Original Research

Incorporating structural features to improve the prediction and understanding of pathogenic amino acid substitutions

Yao Xiong[1,2], Jing-Bo Zhou[1], Ke An[1], Wei Han[1], Tao Wang[3], Zhi-Qiang Ye[1,3,*], Yun-Dong Wu[1,3,4,*]

[1]*State Key Laboratory of Chemical Oncogenomics, Peking University Shenzhen Graduate School, 518055 Shenzhen, Guangdong, China,* [2]*Assisted Reproduction Center, Northwest Women's and Children's Hospital, 710003 Xi'an, Shaanxi, China,* [3]*Shenzhen Bay Laboratory, 518055 Shenzhen, Guangdong, China,* [4]*College of Chemistry and Molecular Engineering, Peking University, 100871 Beijing, China*

TABLE OF CONTENTS

1. Abstract

Background: The wide application of gene sequencing has accumulated numerous amino acid substitutions (AAS) with unknown significance, posing significant challenges to predicting and understanding their pathogenicity. While various prediction methods have been proposed, most are sequence-based and lack insights for molecular mechanisms from the perspective of protein structures. Moreover, prediction performance must be improved. **Methods**: Herein, we trained a random forest (RF) prediction model, namely AAS3D-RF, underscoring sequence and three-dimensional (3D) structure-based features to explore the relationship between diseases and AASs. **Results**: AAS3D-RF was trained on more than 14,000 AASs with 21 selected features, and obtained accuracy (ACC) between 0.811 and 0.839 and Matthews correlation coefficient (MCC) between 0.591 and 0.684 on two independent testing datasets, superior to seven existing tools. In addition, AAS3D-RF possesses unique structure-based features, context-dependent substitution score (CDSS) and environment-dependent residue contact energy (ERCE), which could be applied to interpret whether pathogenic AASs would introduce incompatibilities to the protein structural microenvironments. **Conclusion**: AAS3D-RF serves as a valuable tool for both predicting and understanding pathogenic AASs.

Submitted: 6 July 2021 Revised: 9 October 2021 Accepted: 21 October 2021 Published: 30 December 2021

2. Introduction

The development of high-throughput sequencing technologies continuously accelerates human genome re-sequencing and the identification of variants [1]; a vast quantity of single-nucleotide variants (SNV) have been registered in the dbSNP database to date [2]. Among them, non-synonymous single-nucleotide variants (nsSNV), which lead to amino acid substitutions (AAS) in their protein products, account for more than half of disease-associated genetic lesions [3, 4], and are thus of great interest. While only a small part of nsSNVs has been confirmed to be related to disease or not, the vast majority remain with uncertain significance (variants of uncertain significance, or VUS). In detail, the dbSNP contains about 8 million nsSNVs or AASs currently, but only about 0.1 million (~1.25%) have explicit associations with clinical phenotypic effects according to annotations in ClinVar [5], UniProt [6], and HGMD [7]. As it is impractical to characterize each VUS through experimental approaches, *in silico* prediction of their disease-association and understanding of the molecular basis of their pathogenicity have become in high demand.

The past decades have witnessed the development of various computational predictors aimed at screening disease-associated AASs (daAASs) from neutral ones (nAASs) [8–15]. Many of them adopted machine learning strategies to train predictors based on various features, especially sequence conservation-related features [10–12, 14]. However, the performance of available predictors still leaves room for improvement, and these predictors often fail to provide insights into the molecular basis of daAASs. To advance in these aspects, it will be promising to explore more informative and interpretable features.

Sequence conservation-related features have been demonstrated as a group of powerful features, but they cannot provide in-depth mechanistic hints since the conservation actually serves as a result (but not cause) of natural selection for structural and functional importance. It is the folded three-dimensional (3D) structure that directly fulfills the function. Hence, it is hoped that exploring structural features can provide more interpretable insights to understand the underlying mechanisms of the pathogenicity of daAASs. With a large collection of protein 3D structures that has accumulated in recent years [16], it has proven to be promising to further explore features based on them, and to develop predictors by combining both sequence and the mined structural features. However, it is challenging to put this idea into practice as the complexity of protein structure data makes their high-throughput processing much more difficult than sequence data.

Several recent studies have put massive effort in this direction. PhyreRisk can map AASs to experimental or homology-modeled structures, and the associated Missense3D tool can then list their potential structural impacts according to a set of knowledge-based rules, such as breaking the disulfide bond, burying the hydrogen bond, introducing clash, altering secondary structures, etc. [17, 18]. VarSite, a protein-centered, experimental structure-focused resource, has integrated various features, including tissue-specific expression, disease type, conservation, protein domain, secondary structure, and interaction site [19]. Users can inspect AASs in the contexts of these features to understand their structural basis [19]. The mutfunc tool provides pre-calculated properties associated with datasets curated from ExAC [20] and ClinVar [5], including stability, interaction, post-translational modification, linear motif, and transcription factor binding site [21]. Moreover, it has covered non-coding variants and extended from human to yeast and *E. coli* [21]. MISCAST has analyzed 40 properties associated with the AAS position for all protein classes jointly and for each of 24 protein functional classes separately, and have identified many properties that are significantly associated with pathogenic or neutral AASs, including protein-protein interactions, residue exposure levels, secondary structures, etc. [22]. Further, MISCAST defined the P3DFi scores for each AAS position on the basis of the analyses of these properties, and the trained machine-learning model has shown that the P3DFi scores can offer orthogonal information to improve the prediction of pathogenic AAS compared with the combination of SIFT [8], PolyPhen-2 [11], and CADD [23]. These studies have largely enhanced the interpretation of AASs from the perspective of structures and functions.

Herein, we explored a more comprehensive set of structural and sequence features, selected an optimal feature subset using an automatic pipeline, and trained a machine-learning model for predicting the pathogenicity of AASs. First, we curated a high-quality collection of experimental structures and homology structure models for 5278 and 3682 human proteins, respectively. Second, one training and two testing datasets were constructed with 14,117, 5485, and 5249 AASs that can be mapped to structures, respectively. Third, a total of 212 candidate features for each AAS were extracted based on sequence or structure, and 21 of them were selected by automatic feature selection to train a random forest predictor [24], namely, AAS3D-RF. Fourth, the evaluations demonstrated that AAS3D-RF outperformed seven other popular tools. The structural features selected in this work improve the prediction performance to different degrees in different scenarios, and the nature of their interpretability also provides more understanding of the molecular basis of the daAASs.

3. Materials and methods

3.1 Curation of AAS datasets

For all human proteins in the Swiss-Prot database (UniProtKB/Swiss-Prot, Release 2018, 04) [6], we down-

loaded and curated their available experimental structures from Protein Data Bank (PDB) [16]. For those proteins without available experimental structures, their homology models from ModBase [25] were adopted. The processing details are described in **Supplementary Fig. 1** and **Supplementary Methods**.

The humsavar.txt (2018, 04 of 25 Apr 2018) file provided by the UniProt FTP server contained manually annotated AAS, and is the main source for preparing the AAS datasets [6]. According to its documentation, AASs labeled with "Disease" serve as positive samples (i.e., daAAS), while those labeled with "Polymorphism" (similar to "neutral" or "benign") are negative samples (i.e., nAAS). Another two data sources of AASs are the 1000 Genomes Project (1000G, Release 2 May 2013) [26] and the VariSNP (Release 16 Feb 2017) [27]. After removing those whose mutant allele frequency is less than 1%, the remaining AASs from 1000G and VariSNP were regarded as nAASs. After mapping to experimental structures and keeping only one instance for duplicates, the AASs from these three sources constituted the *TotalDataset*.

The *TotalDataset* was then split into *TrainDataset* and *TestDataset 1* according to the following procedure: After an all-to-all BLASTP (E-value $\leq$0.01) run among the protein sequences in *TotalDataset* [28], the proteins with sequence identity to any other one of less than 30% were selected, and their AASs were regarded as the *TestDataset 1*. The AASs in the remaining proteins served as the *TrainDataset*. On the other hand, the AASs in humsavar.txt located at homology structure models constituted the *TestDataset 2*. The detailed AAS mapping process is also illustrated in **Supplementary Fig. 2**.

A series of subsets were further prepared from these three datasets. In detail, the subsets were organized according to the proportion of daAASs on each protein: For proteins containing only daAASs or nAASs, their AASs constitute the "Pure" subsets; other subsets, namely, "Mixed" subsets at different mixing levels, were constructed by selecting proteins within specific ranges of daAASs proportion, including open interval between 0.0 and 1.0 (denoted as]0.0, 1.0[) and close intervals of [0.1, 0.9], [0.2, 0.8], [0.3, 0.7], and [0.4, 0.6].

3.2 Feature calculation

Based on protein structures, sequences, and sequence alignments, a total of 212 candidate features were calculated to characterize each AAS from various perspectives (gene/protein, AAS site, substitution itself, etc.). The full list of features is described in **Supplementary Methods**.

Each dataset was organized as a feature matrix with each row representing an AAS and each column corresponding to a feature. The missing data were filled with the mean values derived from the *TrainDataset*. Furthermore, each feature column was transformed into Z-score, i.e., all values in the same feature column were standardized with the mean and standard deviation derived from the *TrainDataset*.

3.3 Feature selection and predictor training

Random forest (RF) is a machine learning framework consisting of an ensemble of decision trees, and has been widely applied in classification problems [24]. In the training stage, each tree is trained with only a part of the training samples to decrease over-fitting. In the stage of prediction, the consensus output from the majority of the trees was taken as the final result. Considering its successful application in many bioinformatics studies [29], we chose the RF technique in this work for automatic feature selection and model training.

We adopted a wrapper strategy in the feature selection procedure. That is, we iteratively generated feature subsets, and evaluated them by using the 10-fold cross-validation performance of RF classifiers trained on them accordingly. The brief logic was that we added two features with the most contribution and removed one feature with the least feature importance in each iteration, and this operation would be terminated to deduce the feature redundancy if the adding of new features would not improve the performance of the classifier any more. This procedure has been demonstrated effective in one previous study [30]. Notably, we conducted the cross-validation at the protein level, where AASs from the same protein were required to reside in the same fold. That is, AASs were grouped at the protein level, namely, Group10Fold cross-validation (G10F-CV). The purpose of this operation was to reduce the level of type 2 circularity, which has been discussed in-depth previously [31]: If protein/gene-level features were used, there would exist over-fitting in predicting the disease-association of those AASs whose proteins contain AASs used in the training procedure.

After obtaining the final feature subset, we determined the optimal RF hyper-parameters by finding the maximum area under the ROC curve (AUC) of G10F-CV in a random search of 1000 trials (**Supplementary Table 1**). Then, by specifying the hyper-parameters with the optimal values, the final classifier was re-trained on the entire *TrainDataset*. All these processes were conducted by using the Scikit-learn toolkit v0.19.0 [32].

3.4 Performance evaluation and comparison with other methods

The two independent testing datasets, i.e., *TestDataset 1* and *TestDataset 2*, were utilized to evaluate performance and to compare with other predictors by adopting standard performance measures, including accuracy (ACC), AUC, Matthews correlation coefficient (MCC), sensitivity (Sen), specificity (Spe), positive predictive value (PPV), and negative predictive value (NPV) (definitions in **Supplementary Methods**).

In addition to the evaluations on the full datasets of *TestDataset 1* and *TestDataset 2*, more in-depth evaluations were performed on the 'Pure' and 'Mixed' testing subsets, which are designed to provide an examination of the extent of type 2 circularity [31].

Our predictor was compared with seven popular tools, including SIFT (version 5.2.2) [8], HumDiv-trained PolyPhen-2 (PPH2_HD) (version 2.2.2r405c) and HumVar-trained PolyPhen-2 (PPH2_HV) (version 2.2.2r405c) [11], PROVEAN (version 1.1.5) [13], FATHMM's weighted method (FATHMM-W) and unweighted method (FATHMM-U) (version 2.3) [14], and PANTHER-PSEP (version 1.01) [15]. The settings of these tools are described in detail in **Supplementary Methods**.

Table 1. Overview of the AAS datasets.

Dataset	Class	# of AAS	# of Proteins[1]	# of Structures[1]
TrainDataset	Disease	7385	611	708
	Neutral	6732	1750	1938
	Total	14,117	1979	2239
TestDataset 1	Disease	3772	321	338
	Neutral	1713	733	757
	Total	5485	834	873
TestDataset 2	Disease	2471	383	394
	Neutral	2778	1208	1265
	Total	5249	1418	1482

[1]The number of "Total" is less than the sum of "Disease" and "Neutral" since one protein or structure may contain daAASs and nAASs at the same time.

4. Results

4.1 Overview of the datasets

In summary, 6430 experimental structures of 5278 proteins and 4238 homology models of 3682 proteins were obtained for preparing the AAS datasets (**Supplementary Fig. 1**).

After initial data cleaning, the humsavar dataset retained 29,328 daAASs and 39,679 nAASs. Among them, 11,157 daAASs and 7072 nAASs were mapped to experimental structures, and 2471 daAASs and 2778 nAASs were mapped to ModBase [25] homology models. Overall, the mapped percentages were 46.5% and 24.8% for daAASs and nAASs, respectively. For the VariSNP dataset, the initial cleaning, keeping those with MAF at no less than 1%, and mapping to UniProt canonical sequences, resulted in 18,233 nAASs. Among them, 1281 (7.0%) were mapped to experimental structures. For the 1000G dataset, 3296 of 34,791 (9.5%) nAASs were mapped to experimental structures. After integration and proper data partition described in section 3.1, these mapped AASs were separated into *TrainDataset*, *TestDataset 1*, and *TestDataset 2* (**Supplementary Fig. 2**).

The *TrainDataset* consists of 14,117 AASs mapped on the experimental structures of 1979 proteins, with 7385 daAASs on 611 proteins and 6732 nAASs on 1750 proteins (Table 1). The ratio of daAAS to nAAS is highly balanced (7385:6732), which will ensure that the trained classifier would not suffer from data bias [33, 34].

The *TestDataset 1* contains 5485 AASs mapped on the experimental structures from 834 proteins, with 3772 daAASs on 321 proteins and 1713 nAASs on 733 proteins (Table 1). The *TestDataset 2* contains 5249 AASs mapped on homology models of 1418 proteins, with 2471 daAASs on 383 proteins and 2778 nAASs on 1208 proteins (Table 1). The *TestDataset 1* and *TestDataset 2* were utilized to evaluate the performance of the trained predictor on AASs with features extracted from experimental structures and reliable homology models, respectively. Notably, both the *TestDataset 1* and *TestDataset 2* have no overlap with *TrainDataset* at either the substitution level or the protein level, which would presumably avoid the two types of circularity that may cause overly optimistic estimation of performance [31]. Moreover, the sequence identity of any pairwise comparison between *TestDataset 1* and *TrainDataset* is less than 30%, but 947 of 1418 proteins in *TestDataset 2* have high scoring pairs with sequence identity $\geq$30% with those in *TrainDataset*, indicating that *TestDataset 1* is more rigorous than *TestDataset 2*. The details of *TrainDataset*, *TestDataset 1*, and *TestDataset 2* are provided at http://www.wdspdb.com/AAS3D-RF/.

We inspected the datasets by examining the "Pure" and "Mixed" subsets (Fig. 1A,B). A substantial portion of AASs came from the "Pure" subsets (32.6% in *TestDataset 1* and 65.1% in *TestDataset 2*), i.e., many proteins contributed only daAASs or only nAASs. These "Pure" or "Mixed" testing subsets can provide more detailed comparisons of the predictors' performance, as some predictors confounded by type 2 circularity could not be evaluated properly on the entire testing dataset [31].

A similar pattern can be found in the *TrainDataset*, where a large part of AASs (44.3%) are from the "Pure" subset (**Supplementary Fig. 3**). In addition, considering that some features adopted in the classifier describe the whole gene or protein but not the AAS itself, the trained classifier may be skewed to classify protein but not the AAS (type 2 circularity). As we had adopted protein level cross-validation (G10F-CV in this work) in the training procedure [31], this skewness could presumably be removed.

4.2 The predictor and its performance on the testing datasets

A total of 21 features (**Supplementary Table 2**) were finally selected after the iterative feature selection procedure. After randomly searching 1000 hyper-parameter combinations of RF, we obtained the optimal one (n estimators: 338, max depth: 10, max features: 0.5233959) corresponding to the highest AUC (0.873) in the G10F-CV. By specifying the optimal hyper-parameter combina-

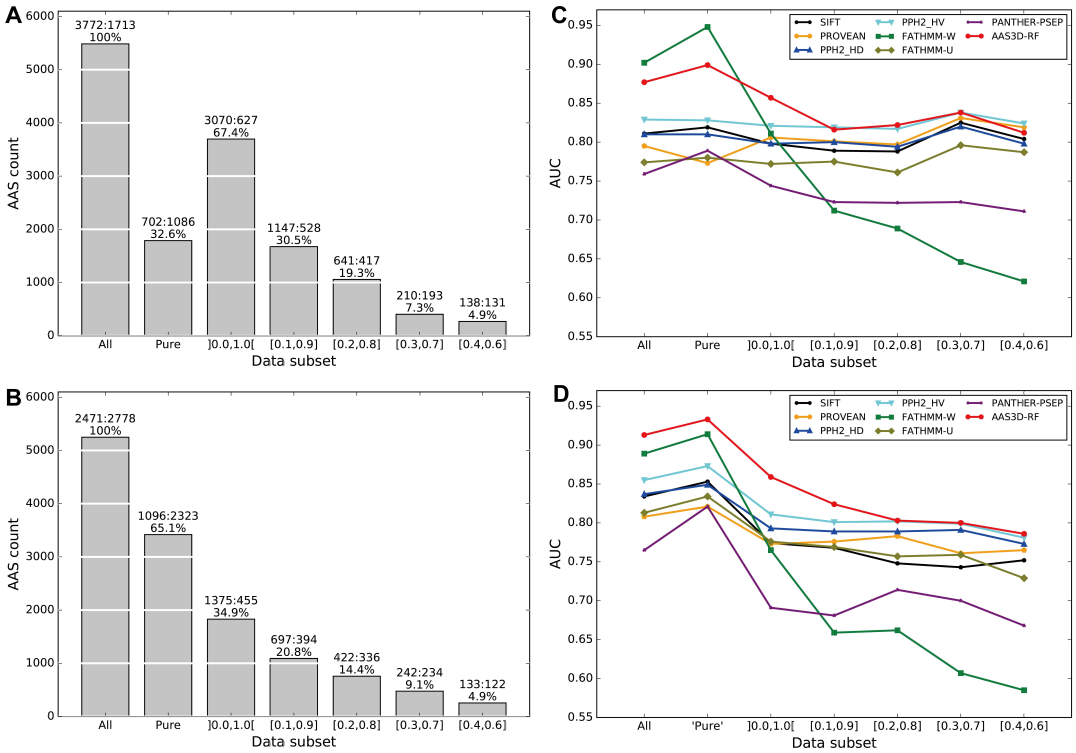

Fig. 1. The composition of different AAS subsets and the performance of eight predictors on them. Each subset was prepared by incorporating the AASs whose proteins harbor a specific range of daAASs proportion (interval labels under the X axis) for *TestDataset 1* (A) and *TestDataset 2* (B). Among them, "All" represents the whole testing dataset, while "Pure" contains the AASs whose proteins harbor either daAASs or nAASs only. The numbers of daAASs and nAASs of each subset are given on top of each bar. The percentage of the AAS number in each subset is also indicated. The AUC scores on different AAS subsets are plotted for *TestDataset 1* (C) and *TestDataset 2* (D).

tion, we re-trained the final RF classifier, namely AAS3D-RF, on *TrainDataset*. We also implemented an automatic pipeline that can calculate the required features for a given AAS, and can then load AAS3D-RF to make predictions. The codes were implemented in Python 2.7 and are available at http://www.wdspdb.com/AAS3D-RF/ and https://github.com/PKU-XiongYao/AAS3D-RF.

The measures of the performance on *TestDataset 1* and *TestDataset 2* are listed in Table 2 and **Supplementary Table 3**. The ACC and MCC values are 0.811 and 0.591 for *TestDataset 1*, and 0.839 and 0.684 for *TestDataset 2*, demonstrating its superior overall performance. The better performance on *TestDataset 2* may partially stem from the similar proteins between *TestDataset 2* and *TrainDataset*, since we did not require that the proteins having HSP with sequence identity $\geq$30% be excluded from *TestDataset 2*. When removing this part of data from *TestDataset 2*, we can observe that the MCC drops from 0.684 to 0.627, supporting our speculation very well (**Supplementary Table 3**). In the real-world application, one may often need to predict the disease-association of an AAS whose protein is similar to some protein in the training set, so the expected performance would presumably be better than reported here.

Table 2. Performance of AAS3D-RF and the structure-removed predictor on the two independent testing datasets.

Predictor	Dataset	ACC	MCC
AAS3D-RF	*TestDataset 1*	0.811	0.591
	TestDataset 2	0.839	0.684
structure-removed	*TestDataset 1*	0.783	0.542
	TestDataset 2	0.835	0.676

4.3 Comparison with other prediction methods

We compared the performance of AAS3D-RF with seven other popular tools. Except FATHMM-W (discussed below), AAS3D-RF outperforms all the other predictors on *TestDataset 1* and *TestDataset 2* in terms of ACC, AUC, MCC, and PPV (Fig. 2A,B, **Supplementary Table 4**). In particular, the MCC values are 8.7 and 11.3 percentage points higher than the second-best predictors (except FATHMM-W) on *TestDataset 1* and *TestDataset 2*, respectively (**Supplementary Table 4**). Notably, many other tools, including SIFT, PPH2_HD, PPH2_HV, PROVEAN, and PANTHER_PSEP, have achieved much higher Sen scores, but their Spe scores are very low, indicating that they are biased to positive predictions and prone to higher false positive rates (1-Spe). As MCC is a performance met-

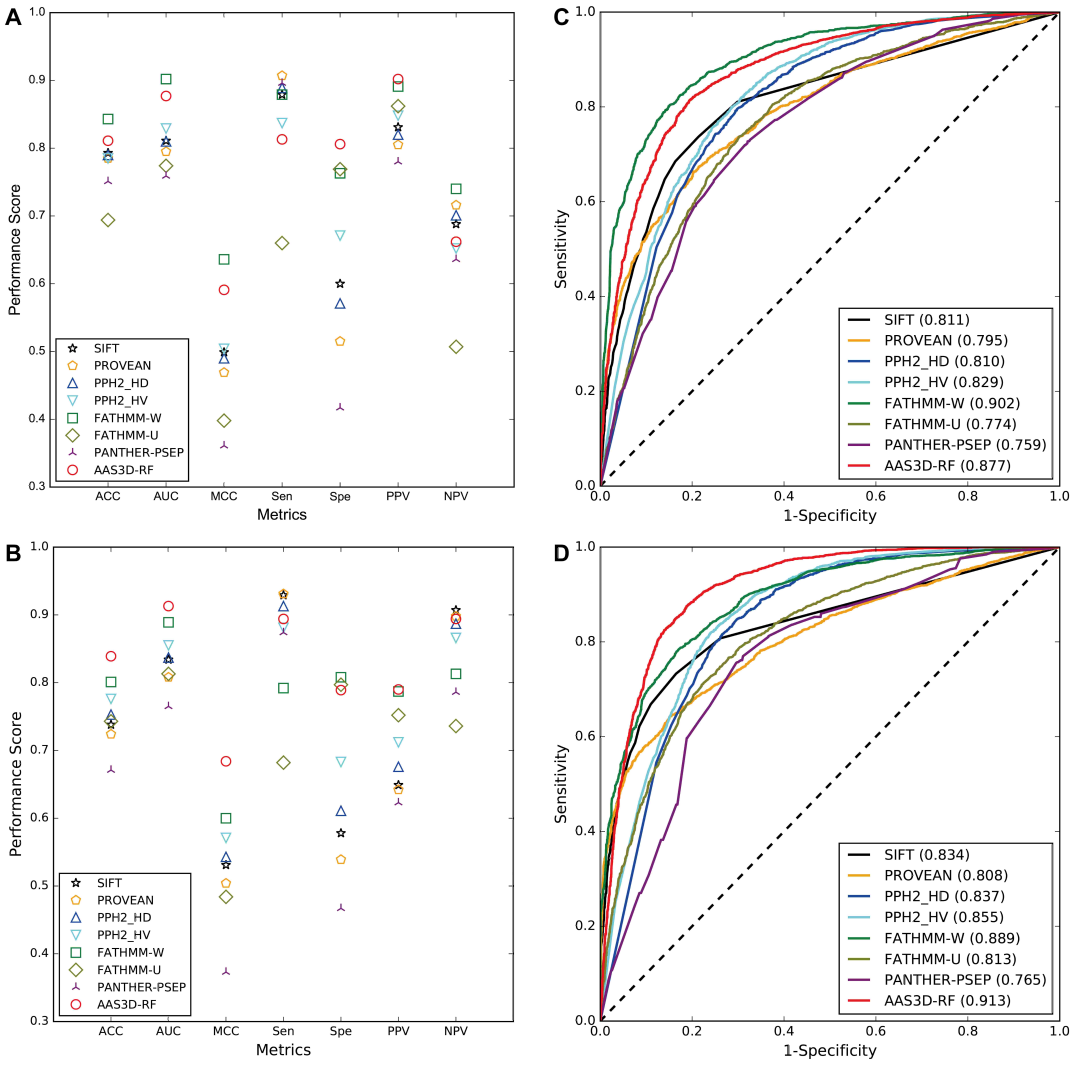

Fig. 2. Comprehensive performance comparison of eight predictors on the two independent testing datasets. (A) and (B) show the performance in terms of seven metrics for *TestDataset 1* and *TestDataset 2*, respectively. (C) and (D) display the ROC curves for *TestDataset 1* and *TestDataset 2*, respectively.

ric considering both positive and negative predictions in balance, we can conclude that AAS3D-RF achieved superior balanced performance without bias to a specific class. The detailed ROC curves demonstrate a similar conclusion: The curve of AAS3D-RF covers the largest area under the curve except FATHMM-W (Fig. 2C,D).

Several previous studies have reported that the prediction performance of FATHMM-W was significantly confounded by type 2 circularity, i.e., it intended to predict all the variants from the same protein as pathogenic or neutral as a whole [31, 35, 36]. In other words, FATHMM-W would perform worse on a dataset with proteins containing a nearly equal number of daAASs and nAASs on each of them (e.g., the subsets with daAAS proportion in the range of [0.4, 0.6]) than on a dataset with proteins containing only daAASs or nAASs (e.g., the "Pure" subsets). Prediction methods sensitive to type 2 circularity would perform poorly in discriminating daAASs from nAASs within a given protein. In our work, the evaluations on "Pure" and different levels of "Mixed" subsets have demonstrated this: FATHMM-W performs the worst on the subsets with daAAS proportion in [0.4, 0.6], but obtains an impressively high AUC on the "Pure" subsets (Fig. 1C,D). As shown, AAS3D-RF consistently ranks at the top tier of all methods among "Mixed" subsets, indicating that the type 2 circularity in AAS3D-RF has been removed at the highest level. As their values are in fact the same for all variants from the same protein, the gene/protein-level features should be the source of type 2 circularity. In our cross-validation and independent testing, the variants from the same protein were either all in the training set or all in the testing sets, hence this grouped cross-validation at the protein level here may have served as an effective strategy to decrease type 2 circularity.

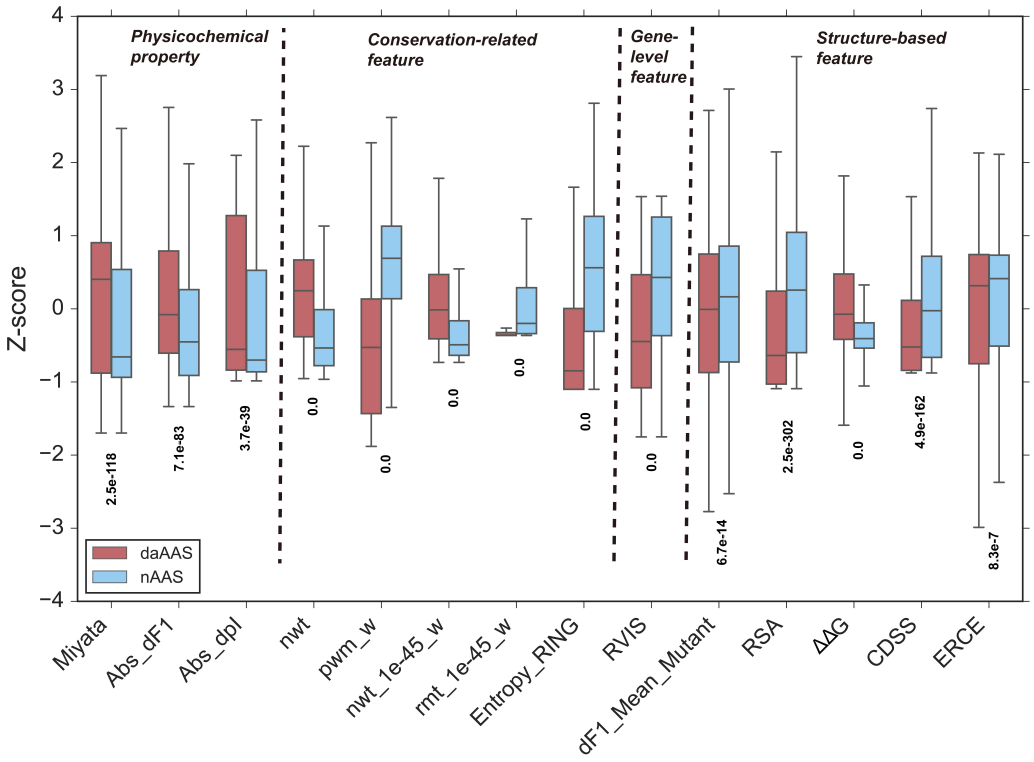

Fig. 3. Z-score distributions of a part of selected features for the daAASs (red) and nAASs (blue) in *TrainDataset*. The *p*-values of two-tailed Mann-Whitney test for each feature comparison are shown.

As published tools were trained on different datasets, it is often difficult to obtain a proper and fair testing dataset that contains enough data and has no overlap with any of their training datasets. Here, *TestDataset 1* and *TestDataset 2* are sufficiently rigorous to our predictor, since there are no overlapped AASs or proteins between them and *TrainDataset*. However, this scenario does not apply to other tools. Hence, the performance comparisons based on *TestDataset 1* and *TestDataset 2* will presumably provide a realistic performance estimate for AAS3D-RF but may supply overly optimistic estimates for other tools. In a truly fair comparison, the improvement of AAS3D-RF over other tools would presumably be larger.

4.4 Structural features improve the interpretability of the pathogenicity of AAS

To interrogate the interpretability of features, which may offer clues for understanding the molecular basis of the daAASs, we plotted the distributions of some features in daAAS and nAAS separately (Fig. 3). As shown, these features can be grouped within four categories: physicochemical properties of residues, conservation-related properties, gene/protein-level features, and structure-based features. While many of them have been analyzed in previous studies, we here mainly focus on several structure-based features.

Five features in the conservation-related group show the most evident contrast between daAASs and nAASs (Fig. 3). Conservation often indicates structural stability or functional importance. As a comprehensive result of many underlying mechanisms, the conservation-related features have superior ability in separating daAASs from nAASs. Many previous studies have also repeatedly revealed the power of this type of features [34, 37–43]. However, these features cannot offer further molecular mechanistic clues of the disease-associated AASs, as mentioned in the Introduction.

Relative solvent accessibility (RSA) and folding free energy change ($\Delta\Delta G$) are two widely used features derived from 3D structures in discriminating daAASs from nAASs [11, 41, 44–48]. Fig. 3 shows that daAASs are more likely to be located at buried sites (small RSA values). The AASs occurring in buried sites may destabilize the protein by distorting the hydrophobic core, thus resulting in pathogenic effects [49, 50]. More directly than RSA, $\Delta\Delta G$ measures the stability change caused by the substitution itself. In our dataset, the $\Delta\Delta G$ values of daAASs are much larger than those of nAASs (Fig. 3), directly demonstrating that daAASs are more prone to large depletion of stability. This also suggests that stability loss resulting from AAS serves as a major mechanism of their disease-association.

The context-dependent substitution score (CDSS) is a substitution score between different amino acids under specific structural environments or contexts. Three features, including Miyata, Abs_dF1, and Abs_dpI, de-

scribe general physicochemical differences independent of residues' structural environments. In contrast, CDSS provides a more specific residue substitution score in given RSA levels and secondary structure states [51]. In Fig. 3, the CDSS values of daAASs are smaller than those of nAASs, indicating that unfavorable substitutions with respect to CDSS tend to be associated with diseases.

Environment-dependent residue contact energy (ERCE) is another selected structural environment-related feature, accounting for both the secondary structure and residue contacts in the 3D structure [52]. The selected ERCE feature for an AAS describes the summation of contacting energy values between the wild-type residue and all its neighboring residues within a distance of 6.5 Å. The lower the contact energy, the more stability it contributes. In Fig. 3, the ERCE values of daAASs are smaller than those of nAASs, suggesting that the residues with higher stability contribution, if substituted, tend to become pathogenic.

To our knowledge, it is the first time that CDSS and ERCE have been utilized in developing methods for predicting the disease-association of AAS. The effectiveness of CDSS and ERCE indicates an understanding that substitutions introducing incompatible residues into a specific structural context will tend to be deleterious to the protein and thus be pathogenic.

4.5 Structural features contribute to the improved prediction performance

In addition to better interpretability of the structural features, to what extent they contribute to the prediction performance is another aspect that must be interrogated. Toward this end, we re-trained an additional predictor with all seven structural features removed according to the same procedure adopted in training AAS3D-RF, namely, the structure-removed predictor. Then, we compared its performance with AAS3D-RF on *TestDataset 1* and *TestDataset 2* (Table 2 and **Supplementary Table 3**).

First, the ACC and MCC of structure-removed predictor were 2.8 and 4.9 percentage points less than the AAS3D-RF on the *TestDataset 1*, respectively (Table 2 and **Supplementary Table 3**), demonstrating that the structure features evidently contributed to the performance.

Second, when evaluated on *TestDataset 2*, the observed improvement of ACC and MCC after adding structure-related features was not as large as that observed on *TestDataset 1* (ΔACC = 0.004 from 0.835 to 0.839, ΔMCC = 0.008 from 0.676 to 0.684) (Table 2 and **Supplementary Table 3**).

What factors have affected the extent of performance improvement of structural features? As several previous studies have proposed that the contribution of structural features is more evident when reliable conservation-related features are unavailable [53, 54], we checked whether this applies in our case. On *TestDataset 2*, by using the number of aligned sequences (the feature of nal_1e-45) as a proxy of reliability of conservation-related features, we observed that the improvement of MCC was indeed more evident in AAS data with nal_1e-45 $<$200 than those $\geq$200 (ΔMCC = 0.147 from 0.645 to 0.792 *vs*. ΔMCC = 0.007 from 0.676 to 0.683) (**Supplementary Table 5**). Similar results were obtained on *TestDataset 1* (0.244 *vs*. 0.041) (**Supplementary Table 5**). The importance of structural features in predicting pathogenic AASs was also frequently emphasized in previous studies, since they can improve the prediction of nAASs in conserved/constrained regions and daAASs in regions with loose constraints, in addition to offering hints of pathogenic mechanisms [55–57].

In summary, incorporation of structural features further improves prediction performance, especially in scenarios wherein conservation-related features are of low reliability.

5. Discussion

Developing tools for predicting daAASs has been a challenge for over a decade. Although a plethora of studies have devoted significant effort and achieved much progress in this field, their performance requires further improvement [37, 58]. As more and more AASs with phenotypic effects are determined, available methods can be re-evaluated and previous machine learning-based tools could be upgraded with new data. With other related data accumulated, such as homologous sequences and structures in public databases, novel predictors could be developed with more accurate feature descriptors, more complete feature space, or brand-new features. Given this background, our work has been carried out to explore new structural features and to combine them with sequence features aiming at improving the prediction and understanding of disease-associated AASs.

Many studies have adopted sequence features and the predicted structural features based on sequence, but only a few directly extract features from protein 3D structures, whether experimental structures or homology models [45, 57, 59, 60]. According to the paradigm of sequence-structure-function, protein structures are more directly related to function than sequence, so the 3D structures should be able to provide greater understanding of pathogenic AASs. However, the structure feature extraction procedure is much more complicated, and only a small part of AASs are covered by structures, which may have restricted the extensive exploration of structural features. Several recent studies have integrated or explored AASs in the structural contexts with much larger datasets. By adopting both experimental and homology-modeled structures, PhyreRisk and Missense3D have increased the number of AASs that can be mapped to 3D structures and can calculate the potential structural impacts based on knowledge-based rules [17, 18]. Similarly, the mutfunc tool has

also incorporated homology-modeled structures, and has provided pre-calculated properties of stability, interaction, post-translational modification, linear motif, and transcription factor binding site [21] for datasets curated from ExAC [20] and ClinVar [5]. By focusing only on experimental structures, VarSite has offered a graphical platform to inspect AASs in the contexts of conservation levels, protein domains, secondary structures, and interaction sites [19]. Through analyzing 40 properties associated with the variant position, MISCAST has identified properties significantly associated with daAASs or nAASs [22]. More specifically, it has also identified significant properties within each of the 24 protein functional groups separately. Accordingly, MISCAST has defined P3DFi scores for each residue position based on either the joint analyses of all protein classes or the separate analyses of each functional class, and these scores can improve the prediction of pathogenic AASs. These studies have largely advanced the interpretation of AASs in the contexts of 3D structures. In this work, we have constructed a more comprehensive set of 212 candidate features, and have adopted an automatic feature selection pipeline considering the redundancy between features [30]. Unlike previous studies that have chosen interpretable features based on knowledge or statistical analyses, our feature selection is more intended to maximize prediction performance and to ignore redundant features.

The main new structural features in this work include ERCE and CDSS. ERCE relies on the secondary structure and structural interacting residues, while CDSS is dependent on the secondary structure and RSA (Fig. 3). The analysis of ERCE and CDSS herein hints that AASs incorporating residues incompatible to its structural environment may be potential reasons for certain daAASs. Another two selected structure features, RSA and $\Delta\Delta G$, are related to stability and are widely recognized in previous studies [11, 41, 44–48]. Our work has also shown that they are highly beneficial according to the *p*-values (Fig. 3). The recent work of MISCAST has highlighted the significance of RSA in its feature group of residue exposure levels as well [22]. The other three selected structural features, dF1_Mean_Mutant, dVol_Mean_Mutant, and varHP_Wild_NB, reflect the physicochemical properties of AAS residues and its structural neighboring residues, and have not previously been explored for predicting pathogenic AASs (**Supplementary Table 2**). Although they are often largely masked by conservation-related features, the performance improvements from structure-related features are evident when reliable conservation-related features are unavailable (**Supplementary Tables 3,5**).

As for physicochemical properties irrelevant to structures, MISCAST has emphasized the properties of the wild-type residues, while AAS3D-RF has mainly focused on the difference between wild-type and mutant residues (Fig. 3). Disulfide bond-related features have also been highlighted in both MISCAST and AAS3D-RF but are slightly different: The former represents the structural neighboring disulfide bonds, while the latter represents the sequence neighboring ones. The post-translational features have been absent in our work, but most of them have been demonstrated as significant in the work of MISCAST. Hence, they should be considered and utilized in future studies of developing predictors. Notably, we have extracted many more features as the candidates (212 in total), including functional regions (calcium binding, DNA-binding, nucleotide phosphate, membrane-spanning, and zinc finger regions), sequence and structural neighboring functional sites (active pocket, chemical group binding, metal ion binding, and disulfide bonded cysteine sites), hydrogen bonds, and the 8-state secondary structures. However, our automatic feature selection pipeline has not incorporated most of them into the final feature subset, possibly because they could not contribute further to maximize the prediction performance. Nevertheless, many of these dropped features are informative in providing clues to understand the molecular basis of pathogenic AASs, as demonstrated in the work of MISCAST, VarSite, mutfunc, PhyreRisk and Missense3D [17–19, 21, 22]. Hence, one may consider strategies to combine manual retaining of certain biologically meaningful features with automatic feature selection together aiming at providing better interpretability in the future.

When developing machine learning-based tools for predicting daAASs, type 2 circularity is an important issue that is worth noting. In our work, Residual Variation Intolerance Score (RVIS), a gene-level metric measuring the relative ability of a human gene tolerating common functional genetic variation in healthy individuals [61], has been chosen during the feature selection. The RVIS scores of AASs from the same gene/protein are identical. In the cross-validation, if AASs from the same gene/protein are separated into the training and the validation data, their identical gene/protein-level feature value will lead to overfitting. To avoid this, we conducted cross-validation at the protein-level, i.e., AASs were first grouped according to their source proteins, and then the training and validation data separation was carried out without splitting any group. The performance evaluation on the ‘Mixed’ datasets has demonstrated that this strategy is effective. In the future, if training data accumulate sufficiently, one can train better predictors based on only ‘Mixed’ datasets with a balanced number of daAASs and nAASs in each protein.

Currently, most predictors and AAS annotators, including AAS3D-RF, are developed for human proteins, and only few can be applied to other organisms. The underlying reasons may be the lack of training/testing samples or proper features suitable for non-human species. For those that can be applied to other organisms, they mainly adopted universal features that are not restricted within specific organisms, such as conservation scores and structural stabil-

ity scores. Such examples include mutfunc [21], Fido-SNP [62], and Envision [63]. As for AAS3D-RF, though it cannot be applied to other organisms due to the human-specific RVIS feature currently, its CDSS and ERCE features will be promising when extrapolating to other organisms.

For this study, we exclusively adopted homology models from ModBase and AASs from humsavar, 1000G, and VariSNP datasets. With the unprecedented progress in protein structure prediction such as the recent AlphaFold, 58% of human proteome residues now have been confidently mapped with 3D conformations [64]. Moreover, ClinVar contains more annotated variants than humsavar and is growing rapidly as well. A study showed that only 8% of ClinVar daAASs and 32% of humsavar daAASs overlap [65]. In addition to public data, the commercial HGMD database holds even more variants with disease annotations [7]. Integrating these larger datasets will be beneficial for improving AAS3D-RF and other related predictors in the future.

6. Conclusions

In this work, we curated a training dataset containing about 15 thousand AASs with known phenotypic effects and mappable to reliable 3D structures. Based on 21 automatically selected sequence and structural features, the RF-based machine-learning model AAS3D-RF was trained using G10F-CV. Evaluation on several independent testing datasets showed that AAS3D-RF achieved ACC of 0.811∼0.839 and MCC of 0.591∼0.684, outperforming seven other tools. Moreover, its unique structure-based features including CDSS and ERCE can offer mechanistic clues for the predicted daAASs. In summary, AAS3D-RF serves as a valuable tool for both predicting and understanding pathogenic AASs.

7. Author contributions

YX, ZQY, and YDW conceived and designed this study; YX, JBZ, and KA performed the calculations; YX, ZQY, YDW, WH, and TW analyzed the data; YX and ZQY drafted the manuscript; all the authors revised and approved the manuscript.

8. Ethics approval and consent to participate

This is a computational study and no subjects of humans or animals were involved in.

9. Acknowledgment

We would like to thank Xinhao Zhang, Fan Jiang, and Xudong Zou for their helpful suggestions. We also thank the reviewers' valuable comments for improving the manuscript. We gratefully acknowledge the support of Shenzhen Bay Laboratory Supercomputing Center and the National Supercomputer Center in Guangzhou (NSCC-GZ) for computing resources.

10. Funding

This work was supported by the Key-Area Research and Development Program of Guangdong Province [2020B0101350001]; the National Natural Science Foundation of China [32070664, 21933004, 31471243]; the Shenzhen Science and Technology Innovation Commission [JCYJ20170818085409785, GXWD20201231165807007-20200812124825001]; and the Shenzhen Municipal Health Commission [SZSM201809085].

11. Conflict of interest

The authors declare no conflict of interest.

12. References

[1] Shendure J, Balasubramanian S, Church GM, Gilbert W, Rogers J, Schloss JA, *et al*. DNA sequencing at 40: past, present and future. Nature. 2017; 550: 345–353.

[2] Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, *et al*. DbSNP: the NCBI database of genetic variation. Nucleic Acids Research. 2001; 29: 308–311.

[3] Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nature Genetics. 2003; 33: 228–237.

[4] Stenson PD, Mort M, Ball EV, Shaw K, Phillips AD, Cooper DN. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Human Genetics. 2014; 133: 1–9.

[5] Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, *et al*. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Research. 2018; 46: D1062–D1067.

[6] UniProt Consortium T. UniProt: the universal protein knowledgebase. Nucleic Acids Research. 2018; 46: 2699.

[7] Stenson PD, Mort M, Ball EV, Evans K, Hayden M, Heywood S, *et al*. The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. Human Genetics. 2017; 136: 665–677.

[8] Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature Protocols. 2009; 4: 1073–1082.

[9] Ramensky V, Bork P, Sunyaev S. Human non-synonymous SNPs: server and survey. Nucleic Acids Research. 2002; 30: 3894–3900.

[10] López-Ferrando V, Gazzo A, de la Cruz X, Orozco M, Gelpí JL. PMut: a web-based tool for the annotation of pathological variants on proteins, 2017 update. Nucleic Acids Research. 2019; 45: W222–W228.

[11] Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, *et al*. A method and server for predicting damaging missense mutations. Nature Methods. 2010; 7: 248–249.

[12] Pejaver V, Urresti J, Lugo-Martinez J, Pagel KA, Lin GN, Nam

HJ, *et al*. Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. Nature Communications. 2020; 11: 5918.
[13] Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. PloS ONE. 2012; 7: e46688.
[14] Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GLA, Edwards KJ, *et al*. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Human Mutation. 2013; 34: 57–65.
[15] Tang HM, Thomas PD. PANTHER-PSEP: predicting disease-causing genetic variants using position-specific evolutionary preservation. Bioinformatics. 2016; 32: 2230–2232.
[16] Rose PW, Prlić A, Altunkaya A, Bi C, Bradley AR, Christie CH, *et al*. The RCSB protein data bank: integrative view of protein, gene and 3D structural information. Nucleic Acids Research. 2017; 45: D271–D281.
[17] Ofoegbu TC, David A, Kelley LA, Mezulis S, Islam SA, Mersmann SF, *et al*. PhyreRisk: a Dynamic Web Application to Bridge Genomics, Proteomics and 3D Structural Data to Guide Interpretation of Human Genetic Variants. Journal of Molecular Biology. 2019; 431: 2460–2466.
[18] Ittisoponpisan S, Islam SA, Khanna T, Alhuzimi E, David A, Sternberg MJE. Can Predicted Protein 3D Structures Provide Reliable Insights into whether Missense Variants are Disease Associated? Journal of Molecular Biology. 2019; 431: 2197–2212.
[19] Laskowski RA, Stephenson JD, Sillitoe I, Orengo CA, Thornton JM. VarSite: Disease variants and protein structure. Protein Science. 2020; 29: 111–119.
[20] Karczewski KJ, Weisburd B, Thomas B, Solomonson M, Ruderfer DM, Kavanagh D, *et al*. The ExAC browser: displaying reference data information from over 60,000 exomes. Nucleic Acids Research. 2017; 45: D840–D845.
[21] Wagih O, Galardini M, Busby BP, Memon D, Typas A, Beltrao P. A resource of variant effect predictions of single nucleotide variants in model organisms. Molecular Systems Biology. 2018; 14:e8430
[22] Iqbal S, Pérez-Palma E, Jespersen JB, May P, Hoksza D, Heyne HO, *et al*. Comprehensive characterization of amino acid positions in protein structures reveals molecular effect of missense variants. Proceedings of the National Academy of Sciences. 2020; 117: 28201–28211.
[23] Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Research. 2019; 47: D886–D894.
[24] Breiman L. Random forests. Machine Learning. 2001; 45: 5–32.
[25] Pieper U, Webb BM, Dong GQ, Schneidman-Duhovny D, Fan H, Kim SJ, *et al*. ModBase, a database of annotated comparative protein structure models and associated resources. Nucleic Acids Research. 2014; 42: D336–D346.
[26] Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, *et al*. A global reference for human genetic variation. Nature. 2015; 526: 68–74.
[27] Schaafsma GCP, Vihinen M. VariSNP, a Benchmark Database for Variations from dbSNP. Human Mutation. 2015; 36: 161–166.
[28] Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, *et al*. BLAST+: architecture and applications. BMC Bioinformatics. 2009; 10: 421.
[29] Touw WG, Bayjanov JR, Overmars L, Backus L, Boekhorst J, Wels M, *et al*. Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle? Briefings in Bioinformatics. 2013; 14: 315–326.
[30] Zhou JB, Xiong Y, An K, Ye ZQ, Wu YD. IDRMutPred: predicting disease-associated germline nonsynonymous single nucleotide variants (nsSNVs) in intrinsically disordered regions. Bioinformatics. 2020; 36: 4977–4983.
[31] Grimm DG, Azencott CA, Aicheler F, Gieraths U, MacArthur DG, Samocha KE, *et al*. The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. Human Mutation. 2015; 36: 513–523.
[32] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, *et al*. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011; 12: 2825–2830.
[33] Wei Q, Dunbrack RL, Jr. The role of balanced training and testing data sets for binary classifiers in bioinformatics. PLoS ONE. 2013; 8: e67863.
[34] Dobson RJ, Munroe PB, Caulfield MJ, Saqi MA. Predicting deleterious nsSNPs: an analysis of sequence and structural attributes. BMC Bioinformatics. 2006; 7: 217.
[35] Wang M, Wei L. IFish: predicting the pathogenicity of human nonsynonymous variants using gene-specific/family-specific attributes and classifiers. Scientific Reports. 2016; 6: 31321.
[36] Ghosh R, Oak N, Plon SE. Evaluation of in silico algorithms for use with ACMG/AMP clinical variant interpretation guidelines. Genome Biology. 2017; 18: 225.
[37] Riera C, Lois S, de la Cruz X. Prediction of pathological mutations in proteins: the challenge of integrating sequence conservation and structure stability principles. Wiley Interdisciplinary Reviews: Computational Molecular Science. 2014; 4: 249–268.
[38] Peterson TA, Doughty E, Kann MG. Towards Precision Medicine: Advances in Computational Approaches for the Analysis of Human Variants. Journal of Molecular Biology. 2013; 425: 4047–4063.
[39] Katsonis P, Koire A, Wilson SJ, Hsu TK, Lua RC, Wilkins AD, *et al*. Single nucleotide variations: Biological impact and theoretical interpretation. Protein Science. 2014; 23: 1650–1666.
[40] Niroula A, Vihinen M. Variation Interpretation Predictors: Principles, Types, Performance, and Choice. Human Mutation. 2016; 37: 579–597.
[41] Ye ZQ, Zhao SQ, Gao G, Liu XQ, Langlois RE, Lu H, *et al*. Finding new structural and sequence attributes to predict possible disease association of single amino acid polymorphism (SAP). Bioinformatics. 2007; 23: 1444–1450.
[42] Steward RE, MacArthur MW, Laskowski RA, Thornton JM. Molecular basis of inherited diseases: a structural perspective. Trends in Genetics. 2003; 19: 505–513.
[43] de Beer TAP, Laskowski RA, Parks SL, Sipos B, Goldman N, Thornton JM. Amino acid changes in disease-associated variants differ radically from variants observed in the 1000 genomes project dataset. PLoS Computational Biology. 2013; 9: e1003382.
[44] Karchin R, Diekhans M, Kelly L, Thomas DJ, Pieper U, Eswar N, *et al*. LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. Bioinformatics. 2005; 21: 2814–2820.
[45] Baugh EH, Simmons-Edler R, Müller CL, Alford RF, Volfovsky N, Lash AE, *et al*. Robust classification of protein variation using structural modelling and large-scale data integration. Nucleic Acids Research. 2016; 44: 2501–2513.
[46] Capriotti E, Altman RB. Improving the prediction of disease-related variants using protein three-dimensional structure. BMC Bioinformatics. 2011; 12: S3.
[47] Yang X, Gao H, Zhang J, Xu X, Liu X, Wu X, *et al*. *ATP1A3* mutations and genotype-phenotype correlation of alternating hemiplegia of childhood in Chinese patients. PLoS ONE. 2014; 9: e97274.
[48] Riera C, Lois S, Domínguez C, Fernandez-Cadenas I, Montaner J, Rodríguez-Sureda V, *et al*. Molecular damage in Fabry disease: characterization and prediction of alpha-galactosidase a pathological mutations. Proteins. 2015; 83: 91–104.
[49] Yue P, Li ZL Moult J. Loss of protein structure stability as a major causative factor in monogenic disease. Journal of Molecular Biology. 2005; 353: 459–473.
[50] Wang Z, Moult J. SNPs, protein structure, and disease. Human Mutation. 2001; 17: 263–270.
[51] Koshi JM, Goldstein RA. Context-dependent optimal substitu-

tion matrices. Protein Engineering. 1995; 8: 641–645.
[52] Zhang C, Kim SH. Environment-dependent residue contact energies for proteins. Proceedings of the National Academy of Sciences of the United States of America. 2000; 97: 2550–2555.
[53] Saunders CT, Baker D. Evaluation of structural and evolutionary contributions to deleterious mutation prediction. Journal of Molecular Biology. 2002; 322: 891–901.
[54] Bao L, Cui Y. Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information. Bioinformatics. 2005; 21: 2185–2190.
[55] Zhang J, Kinch LN, Cong Q, Katsonis P, Lichtarge O, Savojardo C, *et al*. Assessing predictions on fitness effects of missense variants in calmodulin. Human Mutation. 2019; 40: 1463–1473.
[56] Glusman G, Rose PW, Prlić A, Dougherty J, Duarte JM, Hoffman AS, *et al*. Mapping genetic variations to three-dimensional protein structures to enhance variant interpretation: a proposed framework. Genome Medicine. 2017; 9: 113.
[57] Quan L, Wu H, Lyu Q, Zhang Y. DAMpred: Recognizing Disease-Associated nsSNPs through Bayes-Guided Neural-Network Model Built on Low-Resolution Structure Prediction of Proteins and Protein–Protein Interactions. Journal of Molecular Biology. 2019; 431: 2449–2459.
[58] Cooper GM, Shendure J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nature Reviews. Genetics. 2011; 12: 628–640.
[59] 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: 14.
[60] 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: e43847.
[61] Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. Genic intolerance to functional variation and the interpretation of personal genomes. PLoS Genetics. 2013; 9: e1003709.
[62] Capriotti E, Montanucci L, Profiti G, Rossi I, Giannuzzi D, Aresu L, *et al*. Fido-SNP: the first webserver for scoring the impact of single nucleotide variants in the dog genome. Nucleic Acids Research. 2019; 47: W136–W141.
[63] Gray VE, Hause RJ, Luebeck J, Shendure J, Fowler DM. Quantitative Missense Variant Effect Prediction Using Large-Scale Mutagenesis Data. Cell Systems. 2018; 6: 116–124 e113.
[64] Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Žídek A, *et al*. Highly accurate protein structure prediction for the human proteome. Nature. 2021; 596: 590–596.
[65] McGarvey PB, Nightingale A, Luo J, Huang H, Martin MJ, Wu C, *et al*. UniProt genomic mapping for deciphering functional effects of missense variants. Human Mutation. 2019; 40: 694–705.

Supplementary material: Supplementary material associated with this article can be found, in the online version, at https://www.imrpress.com/journal/FBL/26/12/10.52586/5036.

Appendix

During the peer-review of this manuscript, the reviewers suggested further evaluation of AAS3D-RF on new AASs and on previously unmapped AASs that can be mapped now due to the increased coverage of 3D structures since April, 2018. We undertook several steps to fulfil this. First, we compared the human protein entries' annotations in UniProtKB/Swiss-Prot (2021, 03) and UniProtKB/Swiss-Prot (2018, 04), and obtained 943 new experimental structures of 907 proteins. The quality-check and the removal of those that are overlapped with the ModBase structures mentioned in section 3.1 resulted in 631 new structures of 606 proteins. Second, we mapped the humsavar (2021, 03) AASs to all the structures including these newly curated 631 structures, previously curated experimental structures, and previously curated homology-modelled structures (**Supplementary Fig. 1**), and then removed those that have occurred in *TrainDataset*, *TestDataset 1*, or *TestDataset 2*, obtaining a new testing dataset, namely *TestDataset 3*, which contains 2,614 new AASs (1,675 daASs and 939 nAASs). The AASs in *TestDataset 3* stem from 616 structures of 582 proteins (**Supplementary Table 6**).

Third, we ran AAS3D-RF and the other seven tools against *TestDataset 3*. The settings for running these seven tools were the same as those used on *TestDataset 1* and *TestDataset 2*, and are described in **Supplementary Methods**. Lastly, we evaluated the performance of AAS3D-RF on *TestDataset 3*. The ACC and MCC are respectively 0.821 and 0.619 similar to those based on *TestDataset 1* and *TestDataset 2* (**Supplementary Tables 4,7**). In addition, AAS3D-RF has also evidently outperformed other tools with respect to ACC and MCC (**Supplementary Fig. 4A**). The ROC plot shows that AAS3D-RF covers a much larger area (0.888) than others as well (**Supplementary Fig. 4B**). Details of *TestDataset 3* are also provided at http://www.wdspdb.com/AAS3D-RF/. In summary, AAS3D-RF has stably superior performance to many other tools by combining new interpretable structural features and sequence features.

Abbreviations: AAS, Amino acid substitution; SNV, Single-nucleotide variant; VUS, Variants of uncertain significance; RF, Random forest; G10F-CV, Group10Fold cross-validation; AUC, Area under the ROC curve; ACC, Accuracy; MCC, Matthews correlation coefficient; PPV, Positive predictive value; NPV, Negative predictive value; CDSS, Context-dependent substitution score; ERCE, Environment-dependent residue contact energy; RVIS, Residual variation intolerance score.

Keywords: Amino acid substitution; Single-nucleotide variant; Pathogenic; Protein structure; Machine learning

Send correspondence to:

Zhi-Qiang Ye, State Key Laboratory of Chemical Oncogenomics, Peking University Shenzhen Graduate School, 518055 Shenzhen, Guangdong, China, Shenzhen Bay Laboratory, 518055 Shenzhen, Guangdong, China, E-mail: yezq@pku.org.cn

Yun-Dong Wu, State Key Laboratory of Chemical Oncogenomics, Peking University Shenzhen Graduate School, 518055 Shenzhen, Guangdong, China, Shenzhen Bay Laboratory, 518055 Shenzhen, Guangdong, China, College of Chemistry and Molecular Engineering, Peking University, 100871 Beijing, China, E-mail: ydwu@pku.edu.cn