

RESEARCH ARTICLE

Accurate prediction of functional effects for variants by combining gradient tree boosting with optimal neighborhood properties

Yuliang Pan¹, Diwei Liu¹, Lei Deng^{1,2*}

1 School of Software, Central South University, Changsha, China, **2** Shanghai Key Laboratory of Intelligent Information Processing, Shanghai, China

* leideng@csu.edu.cn



OPEN ACCESS

Citation: Pan Y, Liu D, Deng L (2017) Accurate prediction of functional effects for variants by combining gradient tree boosting with optimal neighborhood properties. PLoS ONE 12(6): e0179314. <https://doi.org/10.1371/journal.pone.0179314>

Editor: Quan Zou, Tianjin University, CHINA

Received: April 1, 2017

Accepted: May 27, 2017

Published: June 14, 2017

Copyright: © 2017 Pan et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by National Natural Science Foundation of China under grants No. 61672541 and No. 61309010, China Postdoctoral Science Foundation under grant no. 2015T80886, Specialized Research Fund for the Doctoral Program of Higher Education of China under grant no. 20130162120073 and Shanghai Key Laboratory of Intelligent Information Processing under grant no. IPL-2014-002.

Abstract

Single amino acid variations (SAVs) potentially alter biological functions, including causing diseases or natural differences between individuals. Identifying the relationship between a SAV and certain disease provides the starting point for understanding the underlying mechanisms of specific associations, and can help further prevention and diagnosis of inherited disease. We propose PredSAV, a computational method that can effectively predict how likely SAVs are to be associated with disease by incorporating gradient tree boosting (GTB) algorithm and optimally selected neighborhood features. A two-step feature selection approach is used to explore the most relevant and informative neighborhood properties that contribute to the prediction of disease association of SAVs across a wide range of sequence and structural features, especially some novel structural neighborhood features. In cross-validation experiments on the benchmark dataset, PredSAV achieves promising performances with an AUC score of 0.908 and a specificity of 0.838, which are significantly better than that of the other existing methods. Furthermore, we validate the capability of our proposed method by an independent test and gain a competitive advantage as a result. PredSAV, which combines gradient tree boosting with optimally selected neighborhood features, can return reliable predictions in distinguishing between disease-associated and neutral variants. Compared with existing methods, PredSAV shows improved specificity as well as increased overall performance.

Introduction

Single amino acid variants (SAVs) are single-base changes that result in amino acid changes of the encoded protein [1]. With the rapid development of sequencing and genomic analysis technologies, substantial SAVs between individuals have been uncovered. The 1000 Genomes project [2] and recent sequencing of whole human genomes [3–6] have provided a large number of single-nucleotide polymorphisms (SNPs), insertions, deletions and structural variants in humans. Among these variations, SAVs are recognized as the most common type in the human genome [7, 8], and some are often closely related to particular diseases [9–11].

Competing interests: The authors have declared that no competing interests exist.

According to the previous studies, SAVs may be responsible for the initiation or progression of cancer through aberrant proteins [12]. And the amino acid change can affect, for example, protein stability, interactions and enzyme activity, thereby leading to disease. Therefore, the identification of whether a SAV is neutral or disease-associated is playing an increasingly important role in understanding the underlying mechanisms of specific SAV-disease associations and developing treatment strategies for diseases.

However, experimentally determining the SAV-disease relationship of such a large number of variants is time-consuming and costly. Accurate computational approaches are vital for analysis the relationship between SAV and disease. Current prediction methods typically employ machine learning algorithms [13–16] such as neural networks [17], random forests (RF) [18] and support vector machines (SVMs) [19], and a large variety of properties, including amino acid sequence features [20], position-specific scoring matrices, residue-contact network features and 3-D structure information. This includes methods such as SIFT [21, 22], SNAP [23], Polyphen2 [24], FunSAV [25] and SusPect [26]. SIFT uses sequence homology to predict phenotypic effect based on the assumption that amino acid variants in the evolutionarily conserved regions are more likely to have functional effects [21, 22]. SNAP [23] combines multiple sequence analysis methods with neural networks to predict the functional effects of variants. Polyphen2 [24] predicts the functional impact of a variant by a Naive Bayes classifier trained using sequence, phylogenetic and structural information. FunSAV utilizes a two-stage random forest with a large number of sequence and structural properties to discriminate the SAV-disease links [25]. Yates et al. combine sequence and structural features to build an SVM classifier named SusPect to predict disease-SAV associations [26].

In this work, we develop a novel approach, termed as PredSAV, to predict the phenotypic effects of SAVs by using the Friedman's gradient tree boosting [27, 28] algorithm. PredSAV combines both sequence neighborhood features and structural neighborhood features describing not only the properties of the target residue but also the target residue's neighborhood environment. PredSAV uses a efficient two-step feature selection method to eliminate uninformative properties, which in turn improves the performance and helps to build faster and more cost-effective models. Extensive comparisons of PredSAV with other existing tools on the benchmark dataset and another independent dataset show that PredSAV significantly outperforms the existing state-of-the-art methods, and illustrate the effectiveness and advantage of the proposed approach. The framework of PredSAV is shown in Fig 1.

Materials and methods

Datasets

The dataset, taken from Wang and Song [25], are mainly gathered from the UniProt [29] human sequence variations and the Ensemble human variation database [30]. Disease-associated variants are obtained from the UniProt human sequence variations and non-Mendelian disease variants without any OMIM entry [31] are removed. Neutral variants are extracted from the Ensemble human variation database. All the SAVs emanate from the HapMap project [32]. We remove the redundant homology sequences at the sequence similarity of 40% by using CD-HIT [33]. Then a BLAST search [34] is used to map the remaining sequences to PDB structures [35]. Structures with resolutions lower than 2.5Å are removed. Ambiguity or invalid SAV-disease associations and neutral variants are deleted. Obsolete PDB structures, such as 1VJJ and 2HR0, are removed. Eventually, the dataset consists of 670 proteins containing 1006 disease-associated and 963 neutral variants. A total of 816 disease-associated and 776 neutral variants are randomly selected as the benchmark dataset and the rest are used as the independent dataset including of 190 disease-associated and 187 neutral variants.

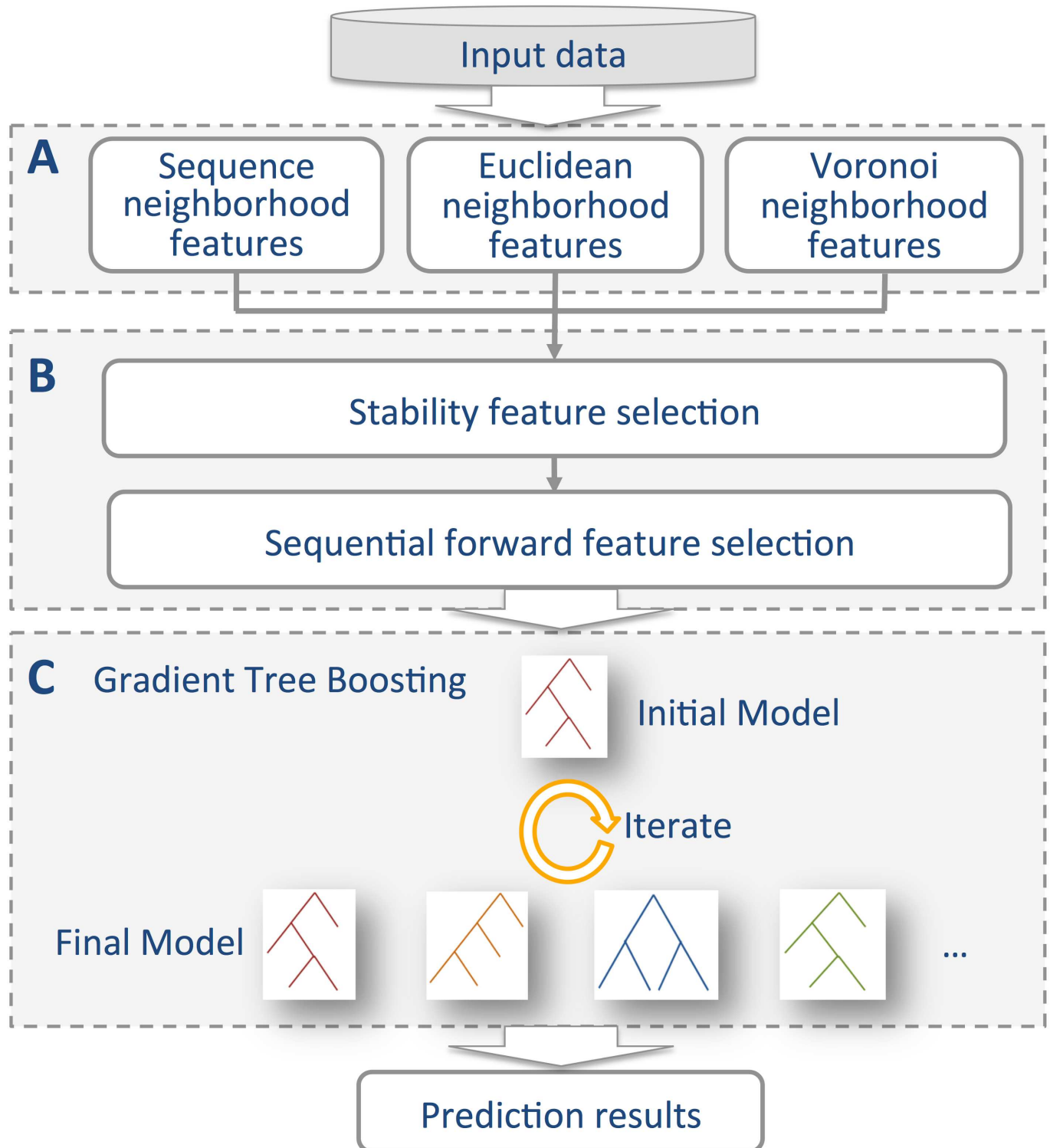


Fig 1. The framework of PredSAV. (A) Feature representation. A total of 1521 sequence, Euclidean and Voronoi neighborhood features are initially generated. (B) Two-step feature selection. Stability selection is used as the first step. We select the top 152 features with score larger than 0.2. The second step is performed using a wrapper-based feature selection. Features are evaluated by 5-fold cross-validation with the GTB algorithm. (C) Prediction model. Gradient boosted trees are finally built for prediction.

<https://doi.org/10.1371/journal.pone.0179314.g001>

Performance evaluation

We evaluate the performance of the proposed method using 5-fold cross-validation and several widely used measures. These measures include sensitivity (SEN/Recall), specificity (SPE), precision (PRE), F1-score (F1), accuracy (ACC), the Matthew's correlation coefficient (MCC) and the area under the ROC curve (AUC).

$$SEN = TP / (TP + FN) \quad (1)$$

$$SPE = TN / (TN + FP) \quad (2)$$

$$PRE = TP / (TP + FP) \quad (3)$$

$$F1 = \frac{2 \times Recall \times Precision}{Recall + Precision} \quad (4)$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (6)$$

Features extraction

In our experiment, a wide variety of sequence and structure features are generated for predicting the phenotypic effects of SAVs. Several novel structural features, including residue-contact network features, solvent exposure features and structural neighborhood features, are calculated. The details of these features are listed as follows.

Sequence features. A large number of sequence features are calculated: 1) position-specific scoring matrices (PSSMs) [34]; 2) predicted solvent accessibility using the SSpro and SSpro8 programs [36]; 3) predicted native disorder by DISOPRED [37]; 4) the dScore that represents the difference between the PSIC [38] scores for the wild type amino acid residue and mutant amino acid residue, calculated by PolyPhen2 [24]; 5) predicted disorder in proteins by DisEMBL [39]; 6) the local structural entropy of a particular residue is computed by LSE [40]; 7) the eight physicochemical properties for each amino acid are obtained from the AAindex database [41]; 8) BLOSUM62 [42] was used to count the relative frequencies of amino acid and their substitution probabilities; 9) solvent accessible surface area, secondary structure and local backbone angles generated by SPIDER2 [43]; 10) predicted the relative solvent accessibility of protein residues by the ACCpro and ACCpro20 from the SCRATCH package [36]; 11) evolutionary conservation scores calculated based on PSSM [34] and Jensen-Shannon divergence [44, 45].

Structure features. Structural features, including secondary structure, four-body statistical pseudo-potential, solvent accessibility and exposure features, are calculated as candidate features for SAV phenotype prediction. We used DSSP [46] to calculate the secondary structure features, including hydrogen bonds, solvent-accessible surface area, C_{α} atom coordinates and backbone torsion angles. The four-body statistical pseudo-potential is based on the Delaunay tessellation of proteins [47]. Delaunay tessellation is an effective way to define the structural

neighbors of a target protein. The potential is defined as follows:

$$Q_{ijkl}^{\alpha} = \log \left[\frac{f_{ijkl}^{\alpha}}{p_{ijkl}^{\alpha}} \right] \quad (7)$$

where $i, j, k,$ and l are the residue identities of the four amino acids in a Delaunay tessellation of the target protein. Each residue is represented by a central point among the atoms in the residue. f_{ijkl}^{α} is the observed frequency of the residue composition ($ijkl$) in a tetrahedron of type α over a set of protein structures. p_{ijkl}^{α} is the expected random frequency.

Energy scores including side-chain energy score, residue energy, conservation, interface propensity, combined1 score, combined2 score and relative solvent accessibility are calculated by using ENDES [48]. Two combined energy scores are also used. The combined1 score is a combination of residue energy, conservation and interface propensity scores. The combined2 score is an optimized combination weights of the three features to get the best prediction of single residue.

Solvent-accessible related features have been shown to be very useful in identifying SAV-disease association [49–51]. We use NACCESS [52] and NetSurfP [53] to calculate solvent accessibility for the protein structures, respectively. The NACCESS program is used to calculate the absolute and relative solvent accessibilities of all atoms. For NetSurfP, the absolute and relative surface accessibility, Z-fit score and secondary structure are computed based on the homology proteins obtained from PSI-BLAST search.

Solvent exposure features, include the coordination number (CN), number of C_{β} atoms in the upper Half-Sphere (HSEBU), number of C_{β} atoms in the lower Half-Sphere(HSEBD) and residue depth (RD), are calculated by HSEpred [54] and the hsexpo program [55]. The hsexpo program uses protein structure information, while the HSEpred uses sequence information to predict these features.

Residue-contact network features. Residue-residue contact networks have been proved to be very beneficial for analyzing and predicting SAV-disease associations [56]. If the distance between the centers of two residues in a structure are within 6.5\AA , an edge exists between the two residues in the network. We use NAPS [57] to compute the residue-residue contact network properties, which describe the local environment of the target variant in the network, including betweenness, closeness, coreness, degree, clustering coefficient, eigenvector centrality, eccentricity and average nearest neighbor degree.

Structural neighborhood features (SNF). Conventional features usually describe only the properties of the current residue itself, cannot represent the real environment well, and thus are insufficient to predict functional effects of SAVs with high precision. Here, we calculate two types of structural neighborhood features (SNF) based on Euclidean distance and Voronoi diagram [58–60], respectively. Surrounding residues located within a sphere of the radius of 5\AA are defined as the Euclidean neighborhood of the central amino acid. The Euclidean distance is computed between any heavy atoms of the surrounding residues and that of the central amino acid. The score of a specific feature i for the central residue r regarding the neighbor n is defined as follows:

$$F_i(r, n) = \begin{cases} \text{the score of feature } i \text{ for residue } r \text{ if } |r - n| \geq 1 \text{ and } d_{r,n} \leq 5\text{\AA}, \\ 0 \quad \text{otherwise} \end{cases} \quad (8)$$

where $d_{r,n}$ is the minimum Euclidean distance between residue r and residue n . The Euclidean

neighborhood feature of the central residue r is defined as:

$$EN_i(r) = \sum_{n=1}^m F_i(r, n), \quad (9)$$

where m is the total number of Euclidean neighbors.

Voronoi neighborhood features are calculated based on Voronoi diagram/Delaunay triangulation. For a 3-D protein structure, individual atoms are divided into Voronoi polyhedra by Voronoi tessellation partition. In the Voronoi diagram (Delaunay triangulation), a pair of residues are defined as Voronoi neighbors if there exists at least one common Voronoi facet between heavy atoms of each residue. The Qhull package [61] is used to calculate Voronoi/Delaunay polyhedra. For the target residue r and its Voronoi neighbors $n \{n = 1, \dots, m\}$, the Voronoi neighborhood property of the feature i is defined as:

$$VD_i = \sum_{n=1}^m P_i(n), \quad (10)$$

where $P_i(n)$ is the score of the residue feature i for neighbor n .

Feature encoding with neighborhood properties. For each sample, a combination of 1,287 (117*11) sequence neighborhood features, 117 Euclidean neighborhood features and 117 Voronoi neighborhood features are calculated. The sequence neighborhood features is generated by applying a sliding window of size 11 to incorporate the evolutionary information from upstream and downstream neighbors in the protein sequence.

Feature selection

The feature selection method improves the performance by removing some redundant features in high-dimensional data [62–65]. In this study, we propose a two-step feature selection approach to select the most important features for predicting the phenotypic effects of SAVs. First, we assess the feature elements using the stability selection [66] calculated by the RandomizedLasso package in the scikit-learn [67]. The idea of stability selection is that a feature selection algorithm is employed on subsample datasets and subsample features. The selection results are merged after repeating a certain number of times. Stronger features have higher scores (close to 1), while weaker features have scores close to 0. The score represents the importance of an individual feature for correctly predicting an SAV-disease association. Here, we select the top 152 features with the score larger than 0.2.

The second step is performed using a wrapper-based feature selection method. The features are evaluated by 5-fold cross-validation with the GTB (gradient tree boosting) algorithm, and correlation features are added by sequential forward selection (SFS). In the SFS scheme, features are sequentially added to a null feature set till an optimal feature subset is obtained. Each added feature is the one whose add maximizes the performance of the classifier. This stepwise feature selection process continues until the AUC score no longer increased. As a result, a set of 44 optimal features are selected as the final optimal feature set.

Gradient tree boosting algorithm

The Gradient Tree Boosting (GTB) [27, 28] is an effective machine learning algorithm that can be utilized for both classification and regression problems. In this study, GTB is implemented under the PredSAV framework as shown in Fig 1 and the prediction of the phenotypic effect of single amino acid variants could be considered as a binary classification problem. For a large number of given input feature vectors \mathcal{X}_i ($\mathcal{X}_i = \{x_1, x_2, \dots, x_n\}$, $i = 1, 2, \dots, N$) with labels

$y_i \in \{-1, +1\}, i = 1, 2, \dots, N$, where “-1” represents neutral variant and “+1” denotes disease-associated variant), the purpose of the GTB algorithm is to build an effective classifier to predict whether a variant is disease-associated or neutral. The GTB algorithm is shown in Algorithm 1.

Algorithm 1 Gradient Tree Boosting Algorithm

Input:

Data set: $D = \{(\chi_1, y_1), (\chi_2, y_2), \dots, (\chi_N, y_N)\}, \chi_i \in \mathcal{X}, \mathcal{X} \subseteq \mathbf{R}, y_i \in \{-1, +1\}$; loss function: $L(y, \Theta(\chi))$; iterations = M ;

Output:

- 1: Initialize $\Theta_0(\chi) = \arg \min_c \sum_i L(y_i, c)$;
- 2: **for** $m = 1$ to M **do**
- 3: Compute the negative gradient as the working response

$$r_i = - \left[\frac{\partial L(y_i, \Theta(\chi_i))}{\partial \Theta(\chi_i)} \right]_{\Theta(\chi) = \Theta_{m-1}(\chi)}, i = \{1, \dots, M\}$$
- 4: The input χ_i is adapted to the classification model r_i by Logistic function and get the estimate α_m of $\beta h(\chi; \alpha)$
- 5: Get the estimate β_m by minimizing $L(y_i, \Theta_{m-1}(\chi_i) + \beta h(\chi_i; \alpha_m))$
- 6: Update $\Theta_m(\chi) = \Theta_{m-1}(\chi) + \beta_m h(\chi; \alpha_m)$
- 7: **end for**
- 8: **return** $\tilde{\Theta}(\chi) = \Theta_M(\chi)$

In the algorithm, the variable $iterations = M$ should be initialized. The logistic function is used as the loss function, which is defined as:

$$L(y, \Theta(\mathbf{x})) = \log(1 + \exp(-y\Theta(\mathbf{x}))), \tag{11}$$

where y is a real class label of variants and $\Theta(\chi)$ is a decision function. The decision function is initialized by the following equation.

$$\Theta_0(\chi) = \arg \min_c \sum_{i=1}^N L(y_i, c), \tag{12}$$

where N is the number of SAVs in the benchmark dataset. Then, GTB constructs m different classification trees $h(\chi, \alpha_1), h(\chi, \alpha_2), \dots, h(\chi, \alpha_m)$ from a number of benchmark datasets. The additive function $\Theta_m(\mathbf{x})$ can be defined as:

$$\Theta_m(\chi) = \Theta_{m-1}(\chi) + \beta_m h(\chi; \alpha_m) \tag{13}$$

Above, the β_m and α_m are a weight and vector of parameters for the m -th classification tree $h(\chi, \alpha_m)$, respectively. In order to minimize the loss function $L(y, \Theta_m(\chi))$, the weight of β_m and the parameter of α_m need to be iterated from $m = 1$ to $m = M$. In the third step, the negative gradient r_i as the working response by the following formula:

$$r_i = - \left[\frac{\partial L(y_i, \Theta(\chi_i))}{\partial \Theta(\chi_i)} \right]_{\Theta(\chi) = \Theta_{m-1}(\chi)}, i = 1, \dots, M \tag{14}$$

Then, the weight of β_m and the parameter of α_m for the m th iteration can be defined as:

$$(\beta_m, \alpha_m) = \arg \min_{\beta, \alpha} \sum_{i=1}^N L(y_i, \Theta_{m-1}(\chi_i) + \beta h(\chi_i; \alpha_m)) \tag{15}$$

However, we do not directly calculate the above equation. In the fourth step, the input χ_i is adapted to the classification model r_i by Logistic function and get the estimate α_m of $\beta h(\chi_i; \alpha)$.

Therefore, we can obtain

$$\alpha_m = \arg \min_{\alpha} \sum_{i=1}^N \frac{1}{1 + e^{r_i h(\chi_i; \alpha)}} \tag{16}$$

In the fifth step, the estimate parameter of β_m is obtained by minimizing the log loss function $L(y, \Theta(\chi))$.

$$\begin{aligned} \min \quad & L(y_i, \Theta_{m-1}(\chi_i) + \beta h(\chi_i; \alpha_m)) \\ = \min \quad & \log(1 + \exp(-y_i(\Theta_{m-1}(\chi_i) + \beta h(\chi_i; \alpha_m)))) \end{aligned} \tag{17}$$

Then, in the sixth step, a new additive function $\Theta_m(\chi)$ is updated by in the eq (8). Finally, we obtain a classification function $\Theta_M(\chi)$ and a useful GTB model $\tilde{\Theta}(\chi)$ as follows:

$$\tilde{\Theta}(\chi) = \Theta_M(\chi) \tag{18}$$

We use a grid search strategy to select the optimal parameters of GTB with 5-fold cross-validation on the benchmark dataset. The optimized number of trees of the GTB is 2000. And the selected depth of the trees is 3. The rest use the default parameters.

The source code and data are available at <http://www.leideng.org/PredSAV/>.

Results and discussion

Benefits of the two-step feature selection

The selection of informative attributes is critically important for building effective and accurate classification models. In total 1521 sequence, Euclidean and Voronoi neighborhood features are initially generated. We apply a two-step feature selection method, consisting of stability selection and sequential forward selection. Stability selection is used as the first attribute selection step for two reasons. First, stability selection can address the difficult variable selection problem with markedly improved error control and structure estimation, especially for high-dimensional problems. Second, stability selection depends little on the chosen initial regularisation and can reduce the risk of overfitting [66]. To assess the utility of the stability selection method, we evaluate the performance by incorporating the GTB classifier with selected attributes that correspond to different cutoffs of stability selection scores. As shown in Table 1, when the number of selected features decreases from 1521 to 152 (the cutoff increases from 0 to 0.2), the highest accuracy of 81.3% is yielded. The other measurements (SEN, SEP, PRE, MCC, AUC)

Table 1. Performance of selected attributes with the two-step feature selection method. The first column lists different cutoffs of stability selection scores.

Features	Number	ACC	SEN	SEP	PRE	MCC	AUC
All features	1521	0.804	0.802	0.807	0.815	0.608	0.881
score>0.1	263	0.808	0.804	0.813	0.812	0.610	0.886
score>0.15	191	0.808	0.809	0.807	0.815	0.616	0.892
score>0.2	152	0.813	0.810	0.817	0.823	0.628	0.896
score>0.25	112	0.810	0.809	0.813	0.819	0.622	0.893
score>0.3	93	0.809	0.810	0.806	0.815	0.618	0.890
score>0.35	84	0.810	0.809	0.808	0.814	0.618	0.889
score>0.4	69	0.809	0.808	0.809	0.814	0.615	0.888
Final optimal features	44	0.826	0.814	0.838	0.839	0.651	0.908

<https://doi.org/10.1371/journal.pone.0179314.t001>

PRE, MCC and AUC) are observed as 0.810, 0.817, 0.823, 0.628 and 0.896, respectively. We select the top 152 features (stability score > 0.2) as the input of the next sequential forward selection step. A set of 44 optimal features is finally selected with the highest AUC score of 0.908. The results of selected features show ~2% and ~5% increase in AUC and MCC over the initial features, respectively.

We compare the proposed two-step feature selection method with three widely used feature selection methods: random forests(RF), maximum Relevance Minimum Redundancy(mRMR) [68] and Recursive Feature Elimination(RFE) [69]. The experiment is based on the benchmark dataset with 5-fold cross-validation. Fig 2 shows the ROC curves of the four feature selection methods. The results are shown in Fig 2. Our two-step feature selection approach obtains the best performance. The results indicate that our two-step feature selection algorithm, which is a composite approach combining the merits of both stability selection and sequential forward selection, can substantially boost the prediction performance with less computational expense and lower risk of overfitting.

Feature importance

The feature importance of these features are calculated by using the gradient tree boosting method. The relative importance and rankings of the optimal features are shown in Fig 3 and

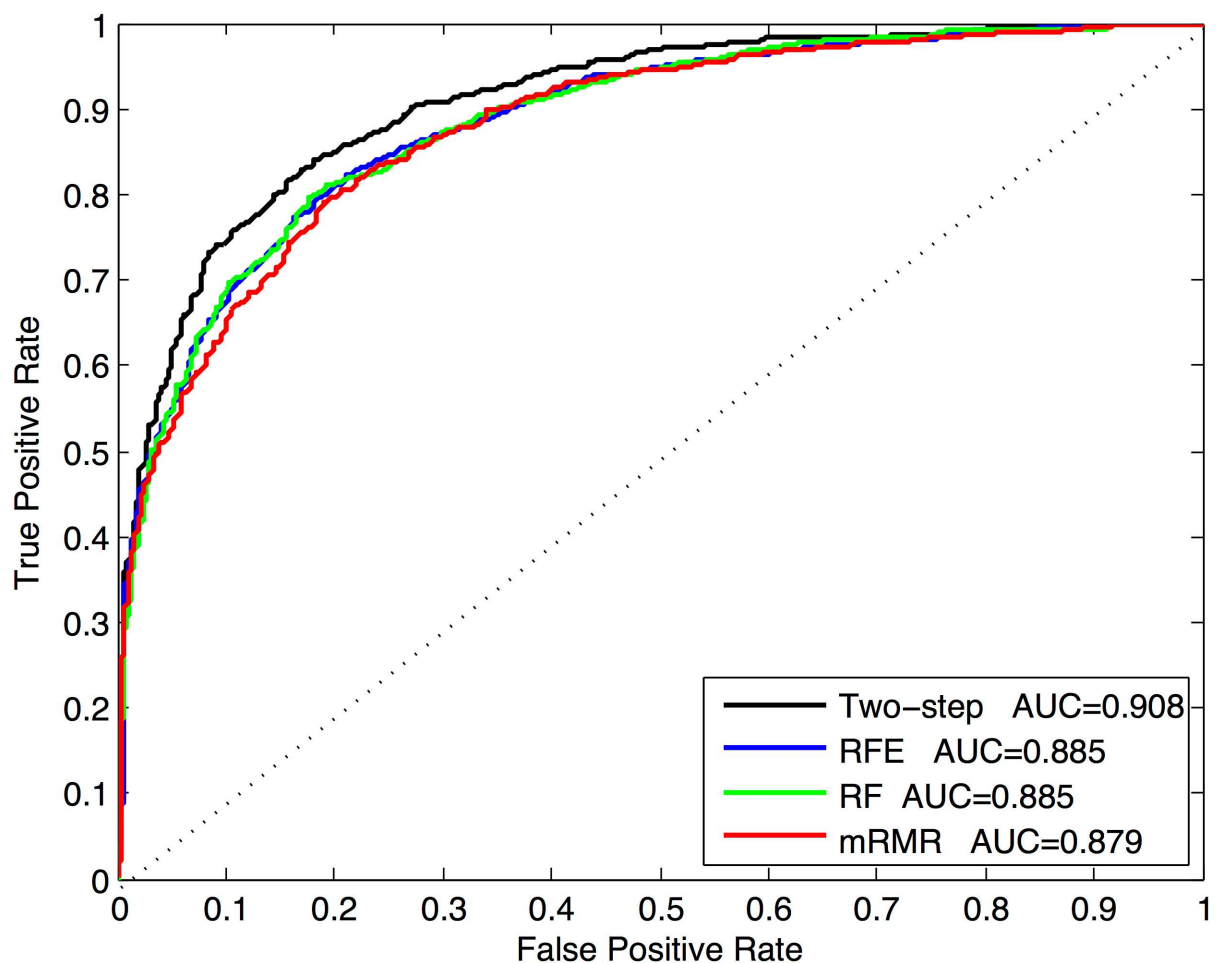


Fig 2. ROC curves of our two-step algorithm and other three existing feature selection methods.

<https://doi.org/10.1371/journal.pone.0179314.g002>

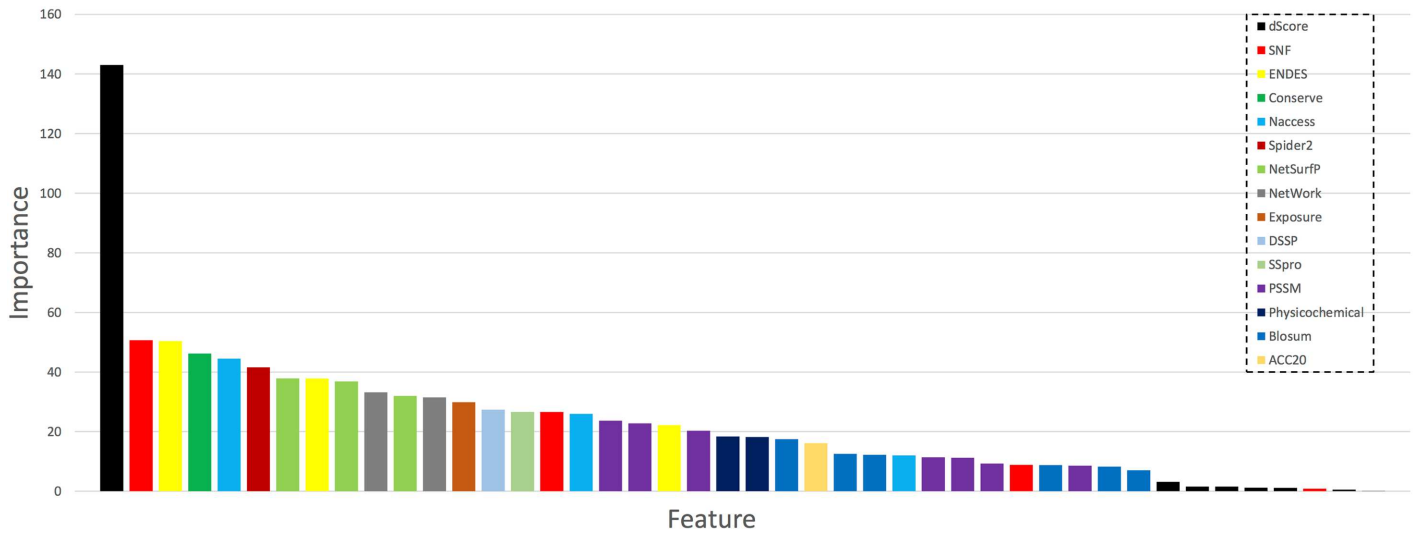


Fig 3. The relative importance and ranking of the optimal feature group, as evaluated by the gradient tree boosting. The bar represents the importance score of the corresponding feature group.

<https://doi.org/10.1371/journal.pone.0179314.g003>

Table 2. Among them, the feature with the highest score (>100) is the dScore feature calculated by PolyPhen2. The dScore represents the difference between the PSIC scores of the wild type amino acid residue and mutant amino acid residue. Another important feature that have not been found useful in previous studies is structural neighborhood features (Euclidean and

Table 2. Rankings of feature importance for the optimal selected features. SN, EN and VN represent sequence neighborhood, Euclidean neighborhood and Voronoi neighborhood, respectively. The numbers in the brackets denote the positions in the sliding window for sequence neighborhood features.

Rank	Feature name	Type	Rank	Feature name	Type
1	dScore(6)	SN	23	Flexibility parameter in physicochemical(5)	SN
2	dScore	VN	24	BLOSUM(H)(7)	SN
3	RA in ENDES(4)	SN	25	ACC20 in SCRATCH(11)	SN
4	Conservation score(1)	SN	26	BLOSUM(N)(3)	SN
5	Non-polar ABS in Naccess(6)	SN	27	BLOSUM(L)(4)	SN
6	ASA in SPIDER2(6)	SN	28	Main-Chain ABS in Naccess(9)	SN
7	Z-fit score in Netsurfp(10)	SN	29	PSSM(C)(9)	SN
8	RA in ENDES(8)	SN	30	PSSM(S)(10)	SN
9	ASA in Netsurfp(2)	SN	31	PSSM(C)(1)	SN
10	Betweenness in NetWork(9)	SN	32	ACC	EN
11	ASA in Netsurfp(6)	SN	33	BLOSUM(L)(6)	SN
12	Closeness in NetWork(6)	SN	34	hydrophobic moment in physicochemical(3)	SN
13	HSEBU in HSEpred(7)	SN	35	BLOSUM(T)(7)	SN
14	KAPPA in DSSP(7)	SN	36	BLOSUM(C)(4)	SN
15	SSpro in the SCRATCH(7)	SN	37	dScore(11)	SN
16	combined2 in ENDES(4)	SN	38	dScore(1)	SN
17	Total-Side REL in Naccess(9)	SN	39	dScore(9)	SN
18	PSSM(N)(10)	SN	40	dScore(3)	SN
19	PSSM(I)(6)	SN	41	dScore(2)	SN
20	PROPENSITY in ENDES(5)	SN	42	dScore	EN
21	PSSM(G)(1)	SN	43	dScore(10)	SN
22	Flexibility parameter in physicochemical(11)	SN	44	dScore(7)	SN

<https://doi.org/10.1371/journal.pone.0179314.t002>

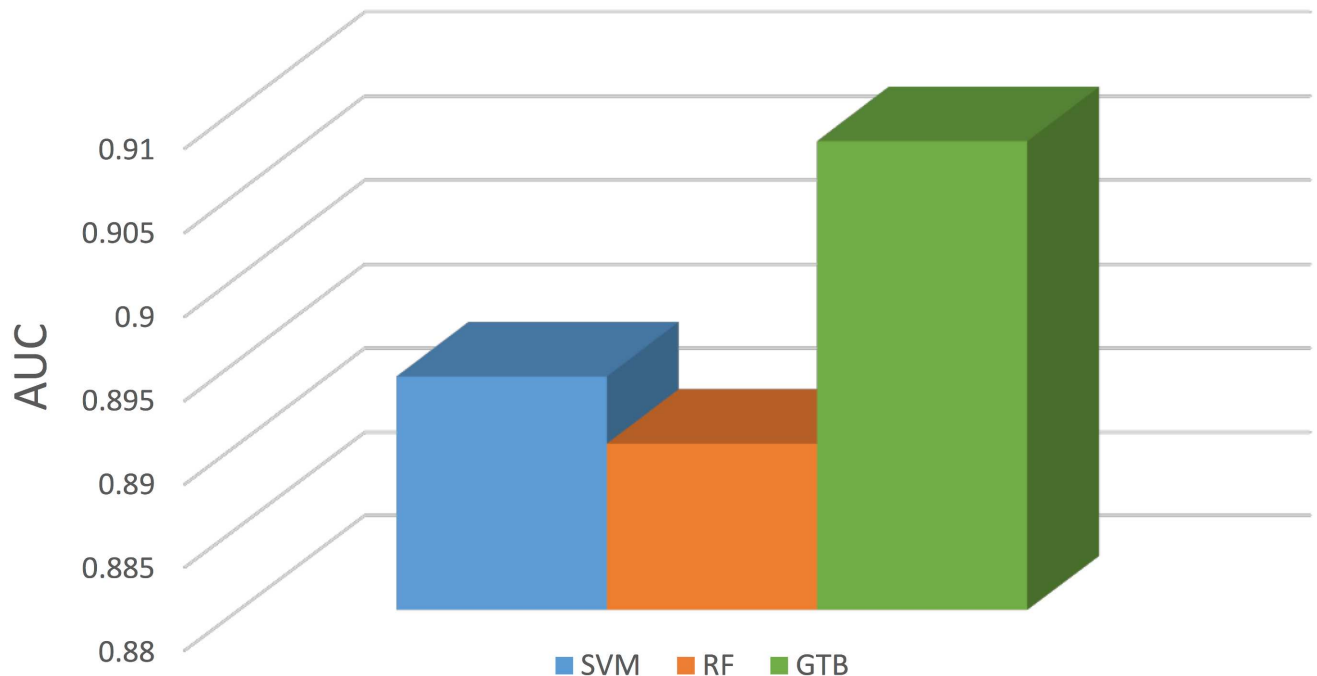


Fig 4. Comparison of the AUC value of the the three methods using 5-fold cross-validation on the benchmark dataset.

<https://doi.org/10.1371/journal.pone.0179314.g004>

Voronoi). We find that structural neighborhood features achieve the second highest score compared with other features, suggesting that structural neighborhood features are critical in distinguishing disease-associated SAVs from neutral SAVs. Solvent accessibilities are also found to be useful in SAVs phenotype prediction. Many solvent accessibility related features, including the solvent accessibility feature calculated by NACCESS, NetSurfP, DSSP, solvent exposure features, ACC20 and the SSpro score, are important and contributive. The results suggest that sequence and structural neighborhood features complement each other quite well and thus collectively make a contribution to the performance enhancement.

Gradient tree boosting improves predictions

PredSAV uses Gradient Tree Boosting (GBT) to build the final model with the 44 optimal features. We compare GBT with Support Vector Machine (SVM) and Random Forests (RF), which are well known to perform fairly well on a variety of tasks. Fig 4 shows the AUC scores of GTB and other machine learning methods on the final optimal feature set. GTB, SVM and RF achieve AUC values of 0.908, 0.894 and 0.890, respectively. Comparing with the other methods, the GTB model can improve the prediction performance. Note that the GBT algorithm is implemented with scikit-learn [67] in this study.

PredSAV outperforms other predictors

To evaluate the performance of the proposed PredSAV, six existing SAVs prediction methods, including FunSAV [25], Polyphen2 [24], SusPect [26], SIFT [21], SNAP [23] and nsSNPAnalyzer [70], are evaluated on the benchmark dataset.

Table 3 and Fig 5 show the detailed results of comparing our method with the existing methods. Overall, our approach shows dominant advantage over the existing methods in six metrics: ACC, SPE, PRE, F1, MCC and AUC. When comparing the AUC score with that of

Table 3. Prediction performance of PredSAV classifiers in comparison with six other prediction tools on the benchmark dataset.

Method	ACC	SEN	SEP	PRE	F1	MCC	AUC
PredSAV	0.826	0.814	0.838	0.839	0.826	0.651	0.908
FunSAV	0.749	0.762	0.736	0.753	0.757	0.508	0.814
PolyPhen2	0.732	0.866	0.590	0.690	0.768	0.476	0.813
SusPect	0.723	0.653	0.798	0.775	0.709	0.455	0.800
SIFT	0.697	0.699	0.695	0.707	0.703	0.394	0.760
SNAP	0.635	0.855	0.395	0.605	0.709	0.284	0.706
nsSNPAnalyzer	0.712	0.745	0.661	0.774	0.759	0.401	-

<https://doi.org/10.1371/journal.pone.0179314.t003>

the existing classifiers, FunSAV (0.814), Polyphen2 (0.813), SusPect (0.800), SIFT (0.760) and SNAP (0.706), our PredSAV classifier (0.908) shows greater improvement by 9%, 9%, 11%, 14% and 20%, respectively. For the remaining measurements ACC, SPE, PRE, F1 and MCC, we can observe similar increases. Especially, the specificity of PredSAV is significant higher than other methods (increased by 10%), which suggests that it has better performance detecting true negatives and may help for reducing experiment cost. Only in SEN, PredSAV is lower than PolyPhen2 and SNAP (0.855 and 0.866 for SNAP and PolyPhen2, respectively). We can observe that PredSAV gains a balanced sensitivity and specificity (0.814 and 0.838, respectively), suggesting that PredSAV has better balance of prediction accuracy between disease-associated and neutral SAVs.

Performance confirmed for independent test

We also validate the performance of PredSAV on the independent test dataset to avoid over-optimistic performance estimates. Results of the independent test are presented in [Table 4](#), which indicate marked improvements for all the performance measures except SEN comparing PredSAV with the exiting methods. [Fig 6](#) shows the ROC curves. The ROC curves indicate the trade-off between the amounts of true positives (TP) and false positives (FP) generated by the classifiers. We observe that PredSAV produces higher true positive rates of prediction across most of the false positive rates. Overall, these observations suggest that the performance of our PredSAV approach is superior to that of the state-of-the-art approaches.

Case study

To further illustrate the effectiveness of PredSAV, we present examples by comparing predictions for variants that are difficult to classify with commonly applied methods. The enzyme phenylalanine hydroxylase (PAH, PDB ID: 1J8U, chain A) [71, 72] is responsible for the conversion of phenylalanine to another amino acid, tyrosine. PAH works with a molecule called tetrahydrobiopterin (BH4) to carry out this chemical reaction. The majority of mutations in PAH result in deficient enzyme activity and cause hyperphenylalaninemia. Some cause phenylketonuria (PKU), others cause non-PKU hyperphenylalaninemia, while still others are silent polymorphisms. As shown in [Fig 7A](#), three PKU-associated SAVs, Q160P (dbSNP: rs199475601), V177L (dbSNP:rs199475602) and V388L (dbSNP:rs62516101), are colored in red. This example illustrates how PredSAV combines gradient tree boosting with optimal neighborhood features to provide better predictions. PredSAV (TP = 3) correctly identified all the three disease-associated variants, compared to Suspect (TP = 0), PolyPhen (TP = 0), SNAP (TP = 2), FunSAV (TP = 0), SIFT (TP = 1) and nsSNP (TP = 0).

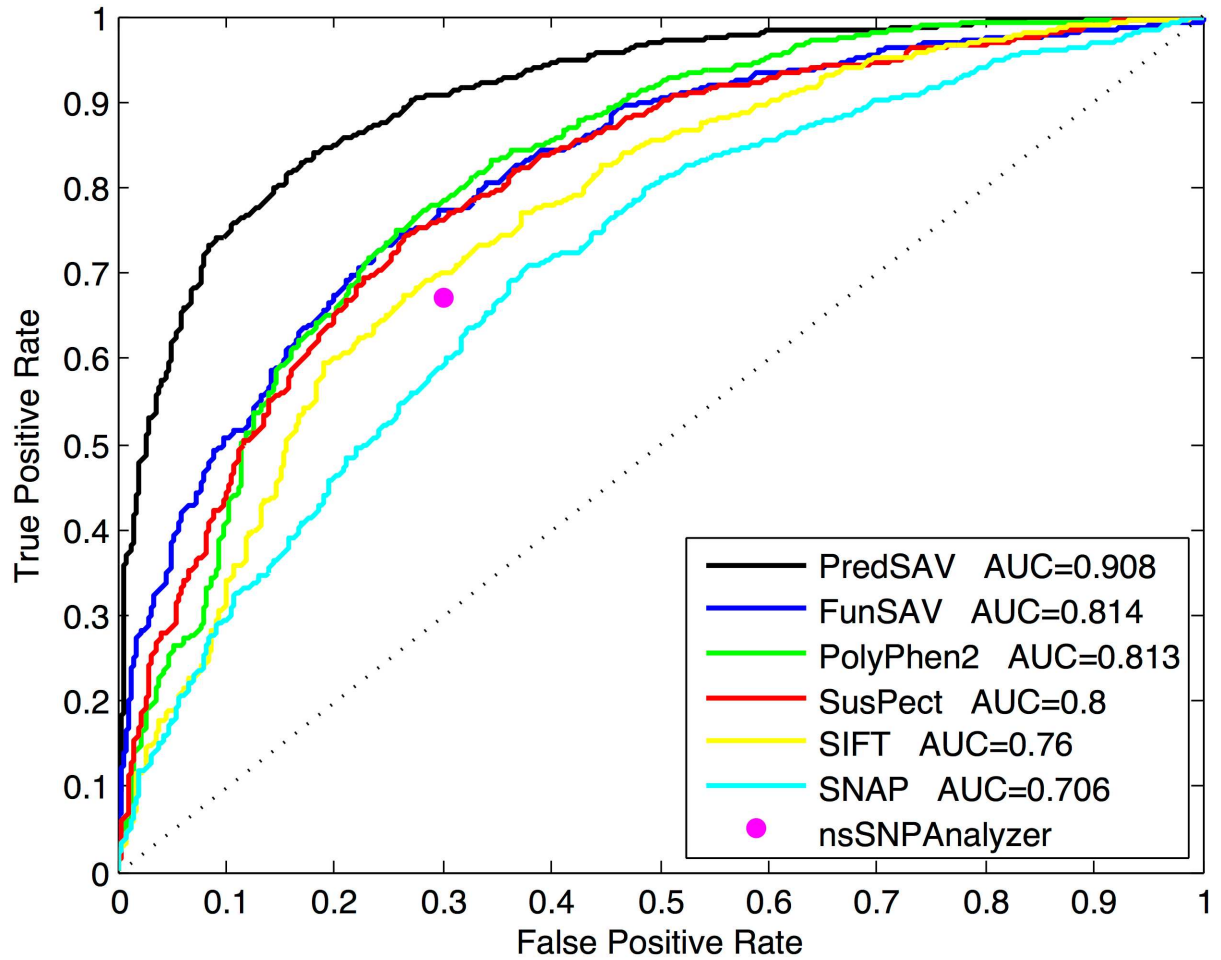


Fig 5. The ROC curves of seven classifiers on the benchmark dataset.

<https://doi.org/10.1371/journal.pone.0179314.g005>

Another example is lanosterol synthase (LSS, PDB ID: 1W6K, chain A) [73, 74], which catalyzes the cyclization of (S)-2,3 oxidosqualene to lanosterol, a reaction that forms the sterol nucleus. Through the production of lanosterol may regulate lens protein aggregation and increase transparency. The variants R614W (dbSNP:rs35785446) and P688L (dbSNP:rs17293705) in LSS are neutral substitutions. From Fig 7B, we can see that PredSAV can predict the neutral SAVs successfully, while other existing methods result in almost completely wrong results (except Suspect in P688L). This suggests that PredSAV has the highest

Table 4. Prediction performance of PredSAV classifiers in comparison with six other prediction tools on the independent test dataset.

Method	ACC	SEN	SEP	PRE	F1	MCC	AUC
PredSAV	0.790	0.780	0.802	0.800	0.789	0.581	0.855
FunSAV	0.731	0.769	0.701	0.679	0.721	0.480	0.792
PolyPhen2	0.727	0.868	0.583	0.679	0.762	0.471	0.806
SusPect	0.716	0.684	0.749	0.734	0.708	0.434	0.774
SIFT	0.729	0.774	0.684	0.714	0.742	0.460	0.786
SNAP	0.594	0.879	0.305	0.562	0.686	0.225	0.671
nsSNPAnalyzer	0.639	0.705	0.550	0.677	0.691	0.258	-

<https://doi.org/10.1371/journal.pone.0179314.t004>

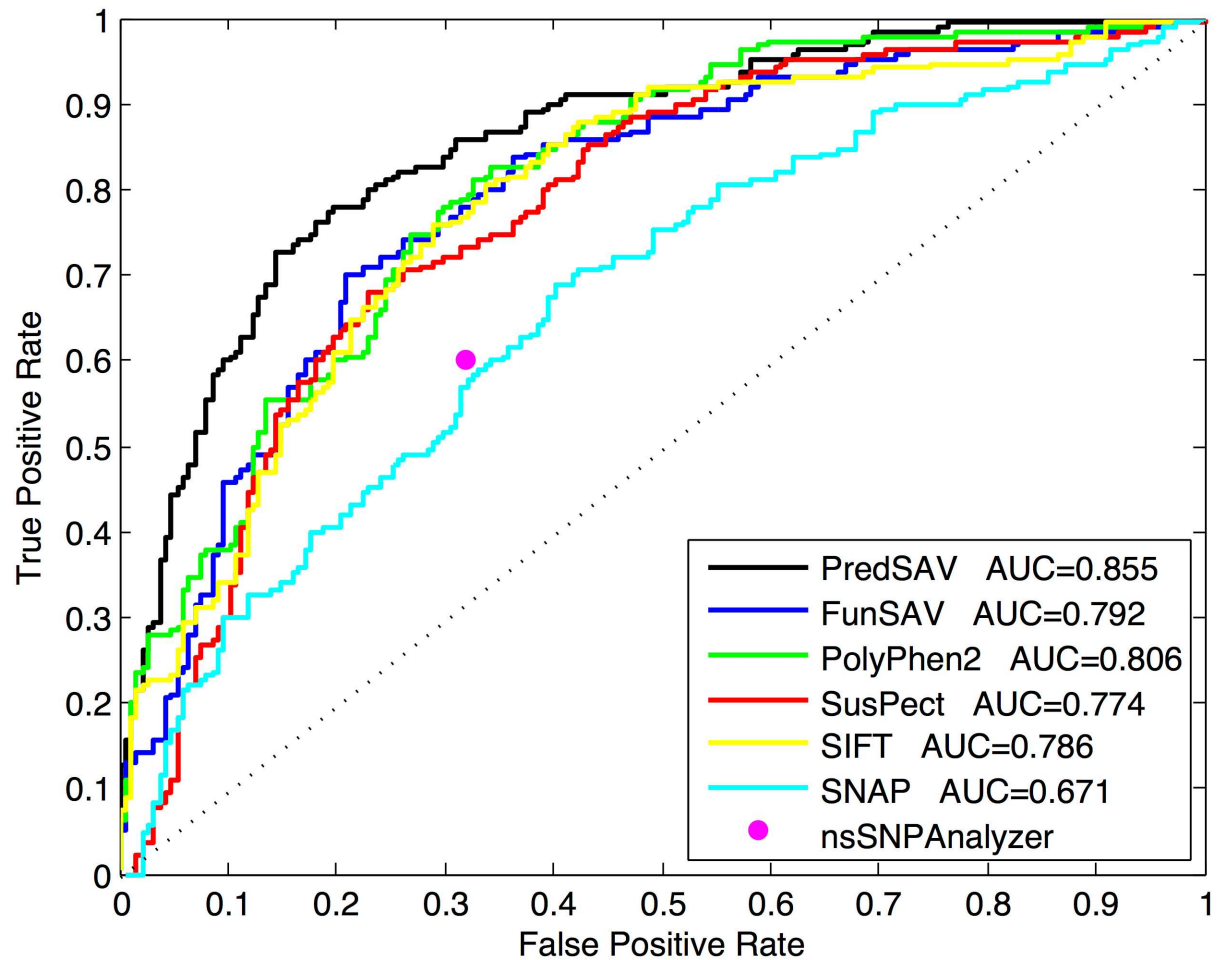


Fig 6. The ROC curves of seven classifiers on the independent test dataset.

<https://doi.org/10.1371/journal.pone.0179314.g006>

specificity, which is desirable for many biological applications since it allows researchers to identify a short list of SAVs for targeted phenotype studies.

Conclusion

In this study, we present a novel approach named PredSAV for producing reliable predictions in distinguishing between effect and neutral variants. To be able to do this, we first extract a very large collection of informative and complementary features, including sequence, structure, network and neighborhood features that describe the local environments proximal to the centered variant and neighboring residues. A two-step feature selection approach, which combines stability selection and sequential forward selection, is utilized to select an optimal subset of features within a reasonable computational cost, and thus improves the prediction performance and reduces the risk of overfitting. Importantly, the use of gradient tree boosting algorithm further attains higher levels of prediction accuracy. We evaluate the PredSAV method with both cross-validation and independent test, and the results indicate that the proposed PredSAV is able to identify disease-associated SAVs with higher overall performance, especially in terms of specificity, when compared with other existing approaches.

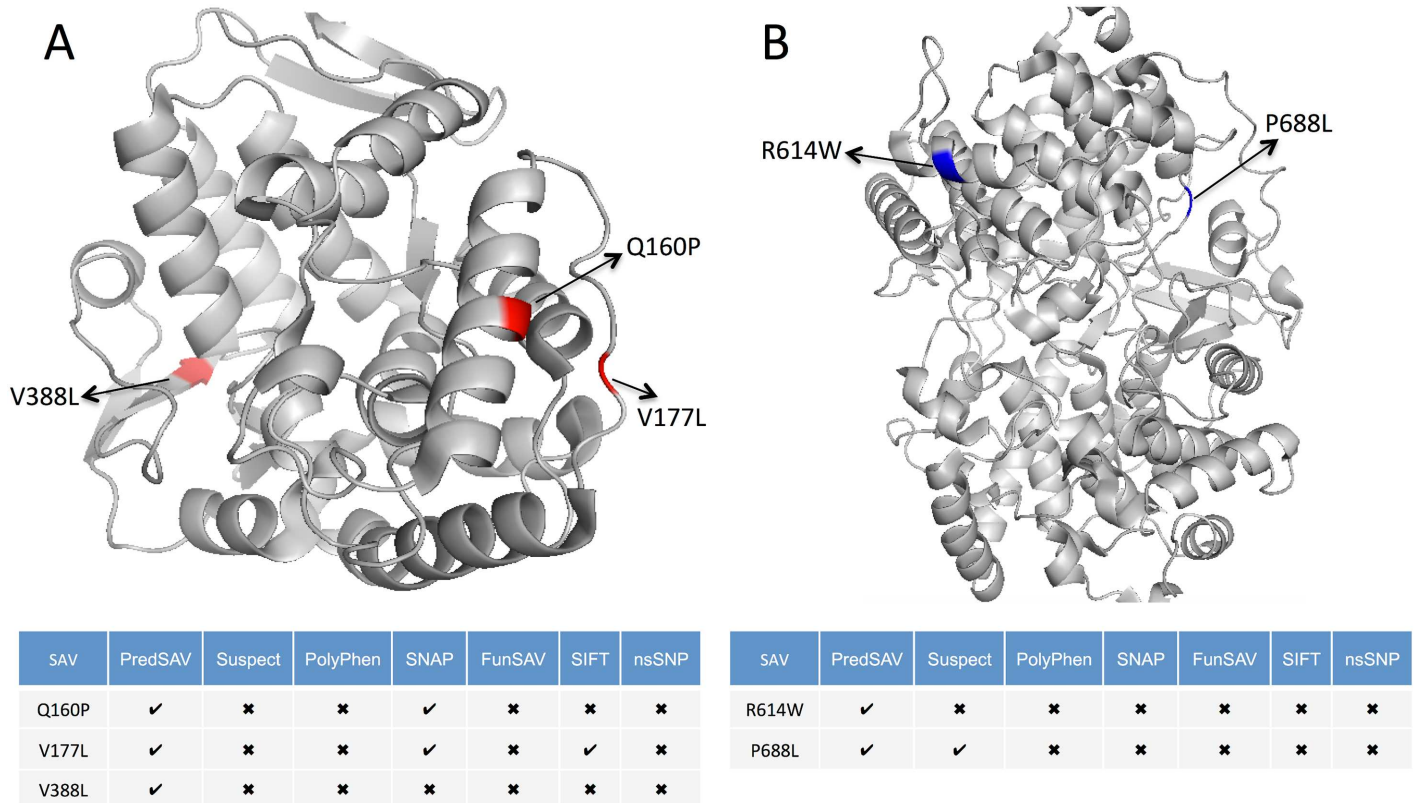


Fig 7. Prediction examples of the functional effects of SAVs in two proteins by PredSAV and other methods. Red color denotes disease-associated variants while blue color represents neutral variants. (A) and (B) represent proteins PAH (PDB ID: 1J8U, chain A) and LSS (PDB ID: 1W6K, chain A), respectively. 3-D structures are rendered using PyMol [75].

<https://doi.org/10.1371/journal.pone.0179314.g007>

A limitation of PredSAV is that it requires the 3D protein structure, which may limit its broader application. However, with the increasing solved protein structures, protein homology modeling projects [76] and predicted 3D structures [77], it is expected that PredSAV can be used as a powerful tool to prioritize the disease-associated variants and help towards the phenotypic effect annotation of these targets.

As for future work, we will explore more efficient features to further improve the performance and learn from other methods [78–82] to provide a web-server for the method proposed in this paper.

Supporting information

S1 File. The disease-associated and neutral variant data used in these experiments. (ZIP)

Author Contributions

Conceptualization: YP LD.

Data curation: YP.

Formal analysis: YP DL.

Funding acquisition: LD.

Investigation: LD.

Methodology: YP DL LD.

Project administration: LD.

Resources: LD YP.

Software: YP DL.

Supervision: LD.

Validation: YP DL LD.

Writing – original draft: YP LD.

Writing – review & editing: YP DL LD.

References

1. Yip YL, Scheib H, Diemand AV, Gattiker A, Famiglietti LM, Gasteiger E, et al. The Swiss-Prot variant page and the ModSNP database: A resource for sequence and structure information on human protein variants. *Human mutation*. 2004; 23(5):464–470. <https://doi.org/10.1002/humu.20021> PMID: 15108278
2. Consortium GP, et al. A map of human genome variation from population-scale sequencing. *Nature*. 2010; 467(7319):1061–1073. <https://doi.org/10.1038/nature09534> PMID: 20981092
3. Cline MS, Karchin R. Using bioinformatics to predict the functional impact of SNVs. *Bioinformatics*. 2011; 27(4):441–448. <https://doi.org/10.1093/bioinformatics/btq695> PMID: 21159622
4. Schuster SC, Miller W, Ratan A, Tomsho LP, Giardine B, Kasson LR, et al. Complete Khoisan and Bantu genomes from southern Africa. *Nature*. 2010; 463(7283):943–947. <https://doi.org/10.1038/nature08795> PMID: 20164927
5. Ahn SM, Kim TH, Lee S, Kim D, Ghang H, Kim DS, et al. The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome research*. 2009; 19(9):1622–1629. <https://doi.org/10.1101/gr.092197.109> PMID: 19470904
6. Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, et al. The diploid genome sequence of an Asian individual. *Nature*. 2008; 456(7218):60–65. <https://doi.org/10.1038/nature07484> PMID: 18987735
7. Jiao YS, Du PF. Predicting Golgi-resident protein types using pseudo amino acid compositions: Approaches with positional specific physicochemical properties. *Journal of theoretical biology*. 2016; 391:35–42. <https://doi.org/10.1016/j.jtbi.2015.11.009> PMID: 26702543
8. Du P, Wang L. Predicting human protein subcellular locations by the ensemble of multiple predictors via protein-protein interaction network with edge clustering coefficients. *PloS one*. 2014; 9(1):e86879. <https://doi.org/10.1371/journal.pone.0086879> PMID: 24466278
9. Li P, Guo M, Wang C, Liu X, Zou Q. An overview of SNP interactions in genome-wide association studies. *Briefings in Functional Genomics*. 2014; p. elu036. PMID: 25241224
10. Zou Q, Wan S, Ju Y, Tang J, Zeng X. Pretata: predicting TATA binding proteins with novel features and dimensionality reduction strategy. *BMC Systems Biology*. 2016; 10(4):401. <https://doi.org/10.1186/s12918-016-0353-5> PMID: 28155714
11. Du P, Tian Y, Yan Y. Subcellular localization prediction for human internal and organelle membrane proteins with projected gene ontology scores. *Journal of theoretical biology*. 2012; 313:61–67. <https://doi.org/10.1016/j.jtbi.2012.08.016> PMID: 22960368
12. Bignell GR, Greenman CD, Davies H, Butler AP, Edkins S, Andrews JM, et al. Signatures of mutation and selection in the cancer genome. *Nature*. 2010; 463(7283):893–898. <https://doi.org/10.1038/nature08768> PMID: 20164919
13. Yuan LF, Ding C, Guo SH, Ding H, Chen W, Lin H. Prediction of the types of ion channel-targeted conotoxins based on radial basis function network. *Toxicology in Vitro*. 2013; 27(2):852–856. <https://doi.org/10.1016/j.tiv.2012.12.024> PMID: 23280100
14. Yang H, Tang H, Chen XX, Zhang CJ, Zhu PP, Ding H, et al. Identification of Secretory Proteins in *Mycobacterium tuberculosis* Using Pseudo Amino Acid Composition. *BioMed Research International*. 2016; 2016. <https://doi.org/10.1155/2016/5413903> PMID: 27597968

15. Chen XX, Tang H, Li WC, Wu H, Chen W, Ding H, et al. Identification of Bacterial Cell Wall Lyases via Pseudo Amino Acid Composition. *BioMed Research International*. 2016; 2016. <https://doi.org/10.1155/2016/1654623> PMID: 27437396
16. Zhao YW, Lai HY, Tang H, Chen W, Lin H. Prediction of phosphothreonine sites in human proteins by fusing different features. *Scientific reports*. 2016; 6. <https://doi.org/10.1038/srep34817> PMID: 27698459
17. Specht DF. Probabilistic neural networks. *Neural networks*. 1990; 3(1):109–118. [https://doi.org/10.1016/0893-6080\(90\)90049-Q](https://doi.org/10.1016/0893-6080(90)90049-Q)
18. Breiman L. Random forests. *Machine learning*. 2001; 45(1):5–32. <https://doi.org/10.1023/A:1010933404324>
19. Cai Yd, Lin SL. Support vector machines for predicting rRNA-, RNA-, and DNA-binding proteins from amino acid sequence. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*. 2003; 1648(1):127–133. [https://doi.org/10.1016/S1570-9639\(03\)00112-2](https://doi.org/10.1016/S1570-9639(03)00112-2) PMID: 12758155
20. Wei L, Xing P, Su R, Shi G, Ma ZS, Zou Q. CPPred-RF: a sequence-based predictor for identifying cell-penetrating peptides and their uptake efficiency. *Journal of Proteome Research*. 2017;. <https://doi.org/10.1021/acs.jproteome.7b00019> PMID: 28436664
21. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic acids research*. 2003; 31(13):3812–3814. <https://doi.org/10.1093/nar/gkg509> PMID: 12824425
22. Sim NL, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic acids research*. 2012; 40(W1):W452–W457. <https://doi.org/10.1093/nar/gks539> PMID: 22689647
23. Bromberg Y, Yachdav G, Rost B. SNAP predicts effect of mutations on protein function. *Bioinformatics*. 2008; 24(20):2397–2398. <https://doi.org/10.1093/bioinformatics/btn435> PMID: 18757876
24. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. *Current protocols in human genetics*. 2013; p. 7–20. <https://doi.org/10.1002/0471142905.hg0720s76> PMID: 23315928
25. Wang M, Zhao XM, Takemoto K, Xu H, Li Y, Akutsu T, et al. FunSAV: predicting the functional effect of single amino acid variants using a two-stage random forest model. *PloS one*. 2012; 7(8):e43847. <https://doi.org/10.1371/journal.pone.0043847> PMID: 22937107
26. Yates CM, Filippis I, Kelley LA, Sternberg MJ. SuSPect: enhanced prediction of single amino acid variant (SAV) phenotype using network features. *Journal of molecular biology*. 2014; 426(14):2692–2701. <https://doi.org/10.1016/j.jmb.2014.04.026> PMID: 24810707
27. Friedman JH. Greedy function approximation: a gradient boosting machine. *Annals of statistics*. 2001; p. 1189–1232. <https://doi.org/10.1214/aos/1013203451>
28. Friedman JH. Stochastic gradient boosting. *Computational Statistics & Data Analysis*. 2002; 38(4):367–378. [https://doi.org/10.1016/S0167-9473\(01\)00065-2](https://doi.org/10.1016/S0167-9473(01)00065-2)
29. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, et al. The universal protein resource (UniProt). *Nucleic acids research*. 2005; 33(suppl 1):D154–D159. <https://doi.org/10.1093/nar/gki070> PMID: 15608167
30. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, et al. Ensembl 2012. *Nucleic acids research*. 2011; p. gkr991.
31. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic acids research*. 2005; 33(suppl 1):D514–D517. <https://doi.org/10.1093/nar/gki033> PMID: 15608251
32. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, et al. A second generation human haplotype map of over 3.1 million SNPs. *Nature*. 2007; 449(7164):851–861. <https://doi.org/10.1038/nature06258> PMID: 17943122
33. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*. 2006; 22(13):1658–1659. <https://doi.org/10.1093/bioinformatics/btl158> PMID: 16731699
34. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*. 1997; 25(17):3389–3402. <https://doi.org/10.1093/nar/25.17.3389> PMID: 9254694
35. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The protein data bank. *Nucleic acids research*. 2000; 28(1):235–242. <https://doi.org/10.1093/nar/28.1.235> PMID: 10592235
36. Cheng J, Randall AZ, Sweredoski MJ, Baldi P. SCRATCH: a protein structure and structural feature prediction server. *Nucleic acids research*. 2005; 33(suppl 2):W72–W76. <https://doi.org/10.1093/nar/gki396> PMID: 15980571

37. Jones DT, Cozzetto D. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics*. 2015; 31(6):857–863. <https://doi.org/10.1093/bioinformatics/btu744> PMID: 25391399
38. Sunyaev SR, Eisenhaber F, Rodchenkov IV, Eisenhaber B, Tumanyan VG, Kuznetsov EN. PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. *Protein engineering*. 1999; 12(5):387–394. <https://doi.org/10.1093/protein/12.5.387> PMID: 10360979
39. Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB. Protein disorder prediction: implications for structural proteomics. *Structure*. 2003; 11(11):1453–1459. <https://doi.org/10.1016/j.str.2003.10.002> PMID: 14604535
40. Chan CH, Liang HK, Hsiao NW, Ko MT, Lyu PC, Hwang JK. Relationship between local structural entropy and protein thermostability. *Proteins: Structure, Function, and Bioinformatics*. 2004; 57(4):684–691. <https://doi.org/10.1002/prot.20263> PMID: 15532068
41. Kawashima S, Kanehisa M. AAindex: amino acid index database. *Nucleic acids research*. 2000; 28(1):374–374. <https://doi.org/10.1093/nar/28.1.374> PMID: 10592278
42. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*. 1992; 89(22):10915–10919. <https://doi.org/10.1073/pnas.89.22.10915> PMID: 1438297
43. Heffernan R, Paliwal K, Lyons J, Dehzangi A, Sharma A, Wang J, et al. Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. *Scientific reports*. 2015; 5. <https://doi.org/10.1038/srep11476> PMID: 26098304
44. Capra JA, Singh M. Predicting functionally important residues from sequence conservation. *Bioinformatics*. 2007; 23(15):1875–1882. <https://doi.org/10.1093/bioinformatics/btm270> PMID: 17519246
45. Miller MP, Kumar S. Understanding human disease mutations through the use of interspecific genetic variation. *Human molecular genetics*. 2001; 10(21):2319–2328. <https://doi.org/10.1093/hmg/10.21.2319> PMID: 11689479
46. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*. 1983; 22(12):2577–2637. <https://doi.org/10.1002/bip.360221211> PMID: 6667333
47. Liang S, Grishin NV. Effective scoring function for protein sequence design. *Proteins: Structure, Function, and Bioinformatics*. 2004; 54(2):271–281. <https://doi.org/10.1002/prot.10560> PMID: 14696189
48. Liang S, Meroueh SO, Wang G, Qiu C, Zhou Y. Consensus scoring for enriching near-native structures from protein–protein docking decoys. *Proteins: Structure, Function, and Bioinformatics*. 2009; 75(2):397–403. <https://doi.org/10.1002/prot.22252> PMID: 18831053
49. Dobson RJ, Munroe PB, Caulfield MJ, Saqi MA. Predicting deleterious nsSNPs: an analysis of sequence and structural attributes. *BMC bioinformatics*. 2006; 7(1):1. <https://doi.org/10.1186/1471-2105-7-217> PMID: 16630345
50. David A, Sternberg MJ. The contribution of missense mutations in core and rim residues of protein–protein interfaces to human disease. *Journal of molecular biology*. 2015; 427(17):2886–2898. <https://doi.org/10.1016/j.jmb.2015.07.004> PMID: 26173036
51. Saunders CT, Baker D. Evaluation of structural and evolutionary contributions to deleterious mutation prediction. *Journal of molecular biology*. 2002; 322(4):891–901. [https://doi.org/10.1016/S0022-2836\(02\)00813-6](https://doi.org/10.1016/S0022-2836(02)00813-6) PMID: 12270722
52. Hubbard SJ, Thornton JM. Naccess. Computer Program, Department of Biochemistry and Molecular Biology, University College London. 1993; 2(1).
53. Petersen B, Petersen TN, Andersen P, Nielsen M, Lundegaard C. A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC structural biology*. 2009; 9(1):1. <https://doi.org/10.1186/1472-6807-9-51>
54. Song J, Tan H, Takemoto K, Akutsu T. HSEpred: predict half-sphere exposure from protein sequences. *Bioinformatics*. 2008; 24(13):1489–1497. <https://doi.org/10.1093/bioinformatics/btn222> PMID: 18467349
55. Hamelryck T. An amino acid has two sides: a new 2D measure provides a different view of solvent exposure. *Proteins: Structure, Function, and Bioinformatics*. 2005; 59(1):38–48. <https://doi.org/10.1002/prot.20379> PMID: 15688434
56. Li Y, Wen Z, Xiao J, Yin H, Yu L, Yang L, et al. Predicting disease-associated substitution of a single amino acid by analyzing residue interactions. *BMC bioinformatics*. 2011; 12(1):1. <https://doi.org/10.1186/1471-2105-12-14> PMID: 21223604
57. Chakrabarty B, Parekh N. NAPS: Network Analysis of Protein Structures. *Nucleic acids research*. 2016; p. gkw383. <https://doi.org/10.1093/nar/gkw383> PMID: 27151201

58. Deng L, Guan J, Dong Q, Zhou S. Prediction of protein-protein interaction sites using an ensemble method. *BMC bioinformatics*. 2009; 10(1):1. <https://doi.org/10.1186/1471-2105-10-426> PMID: [20015386](https://pubmed.ncbi.nlm.nih.gov/20015386/)
59. Deng L, Guan J, Wei X, Yi Y, Zhang QC, Zhou S. Boosting prediction performance of protein–protein interaction hot spots by using structural neighborhood properties. *Journal of Computational Biology*. 2013; 20(11):878–891. <https://doi.org/10.1089/cmb.2013.0083> PMID: [24134392](https://pubmed.ncbi.nlm.nih.gov/24134392/)
60. Chao F, Liu D, Rui H, Chen Z, Lei D. PredRSA: a gradient boosted regression trees approach for predicting protein solvent accessibility. *Bmc Bioinformatics*. 2016; 17 Suppl 1(S1):85–95. <https://doi.org/10.1186/s12859-015-0851-2> PMID: [26818760](https://pubmed.ncbi.nlm.nih.gov/26818760/)
61. Barber CB, Dobkin DP, Huhdanpaa H. The quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software (TOMS)*. 1996; 22(4):469–483. <https://doi.org/10.1145/235815.235821>
62. Zou Q, Zeng J, Cao L, Ji R. A novel features ranking metric with application to scalable visual and bioinformatics data classification. *Neurocomputing*. 2016; 173:346–354. <https://doi.org/10.1016/j.neucom.2014.12.123>
63. Lin H, Chen W. Prediction of thermophilic proteins using feature selection technique. *Journal of microbiological methods*. 2011; 84(1):67–70. <https://doi.org/10.1016/j.mimet.2010.10.013> PMID: [21044646](https://pubmed.ncbi.nlm.nih.gov/21044646/)
64. Wei L, Xing P, Shi G, Ji ZL, Zou Q. Fast prediction of protein methylation sites using a sequence-based feature selection technique. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. 2017;. <https://doi.org/10.1109/TCBB.2017.2670558> PMID: [28222000](https://pubmed.ncbi.nlm.nih.gov/28222000/)
65. Ding H, Feng PM, Chen W, Lin H. Identification of bacteriophage virion proteins by the ANOVA feature selection and analysis. *Molecular BioSystems*. 2014; 10(8):2229–2235. <https://doi.org/10.1039/C4MB00316K> PMID: [24931825](https://pubmed.ncbi.nlm.nih.gov/24931825/)
66. Meinshausen N, Bühlmann P. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2010; 72(4):417–473. <https://doi.org/10.1111/j.1467-9868.2010.00740.x>
67. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*. 2011; 12:2825–2830.
68. Peng H, Long F, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*. 2005; 27(8):1226–1238. <https://doi.org/10.1109/TPAMI.2005.159> PMID: [16119262](https://pubmed.ncbi.nlm.nih.gov/16119262/)
69. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. *Machine learning*. 2002; 46(1–3):389–422. <https://doi.org/10.1023/A:1012487302797>
70. Bao L, Zhou M, Cui Y. nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms. *Nucleic acids research*. 2005; 33(suppl 2):W480–W482. <https://doi.org/10.1093/nar/gki372> PMID: [15980516](https://pubmed.ncbi.nlm.nih.gov/15980516/)
71. Flydal MI, Martinez A. Phenylalanine hydroxylase: function, structure, and regulation. *IUBMB life*. 2013; 65(4):341–349. <https://doi.org/10.1002/iub.1150> PMID: [23457044](https://pubmed.ncbi.nlm.nih.gov/23457044/)
72. Andersen OA, Flatmark T, Hough E. High resolution crystal structures of the catalytic domain of human phenylalanine hydroxylase in its catalytically active Fe (II) form and binary complex with tetrahydrobiopterin. *Journal of molecular biology*. 2001; 314(2):279–291. <https://doi.org/10.1006/jmbi.2001.5061> PMID: [11718561](https://pubmed.ncbi.nlm.nih.gov/11718561/)
73. Baker CH, Matsuda SP, Liu DR, Corey E. Molecular-cloning of the human gene encoding lanosterol synthase from a liver cDNA library. *Biochemical and biophysical research communications*. 1995; 213(1):154–160. <https://doi.org/10.1006/bbrc.1995.2110> PMID: [7639730](https://pubmed.ncbi.nlm.nih.gov/7639730/)
74. Thoma R, Schulz-Gasch T, D'arcy B, Benz J, Aebi J, Dehmow H, et al. Insight into steroid scaffold formation from the structure of human oxidosqualene cyclase. *Nature*. 2004; 432(7013):118–122. <https://doi.org/10.1038/nature02993> PMID: [15525992](https://pubmed.ncbi.nlm.nih.gov/15525992/)
75. DeLano WL. The PyMOL molecular graphics system. 2002;.
76. Schwede T, Kopp J, Guex N, Peitsch MC. SWISS-MODEL: an automated protein homology-modeling server. *Nucleic acids research*. 2003; 31(13):3381–3385. <https://doi.org/10.1093/nar/gkg520> PMID: [12824332](https://pubmed.ncbi.nlm.nih.gov/12824332/)
77. Hardin C, Pogorelov TV, Luthey-Schulten Z. Ab initio protein structure prediction. *Current opinion in structural biology*. 2002; 12(2):176–181. [https://doi.org/10.1016/S0959-440X\(02\)00306-8](https://doi.org/10.1016/S0959-440X(02)00306-8) PMID: [11959494](https://pubmed.ncbi.nlm.nih.gov/11959494/)
78. Zhang CJ, Tang H, Li WC, Lin H, Chen W, Chou KC. iOri-Human: identify human origin of replication by incorporating dinucleotide physicochemical properties into pseudo nucleotide composition. *Oncotarget*. 2016; 7(43):69783–69793. <https://doi.org/10.18632/oncotarget.11975> PMID: [27626500](https://pubmed.ncbi.nlm.nih.gov/27626500/)
79. Lin H, Liang ZY, Tang H, Chen W. Identifying sigma70 promoters with novel pseudo nucleotide composition. *IEEE/ACM transactions on computational biology and bioinformatics*. 2017;. <https://doi.org/10.1109/TCBB.2017.2666141> PMID: [28186907](https://pubmed.ncbi.nlm.nih.gov/28186907/)

80. Liang ZY, Lai HY, Yang H, Zhang CJ, Yang H, Wei HH, et al. Pro54DB: a database for experimentally verified sigma-54 promoters. *Bioinformatics*. 2016; p. btw630. <https://doi.org/10.1093/bioinformatics/btw630>
81. Deng L, Zhang QC, Chen Z, Meng Y, Guan J, Zhou S. PredHS: a web server for predicting protein-protein interaction hot spots by using structural neighborhood properties. *Nucleic Acids Research*. 2014; 42(Web Server issue):290–5. <https://doi.org/10.1093/nar/gku437> PMID: 24852252
82. Garzón JI, Deng L, Murray D, Shapira S, Petrey D, Honig B. A computational interactome and functional annotation for the human proteome. *Elife*. 2016; 5:e18715. <https://doi.org/10.7554/eLife.18715> PMID: 27770567