted to numerical values. All variants shown in (Table S1) via all atom MD. Features derived from MD were done using VMD plugins [48], for which two data points per ns were used. Solvation energy was computed using the Adaptive Poisson-Boltzman Solver (APBS) at 0.15 M NaCl with solvent and reference dielectric constants of 78.54 and 1.0, respectively. A randomized seed was used to split the data into training (70%) and testing data sets (30%). A representative excerpt of the dataset is provided in Table S1. Bootstrapping of the performance metrics described in Section 6.4 used 500 randomized splits of the training and testing data.

6.3 Summary of ML and statistical classifiers

Machine learning and statistics-based classifiers were with molecular dynamics- and bioinformatics-based feature sets. We used the following statistical and machine learning approaches as classifiers for our data.

6.3.1 Conditional probability

We collect m features that are normalized as z-scores for n variants. The variant i's z-score for its jth feature is defined as z'_{ij}. We shift the z-scores for each jth feature by the corresponding value for the wild-type structure, $z'_{wt,j}$, e.g.

$$z_{ij} \equiv z'_{ij} - z'_{wt,j},$$

such that $z_{wt,j} = 0 \; \forall j \in \{j_1, \ldots, j_n\}$. We then define a conditional probability

$$p_{ij}(NT|z_{ij} > \lambda_j), \tag{1}$$

that reflects the likelihood for a variant to be non-trafficking (NT) if its z-score z_{ij} exceeds some threshold λ_j. Eq. 1 is evaluated for all λ_j spanning the minimum and maximum values of z_{ij} and normalized such that its minimum and maximum are bounded by 0 and 1. We then define a 'total' NT probability for the variant i as the product of values from Eq. 1 that are determined for each feature j

$$P_i(NT|\{z_{i1}, \ldots z_{in}\}) \equiv \prod_j p_{ij}(NT|z_{ij} > \lambda_j). \tag{2}$$

Finally, a variant i is deemed non-trafficking if $P_i(NT|\{z_{i1}, \ldots z_{in}\}) \geq \Lambda$, otherwise the variant is classified as trafficking. Λ may be tuned to achieve the desired true and false positive rates.

6.3.2 Decision tree

This is a supervised machine learning technique (SML) [49], for which the target labels are already known, to classify the variants into trafficking/non-trafficking. We choose this method given that DTs exhibit robust performance [7], are easy to interpret, and that they are nonparametric, meaning that they make no distributional assumptions on the data[50, 51]. Disadvantages include that DTs can be inaccurate, particularly with noisy data, and suffer from overfitting [51].

We highlight our parameter choices for this model below. Entropy was used as a selection measure[52]. Hyperparameters for fitting the bioinformatics scores were:criterion=entropy, random state=50, max. depth=None, min. samples leaf=1. ROC curves for individual features were generated by fitting them individually against the target data using the functions provide with SCIKIT-LEARN.

Hyperparameters for fitting the MD models included: criterion=entropy, max depth=None, max.leaf nodes=None, min. samples leaf=1, min. samples split=2, max. features=3, and random state=50.

We observed that when all features were used for splitting at a given node with a decision tree, the DT outputs were biased toward a narrow set of features. Therefore, we have used a maximum of three features (i.e., for the MD dataset), from which one is selected for splitting the data at a given node. The hyperparameters for fitting against all feature data set were: criterion=entropy, max. depth=None, max. leaf nodes=None, min. samples leaf=1,min. samples split=2, max. features=4 and random state=50. These parameters indicate that four were the maximum number of features for splitting at a node.

6.3.3 Random forest

The RF approach is similar to DT, however, instead of fitting the data to a single DT, the models are trained on multiples of DTs, for which each DT is trained on a randomly drawn subset of data. This is repeated 10,000 times, after which the aggregate values of all DTs are estimated [8, 9] The aggregated model is used to classify the test data. In addition to the training data, subsets of features are drawn randomly to fit a given DT. Random forests are a representative ensemble method that has been shown to perform better than the DT when the hyperparameters are appropriately tuned to fit the training data [8, 9]. Other advantages include that they are less prone to overfitting than DT, can work with high dimensionality data efficiently, overcome data points with an incomplete feature set, and identify outliers [8, 9].

The hyperparameters we chose for the RF model when fitting against bioinformatics data include: 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. A maximum of two features were drawn randomly for splitting a node in RF model.

The hyperparameters used to fit RF models with md data include: 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 up to a maximum of three randomly selected features in RF model for node splitting.

Lastly, the hyperparmeters for all feature set 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.

6.3.4 Support vector machines

SVM models were constructed similar to the procedures described above. 70% of the data was used to train the classifier and 30% for the testing the trained models. Unlike the others ML methods explained above, SVM generates a hyperplane to separate the classes in the high-dimensional feature space[53]. Hyperparamters used to fit the SVM models (for all three data sets) were: gamma='auto', random state=50, max. iter=-1, kernel='linear', probability=True. A default value of 1 was used for C (also called as the regularization parameter). C and regularization are inversely proportional, i.e., greater the value of C lower is the regularization for a model[53]. Gamma is the kernel coefficient and gamma='auto' enables the use of 1/number of features for a given model[54]. Typically, a large value of gamma leads to large bias and low variance and vice versa. Max. iter = -1 indicates that there is no limit on the maximum iterations.

6.4 Model performance assessments

We utilize several metrics to assess model performance. We report the models' accuracy, which is formally defined as

$$\text{Accuracy} = \frac{\text{\# correct predictions (TP+TN)}}{\text{\# predictions (TP+TN+FP+FN)}} \tag{3}$$

where TP, TN, FP, FN are true positives, true negatives, false positives, and false negatives, respectively. The accuracy provides a simple assessment for the probability of correct positive assignments. However, this metric can yield misleading results, especially for imbalanced datasets[55], therefore we also report the precision. The precision reflects the validity of the model, that is, the percentage of positives that are true positives:

$$\text{Precision of TP} = \frac{\text{TP}}{\text{TP + FP}} \tag{4}$$

The recall, which is also frequently referred to as sensitivity, indicates the percentage of true positives that were correctly identified:

$$\text{Recall} = \frac{\text{TP}}{\text{TP + FN}} \tag{5}$$

Lastly, we report the F1-score, which is the harmonic average of recall and precision.

$$\text{F1-score} = \frac{2}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}} \tag{6}$$

The F1 score measures the balance between the precision and recall of the model and is the de facto standard for comparing ML models.

We also accompany these data with a confusion matrix, which is a N $\times$ N matrix used to evaluate the performance of a model/classifier on test data when the true values are known. N is the number of target

classes, which typically are four in the case of binary classification (false versus true). Performance was also assessed via ROC curves that report the true positive rate (TPR) and false positive rate (FPR) as a function of threshold. The Area under the curve (AUC) for each curve designates the model performance. Namely, an inefficient model for which TPR=FPR will yield an AUC of 0.5 relative to an optimal model, for which AUC = 1.0 (FPR=0% and TPR=100%) All calculations were conducted using NUMPY and the SCIKIT-LEARN package.

7 Acknowledgements

Research reported in this publication release was supported by the American Heart Association under grant number 20IPA35320141.

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] B. Oliveira-Mendes et al. "A standardised hERG phenotyping pipeline to evaluate KCNH2 genetic variant pathogenicity". In: *Clinical and Translational Medicine* 11.11 (2021), e609.

[3] 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.

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

[5] 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.

[6] Y. Xu and R. Goodacre. "On splitting training and validation set: a comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning". In: *Journal of Analysis and Testing* 2.3 (2018), pp. 249–262.

[7] 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.

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

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

[10] C. Cortes and V. Vapnik. "Support-vector networks". In: *Machine learning* 20.3 (1995), pp. 273–297.

[11] M. M. Gepp and M. C. Hutter. "Determination of hERG channel blockers using a decision tree". In: *Bioorganic & medicinal chemistry* 14.15 (2006), pp. 5325–5332.

[12] Y. Zhang et al. "Prediction of hERG K+ channel blockage using deep neural networks". In: *Chemical biology & drug design* 94.5 (2019), pp. 1973–1985.

[13] M. Liu et al. "Prediction of hERG potassium channel blockage using ensemble learning methods and molecular fingerprints". In: *Toxicol. Lett.* 332 (Oct. 2020), pp. 88–96. ISSN: 03784274.

[14] H.-M. Lee et al. "Computational determination of hERG-related cardiotoxicity of drug candidates". In: *BMC bioinformatics* 20.10 (2019), pp. 67–73.

[15] B. Oliveira-Mendes et al. "A standardised hERG phenotyping pipeline to evaluate KCNH2 genetic variant pathogenicity". In: *Clin. Transl. Med.* 11.11 (Nov. 2021). ISSN: 2001-1326.

[16] W. Wang and R. MacKinnon. "Cryo-EM Structure of the Open Human Ether-à-go-go -Related K + Channel hERG". In: *Cell* 169.3 (Apr. 2017), 422–430.e10. ISSN: 00928674.

[17] A. Pavlova et al. "Machine Learning Reveals the Critical Interactions for SARS-CoV-2 Spike Protein Binding to ACE2". In: *The journal of physical chemistry letters* 12 (2021), pp. 5494–5502.

[18] V. Gurunathan et al. "Data Mining of Molecular Simulations Suggest Key Amino Acid Residues for Aggregation, Signaling and Drug Action". In: *Biomolecules* 11.10 (2021), p. 1541.

[19] L. Heo and M. Feig. "High-accuracy protein structures by combining machine-learning with physics-based refinement". In: *Proteins: Structure, Function, and Bioinformatics* 88.5 (2020), pp. 637–642.

[20] S. Jamal et al. "Artificial Intelligence and Machine learning based prediction of resistant and susceptible mutations in Mycobacterium tuberculosis". In: *Scientific reports* 10.1 (2020), pp. 1–16.

[21] G. Mugumbate et al. "Application of Computational Methods in Understanding Mutations in Mycobacterium tuberculosis Drug Resistance". In: *Frontiers in Molecular Biosciences* (2021), p. 837.

[22] M. Latallo et al. "Predicting allosteric mutants that increase activity of a major antibiotic resistance enzyme". In: *Chemical science* 8.9 (2017), pp. 6484–6492.

[23] G. Li et al. "Machine Learning Enables Selection of Epistatic Enzyme Mutants for Stability Against Unfolding and Detrimental Aggregation". In: *ChemBioChem* 22.5 (2021), p. 904.

[24] G. A. Babbitt et al. "DROIDS 3.0—Detecting genetic and drug class variant impact on conserved protein binding dynamics". In: *Biophysical journal* 118.3 (2020), pp. 541–551.

[25] A. Garg and D. Pal. "Exploring the use of molecular dynamics in assessing protein variants for phenotypic alterations". In: *Human mutation* 40.9 (2019), pp. 1424–1435.

[26] B. Tsai, Y. Ye, and T. A. Rapoport. "Retro-translocation of proteins from the endoplasmic reticulum into the cytosol". In: *Nat. Rev. Mol. Cell Biol.* 3.4 (Apr. 2002), pp. 246–255. ISSN: 1471-0072.

[27] Q. Gong et al. "Degradation of trafficking-defective long QT syndrome type II mutant channels by the ubiquitin-proteasome pathway." In: *J. Biol. Chem.* 280.19 (May 2005), pp. 19419–25. ISSN: 0021-9258.

[28] Y. Haitin, A. E. Carlson, and W. N. Zagotta. "The structural mechanism of KCNH-channel regulation by the eag domain". In: *Nature* 501.7467 (2013), pp. 444–448.

[29] D. Flöck et al. "Misfolding of the amyloid β-protein: A molecular dynamics study". In: *Proteins: Structure, Function, and Bioinformatics* 62.1 (2006), pp. 183–192.

[30] R. L. Redler et al. "Computational approaches to understanding protein aggregation in neurodegeneration". In: *Journal of molecular cell biology* 6.2 (2014), pp. 104–115.

[31] C. A. Harley et al. "Changes in Channel Trafficking and Protein Stability Caused by LQT2 Mutations in the PAS Domain of the HERG Channel". In: *PLoS One* 7.3 (Mar. 2012). Ed. by Z. Zhang, e32654. ISSN: 1932-6203.

[32] 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.

[33] R. Adaixo et al. "Structural properties of PAS domains from the KCNH potassium channels". In: *PloS one* 8.3 (2013), e59265.

[34] T. Vaccari et al. "The human gene coding for HCN2, a pacemaker channel of the heart". In: *Biochim. Biophys. Acta - Gene Struct. Expr.* 1446.3 (Sept. 1999), pp. 419–425. ISSN: 01674781.

[35] 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.

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

[37] 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.

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

[39] 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.

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

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

[42] 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.

[43] 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.

[44] P. Huilgol. *Accuracy vs. F1-Score*. en. Aug. 2019.

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

[46] 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.

[47] 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.

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

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

[50] 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.

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

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

[53] D. Bzdok, M. Krzywinski, and N. Altman. "Machine learning: supervised methods". In: *Nature methods* 15.1 (2018), p. 5.

[54] *sklearn.svm.SVC*. en.

[55] S. A. Josephine. "Predictive Accuracy: A Misleading Performance Measure for Highly Imbalanced Data Classified negative". In: *SAS Global Forum*. 2017.

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

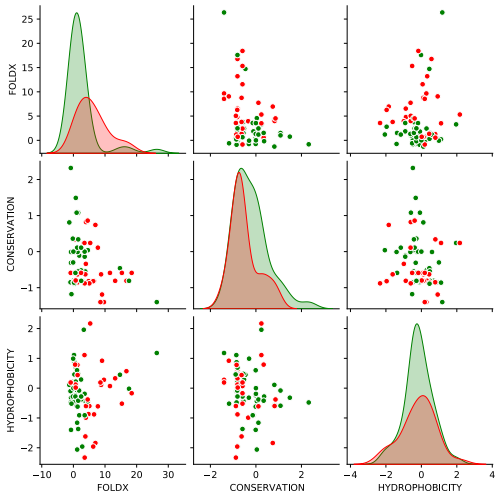

Figure S1: Bioiformatics data distribution .

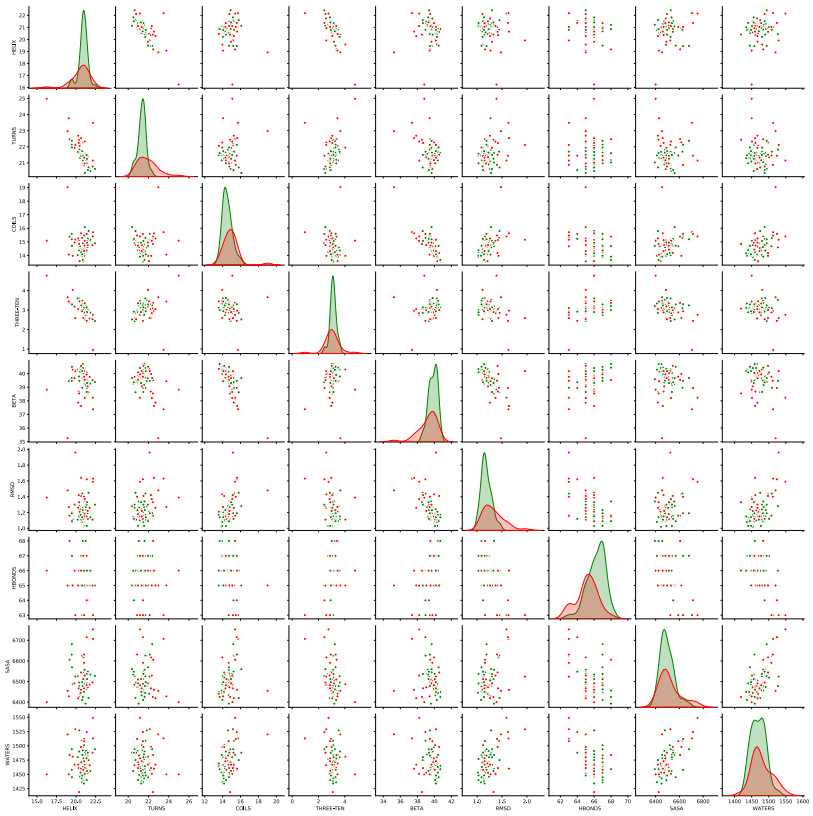

Figure S2: MD data distribution .

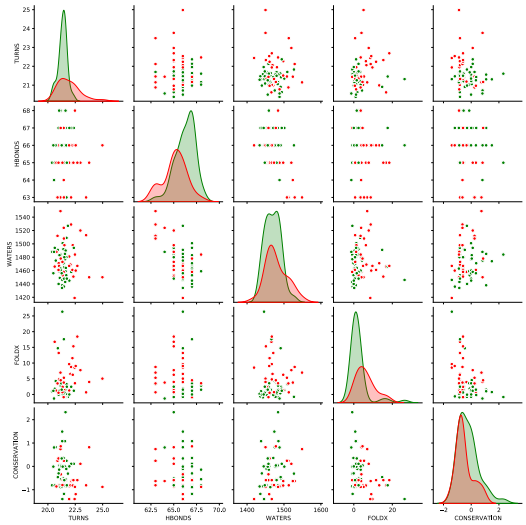

Figure S3: Distribution of best features .

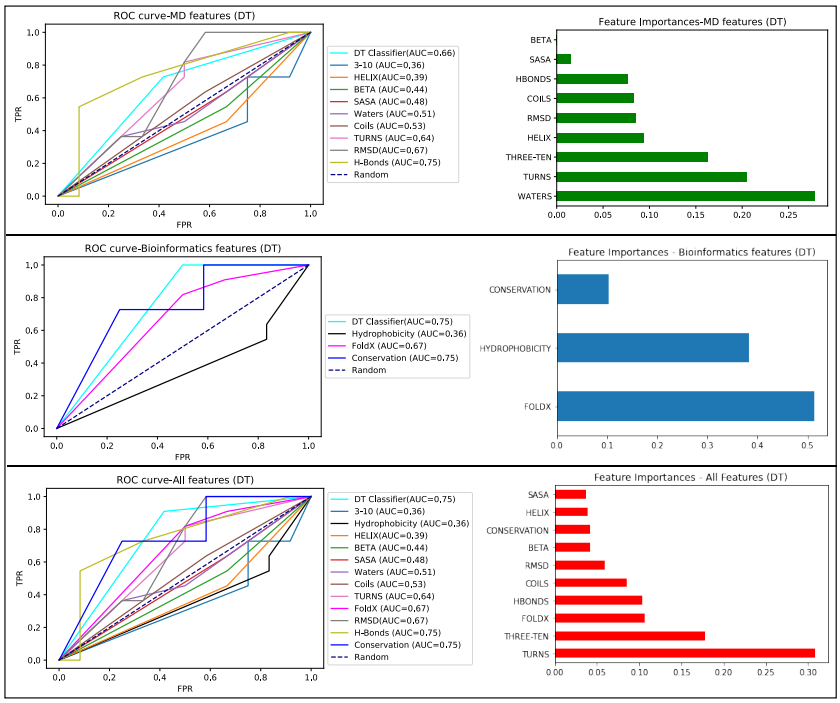

Figure S4: Comparing DT performance on MD, bioinformatics and combination of both feature sets .

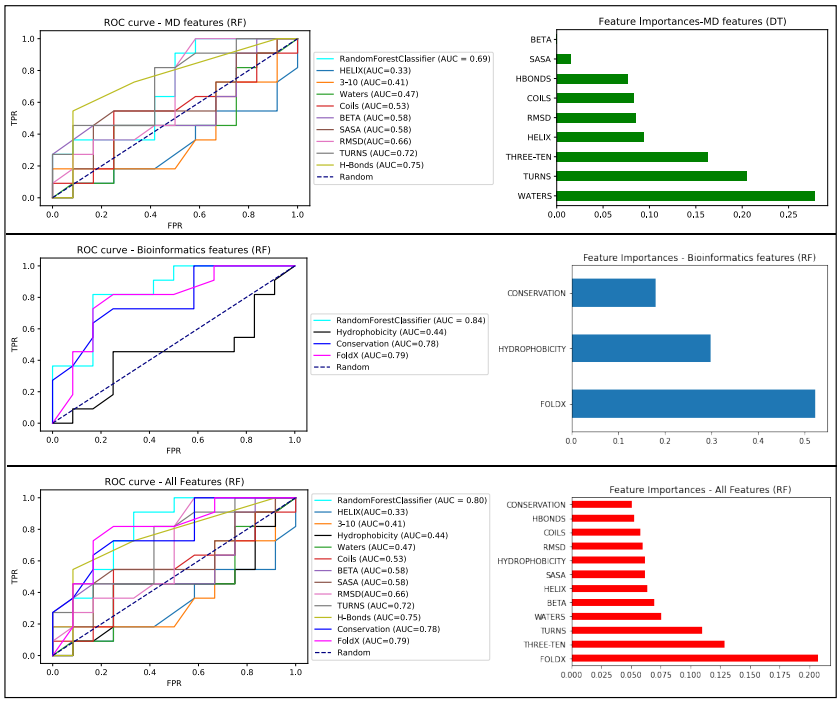

Figure S5: Comparing RF performance on MD, bioinformatics and combination of both feature sets .

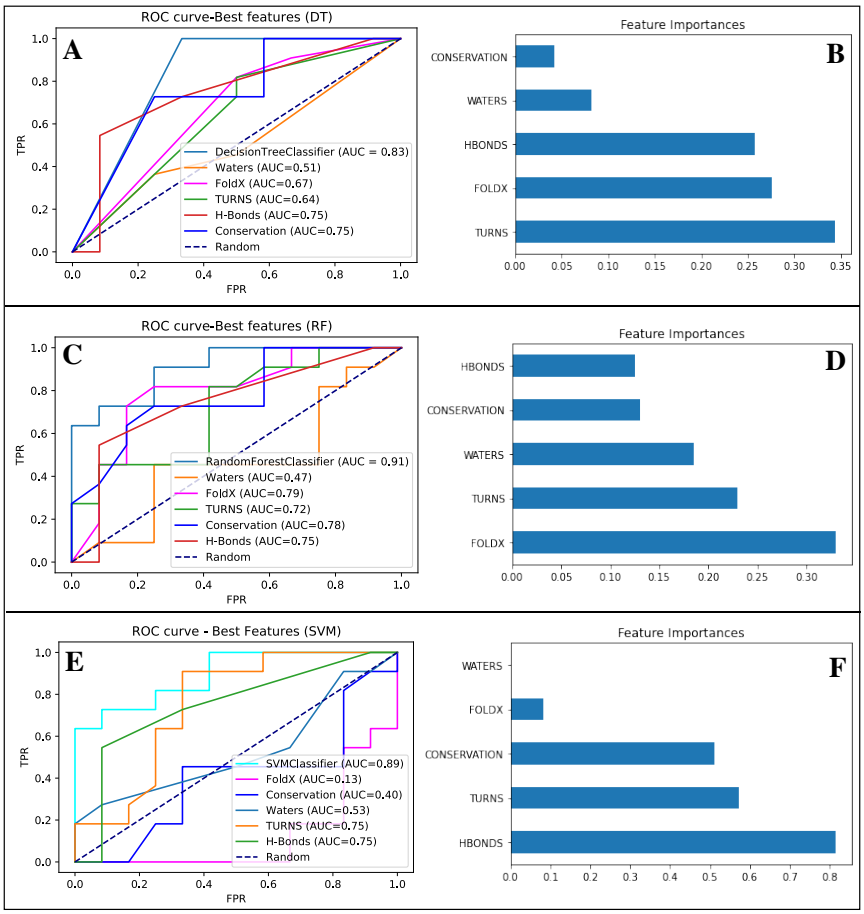

Figure S6: The best five features (conservation, H-Bonds, waters, FoldX, and turns) were used to train DTs and RFs. Top panels are DTs (A,B) and bottom panels RFs (C,D). Maximum two features were allowed to split at a node .

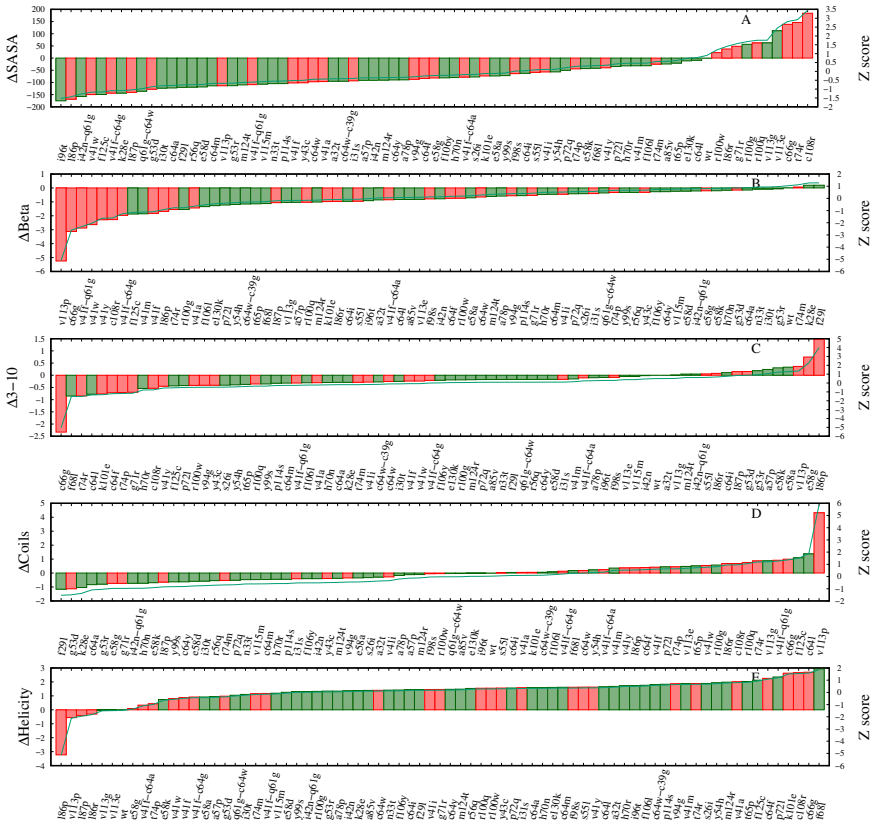

Figure S7: MD features compared .

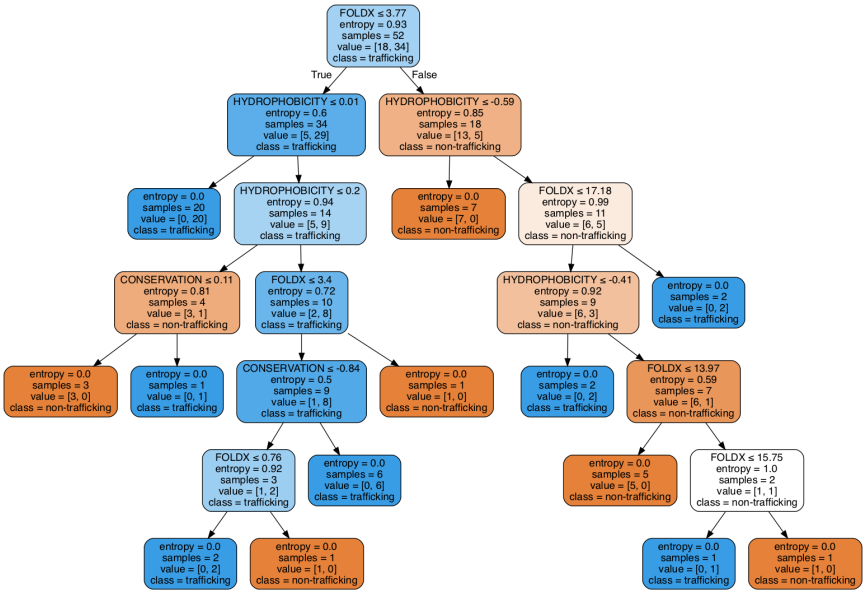

Figure S8: Decision tree applied to bioinformatics scores. Orange and blue boxes represent entropies of 0 and 1, respectively .

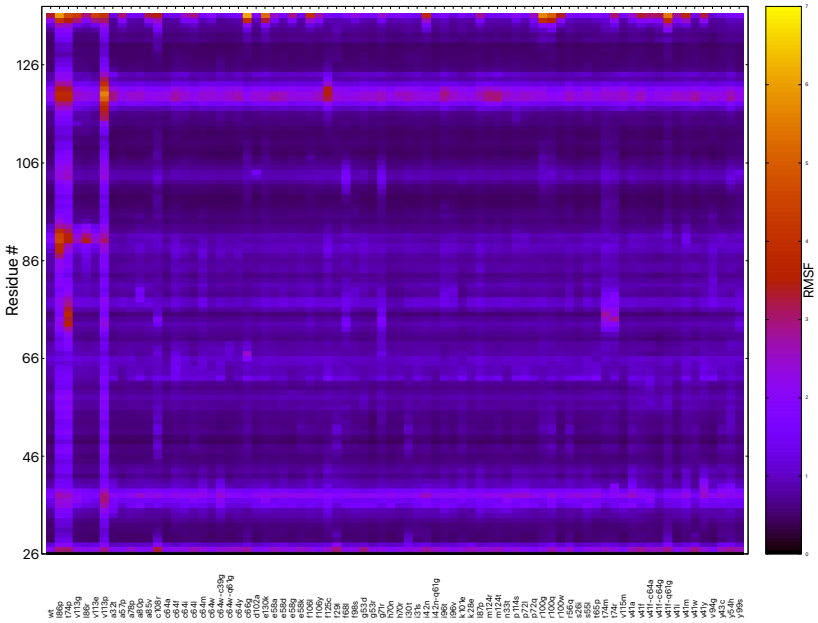

Figure S9: Comparison of RMSFs. Color bar on the right reflects the RMSF .

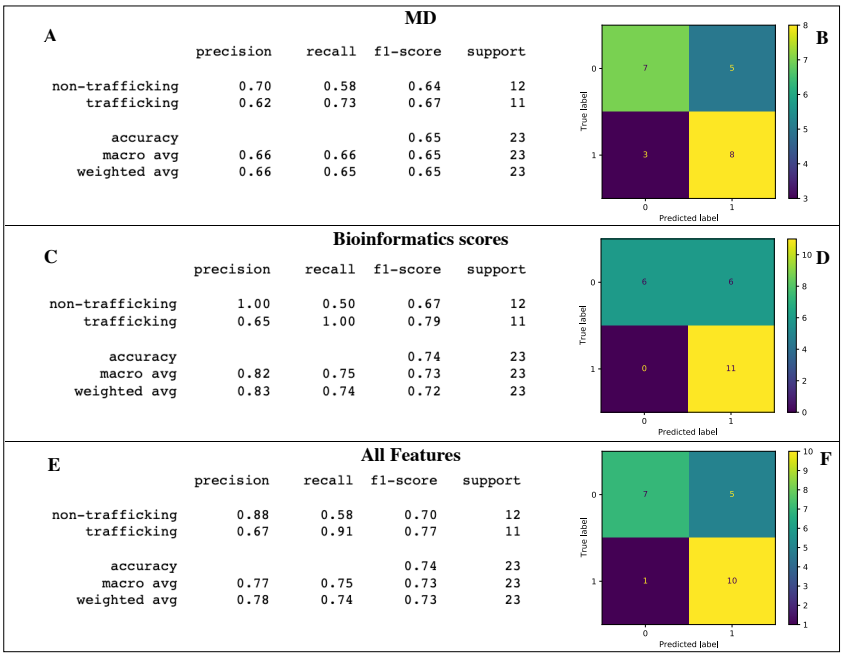

Figure S10: Evaluating the DT model performance on three data sets (MD, bioinformatics and combination) of PAS domain. Shown on right are the confusion matrices that help evaluate the models performance when the true values of the test data are known. The target classes in the confusion matrices are true positive, true negative, false positive and false negative Support defines the number of data points in the test data set .

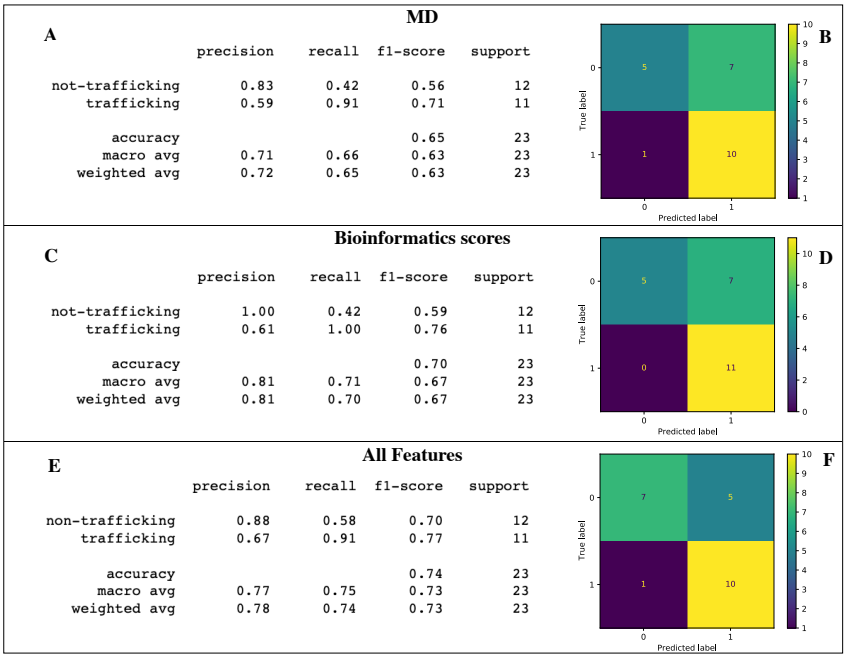

Figure S11: Evaluation of RF performance on three PAS data sets.

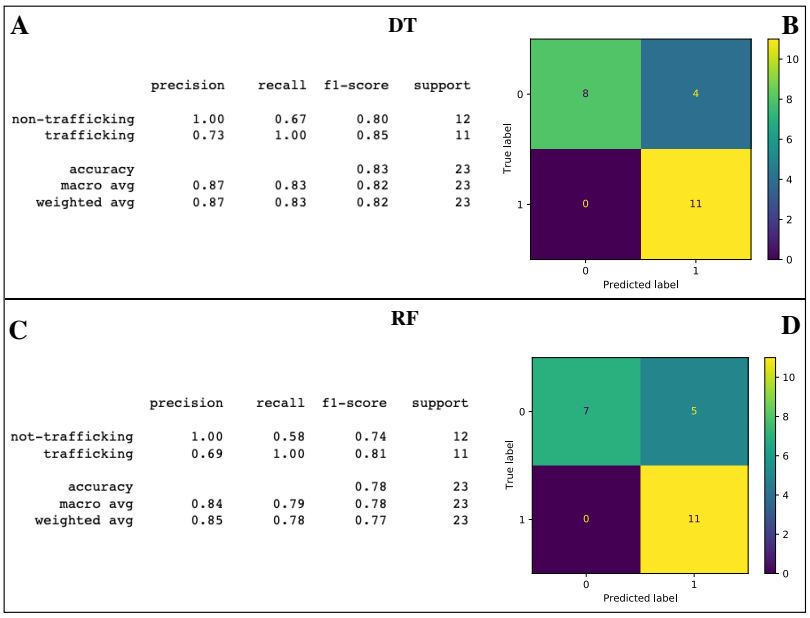

Figure S12: Performance of DT and RF on the top five features (Conservation, Foldx scores, #waters, #H-Bonds, and turns).

Table S1: All PAS variants included in this study

Variant	Class (Pathogenic)	MD	Potential benign (kal predictions)	Clinvar (pathogenic)	Anderson Variants	Class (Trafficking)
A32T	1	1			1	1
A57P	0	1			1	1
A57T		1		1		
A57V		1	1			
A78P	1	1			1	1
A80P		1			1	-1
A85V	1	1			1	1
C108R	1	1			1	0
C44F	1					
C49Y	1					
C64A	1	1			1	1
C64F	1	1			1	0
C64I	1	1			1	1
C64L	1	1			1	1
C64M	1	1			1	1
C64W	1	1			1	0
C64Y	1	1			1	1
C64W-C39G	1	1			1	1
C64W-Q61G	0	1			1	1
C66G	1	1			1	1
D102A		1			1	-1
E130K	0	1			1	1
E58A	1	1			1	1
E58D	0	1			1	1
E58G	1	1			1	0
E58K	1	1			1	1
F106L	1	1			1	1
F106Y	1	1			1	1
F125C	0	1			1	1
F29L	1	1			1	1
F68L	1	1			1	1
F98S	1	1			1	0
G47V	1					
G53D	1	1			1	0
G53R	1	1			1	1
G71R	1	1			1	0
H70N	1	1			1	1
H70R	1	1			1	1
I30T	1	1			1	1
I31S	1	1			1	0
I42L		1	1			
I42N	1	1			1	1
I42T		1		1		
I42N-Q61G	0	1			1	1
I96T	1	1			1	1
I96V		1			1	-1
K101E	1	1			1	0
K28E	1	1			1	1
L86P	1	1			1	0
L86R	1	1			1	0
L87P	1				1	0
M124R	1	1			1	1
M124T	1	1			1	1
N33T	0	1			1	1
N45Y	1					
P114S	1	1			1	1
P72L	1	1			1	1
P72Q	1	1			1	1
R100G	1	1			1	1
R100Q	1	1			1	0
R100W	1	1			1	0
R56K		1	1			
R56L		1		1		
R56Q	0	1			1	1
S26I	1	1			1	1
S55L		1			1	0
S55T		1	1			
T65P		1			1	1
T74M	1	1			1	0
T74P	1	1			1	0
T74R	1	1			1	0
V41A	1	1			1	0
V41F	1	1			1	0
V41F-C39G		1			1	
V41F-C64A	1	1			1	0
V41F-Q61G	1	1			1	0
V41F-C64G		1			1	0
V41I	1	1			1	0
V41M	0	1			1	1
V41W	1	1			1	0
V41Y	1	1			1	0
V94G	1	1			1	0
V113G		1			1	1
V113E	0	1			1	1
V113P		1			1	0
V115M	0	1			1	1
WT	0	1			1	0
Y43D		1		1		
Y43F		1	1			
Y43W		1	1			
Y43C	1	1			1	0
Y54H	1	1			1	1
Y99S	1	1			1	1

'Anderson variants' were reported in the reference[3] and supposedly have experimental solubility's reported. 'Potential benign'