**Predicting Bedaquiline Resistance Based on Known Mutations for Rv0678**

Results

test\_precision 0.836190

test\_recall 0.783333

test\_F1 0.801076

test\_AUPRC 0.804643

A gradient-boosted tree classifier was trained and optimised with mutations associated with resistance and susceptibility phenotypes. This was done to determine if mutations in *Rv0678* are useful in predicting the resistance phenotype. Features were engineered to represent various amino acid property changes due to mutation. The optimised model performed reasonably well with an area under the precision-recall curve (AUPRC) of 0.805, suggesting that information about the resistance phenotype is contained within the features.

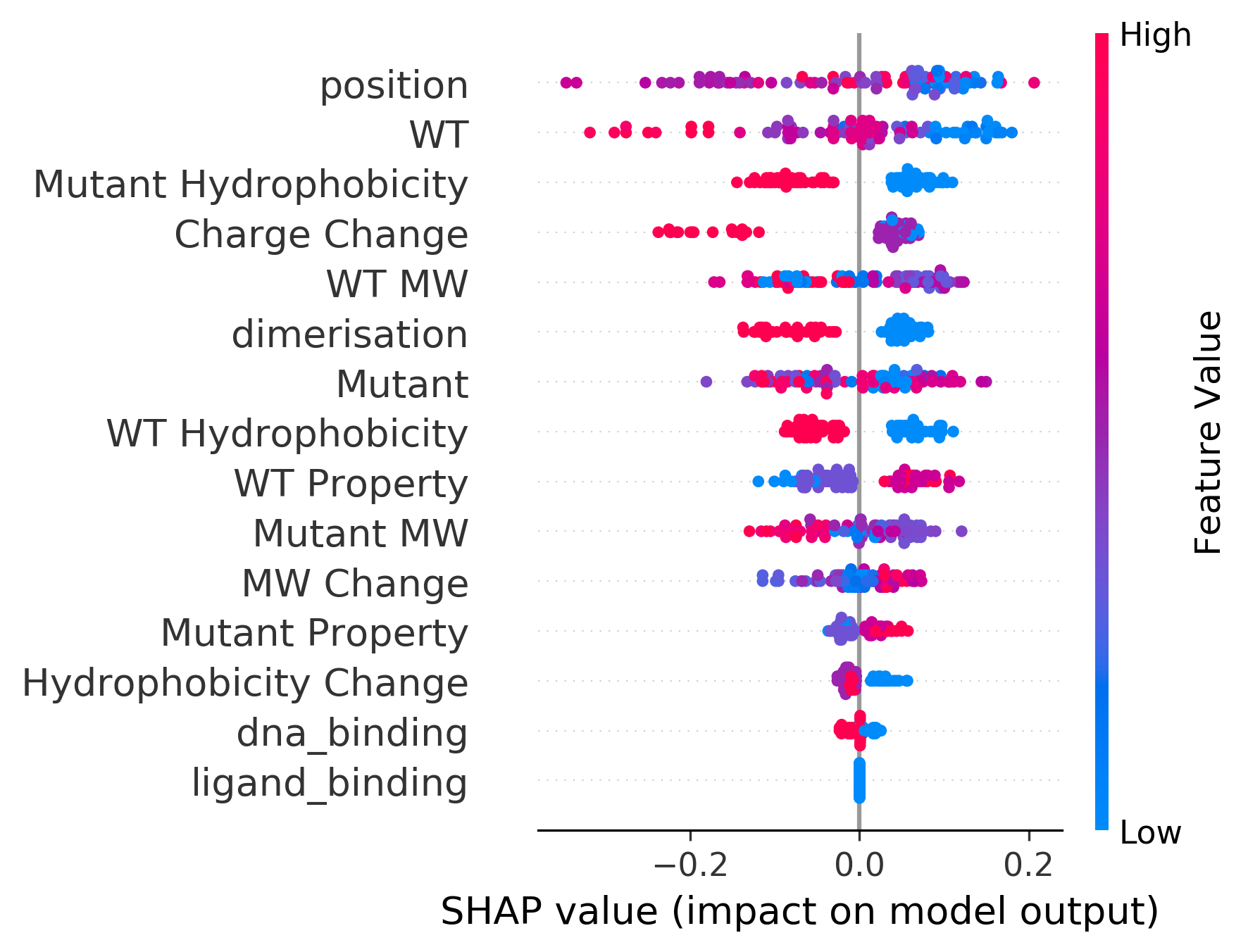
Models were interpreted using *TreeExplainer* (lundberg), which decomposes the predicted probability score of each prediction into individual feature contributions. That is, each feature is assigned a SHAP value which represents the change in predicted probability score in each prediction. The importance of features in “black box” models can thus be inferred from their mean absolute SHAP values. The results suggest that 5’ end mutations, mutations in the DNA binding domain and of polar and positively charged residues are associated with resistance. Conversely, mutations in the dimerisation domains, transitions from negatively to positively charged residues, and mutations involving hydrophobic wild-type or variant residues are associated with susceptibility.

Figure 1. Summary plot of SHAP values. Each point represents the SHAP value of a single prediction for a particular feature. Points are stacked vertically using density estimation. ‘WT’, ‘mutant’ and ‘MW’ denotes the wild-type amino acid, amino acid variant and molecular weight (Da) respectively. ‘Property’ refers to whether an amino acid was non-polar, polar, positively charged or negatively charged. ‘Ligand\_binding’, ‘dna\_binding’ and ‘dimerisation’ refer to whether the the amino acid residue is involved in ligand binding, DNA binding or dimerisation. ‘Position’ refers to the integer 5’-3’ position of the variant. Positive SHAP values imply an increase in the predicted probability of resistance due to the presence of the feature.

Methods

A gradient-boosted tree classifier was developed using the *XGBoost API* (v1.0.2) (<https://dl.acm.org/doi/abs/10.1145/2939672.2939785>) and interpreted using *TreeExplainer* as part of the *shap* API (v0.35.0) (<https://www.nature.com/articles/s42256-019-0138-9>) to determine whether the phenotypes of genomes with unknown mutations could be predicted based on the change in amino acid properties for known mutations. The task of the classifier is to predict whether a mutation results in a susceptible (0) or resistant (1) phenotype given the change in AA properties. Fifteen features were engineered based on the wild-type and mutant residues of each mutation. Two features represent the amino acid residue of the wild-type and of the mutant. Two features encode whether they are non-polar, polar, positively charged or negatively charged. Two features represent whether the wild-type and the mutant residues are hydrophobic or hydrophilic, based on the hydrophobicity scale proposed by Janin (<https://www.nature.com/articles/277491a0>). Two features encode the molecular weight of wild-type and mutant AA. Two features represent the change in charge or molecular weight from from the wild-type to mutant, where non-polar and polar residues are assumed to contribute a charge of zero. One feature represents the change in hydrophobicity where a hydrophobic→hydrophilic residue change is coded as +1 and the reverse as -1. Three features represent mutations in the DNA-binding domain, in the dimerisation domain, and to the residues in contact with 2-stearoylglycerol. The last feature represented the 5’- 3’ position of amino acid mutations. Only mutations that are associated with resistant or susceptible phenotypes were used, resulting in 59 resistance and 32 susceptibility mutations. Model parameters were optimised to maximise the F1 score and the test error was estimated using a nested, stratified, 10 x 10 cross-validation procedure. AUPRC was used for model evaluation due to class imbalance in the dataset ([10.1371/journal.pone.0118432](https://dx.doi.org/10.1371%2Fjournal.pone.0118432)). All scripts used for the analysis here is hosted on GitHub (https://github.com/cednotsed/TB-Bedaquiline-Resistance-Modelling.git).