Supplementary Methods

1. 3D Structure acquiring and data cleaning

The overall pipeline is shown in **Fig. S1**. In detail, we retrieved all human canonical protein sequences and their annotations from uniprot_sprot.fasta and uniprot_sprot.dat downloaded from the UniProt FTP server (Release 2018_04). Among the 20,340 protein entries (TITIN (Q8WZ42) was excluded because it is too long to extract related features in a reasonable time limit), 6,328 have at least one PDB item found in the "DR" fields according to their UniProt annotations. Since one protein may have several overlapping PDB structures, we aimed at selecting the non-overlapping PDB structures with the largest sequence coverage and best resolution. We first parsed out all the PDB items with X-ray crystallography diffraction, NMR and EM methods according to the annotations in the "DR" fields. For each protein, we split these PDB items into two lists based on whether the experimental resolution is given. For the list with resolution information available, we only retained the PDB items with (1) resolution equal to or less than 3.0 Å, (2) length longer than 50 or spanning over 50% of the full length to ensure high quality and sufficient size. After sorting it by structure length and resolution, we obtained a list, which was used to select the non-overlapping PDB items (LIST1) with higher priority for longer structure length and better resolution. Considering that certain structures, *e.g.*, NMR structures, are not accompanied by resolution information but also very informative, a part of them were selected to supplement LIST1 according to this procedure: Specifically, we retained the PDB items with length longer than 50 or spanning more than 50% of the full-length sequence, and then selected a list of non-overlapping PDB items with higher priority for longer structure length, namely LIST2. For each item in LIST2, if it had no overlap with any item in LIST1, we added it to LIST1. These steps led to the final LIST1 with 6,529 PDB items for 5,371 proteins. The information used above for data filtering such as structure determination method, resolution, protein region that a PDB item covers was obtained from the "DR" fields in uniprot_sprot.dat.

It is worth noting that these PDB entries may be protein complexes, or may contain expression tags, cloning artifacts or insertion codes in PDB chains, which would thus hinder the subsequent batch mining of the structural features. To smooth the process of structural analyses, we technically adopted a self-modeling strategy to obtain the coordinates only for protein regions that the PDB items cover. In detail, from the full-length UniProt reference sequences, we first obtained the sequence fragments that matched with those PDB items in LIST1, and then we modelled the structures of these sequence fragments using Modeller v9.19 by specifying the corresponding chosen PDB items as templates (1). In the process of running Modeller, the sequence fragment and chosen PDB template were aligned using local alignment and id.sim.mat similarity scoring matrix. Once the alignment was constructed, we generated the model using Modeller's automodel class and refined

the model using the predefined function refine.very_slow in order to get an energy optimized conformation.

Eighty-eight out of 6,529 chosen PDB items do not have models because of missing templates in PDB format (n = 82) and residue type "U" in the target that Modeller could not handle (n = 6). For the remaining 6,441 structures, we removed one containing non-standard residue "ASX" and ten whose sequence identities between the aligned target and template are below 50%, which might be due to some annotation errors in the UniProtKB/Swiss-Prot "DR" fields. Finally, there are 6,430 experimental structures for 5,278 proteins.

As for the 14,012 proteins without PDB items found in the "DR" fields, we obtained their comparative structure models in xml format from ModBase (2). Except that 485 protein entries' xml files lack model information, 342,868 models of 13,527 proteins were downloaded on September 8, 2018. According to its documentation, a model is considered reliable and has an acceptable fold assignment if its model score is equal to or greater than 0.7 and the E-value is equal to or below 0.0001. The models were filtered based on the following six criteria: (i) model score $\geqslant$ 0.7; (ii) E-value $\leqslant$ 0.0001; (iii) sequence identity $\geqslant$ 50%; (iv) model length > 50 or model spans > 50% of the protein sequence; (v) containing only standard residue types; (vi) the sequence parsed from the model's ATOM section exactly matches the corresponding UniProt sequence region. This filtering resulted in a list of 26,111 structure model candidates. For each protein that has structure models in the candidate list, we sorted the models by model length and sequence identity, and then selected the non-overlapped structure models with higher priority for longer model length and larger sequence identity. As a result, a total of 4,238 comparative homology models covering 3,682 proteins were obtained.

To resolve the residue numbering inconsistencies between the UniProt sequences and ModBase structure models, we aligned the sequences parsed out from the ATOM sections in structure files with the corresponding UniProt sequences using Biopython's pairwise2 module (3). According to the alignments, we acquired the corresponding residue position shifts and reassigned the residue numbers for all chosen ModBase structure models.

These curated experimental and modelled structure files in PDB format and their information are available at http://www.wdspdb.com/AAS3D-RF/ .

2. Candidate feature calculation

All of the candidate feature descriptions are available at http://www.wdspdb.com/AAS3D-RF/ .

2.1 PSI-BLAST-based sequence features

We ran PSI-BLAST (4) locally against UniRef90 database (Release 2018_06) using ncbi-blast-2.2.29+ with three iterations and an E-value of 0.0001 (5) to find

homologous sequences. And then according to PyMut (6), the alignment was filtered in four different ways to generate different final alignments: (i) taking all the sequences, (ii) retaining only the human sequences, (iii) excluding all the human sequences, and (iv) taking matches under a stricter E-value threshold of 10^{-45}. Next, for each of the four alignments above, features were computed, including: (i) number of sequences in the alignment, (ii) number of amino acids in the aligned position, (iii) total and relative number of aligned wild-type amino acids, (iv) total and relative number of aligned mutant amino acids, (v) position weight matrix score, both in a weighted (using the BLAST bit-score) and unweighted fashion, resulting in a total of 56 features.

Shannon's entropy measures the randomness of residues at a specific column in a multiple sequence alignment, reflecting the conservation level of the site. The entropy was calculated as a feature using in-house scripts based on the alignment of all the PSI-BLAST hit sequences described above:

$$\text{Shannon}' \text{s entropy} = -\sum_{i=1}^{20} p_i \, log_{20} p_i \ ,$$

where p_i represents the frequency of amino acid i at the AAS position in the alignment. Another entropy feature was directly derived from the results of RING (version 2.0) (7, 8) after running PSI-BLAST (4) for each protein sequence segment parsed from the 3D structure. The same parameter settings as above were adopted to get PSI-BLAST alignments in RING.

We also extracted six features directly from the PSSM (Position-Specific Scoring Matrix) profiles (9), generated in the PSI-BLAST runs. These included the PSSM scores of wild-type and mutant amino acids along with their difference, the weighted percentages of wild-type and mutant amino acids, and also the position's information content.

2.2 Gene-level features

These features originally aim to prioritize genes that are plausible candidates for inherited diseases, and will thus provide information for further prioritizing disease variants. Here, we integrated Gene Damage Index (GDI) (10) and Residual Variation Intolerance Score (RVIS) (11).

GDI is a genome-wide, population-based metric to describe the cumulative mutational damage for each human protein-coding gene. A gene with a low GDI is less damaged and tends to be under stronger purifying selection. Therefore, variants in those genes are more likely to be associated with diseases (10). RVIS ranks human genes according to their ability to tolerate common functional genetic variation in healthy individuals. Genes with significantly less common functional variations than expected tend to have higher constraint and lower RVIS. That is, the lower the RVIS, the more likely the gene causes certain kinds of disease (11). Here, we used the RVIS gene score downloaded from http://genic-intolerance.org, which is based on ExAC v2 (release 2.0).

2.3 Substitution matrix scores

For each AAS, the scores in BLOSUM62 (12), Grantham (13), and Miyata (14) matrices were used as features. BLOUSM62 describes the substitution bias from evolutionary perspective, and is widely used as default in many alignment algorithms. Grantham and Miyata matrices are used to describe physicochemical dissimilarities between amino acid, with Grantham's distance based on polarity, molecular volume, and composition, and Miyata's distance based only on polarity and molecular volume.

2.4 Differences of physicochemical properties

Several pieces of work have proposed that AASs with relatively small physicochemical changes occur more frequently than those with large changes (13, 15-17). Disease-associated AASs typically have more drastic changes than neutral AASs. Here, we considered three biologically meaningful properties: (i) Grantham molecular volume (13), (ii) Kyte-Doolittle hydropathy index (18), and (iii) isoelectric point (19). Another five highly interpretable numerical factors were considered, which reflect polarity, secondary structure, molecular size and volume, amino acid composition and codon usage, and electrostatic charge, derived from factor analysis of 494 amino acid attributes (20). We calculated the difference and absolute difference between the mutant and wild-type amino acid for each of these eight properties.

2.5 The composition of nearby sequence

We used a vector with 20 components to encode the composition of 20 residue types for a sequence window of 19, which is centered at the AAS position and spans nine residues either to the N-terminus or to the C-terminus (21).

2.6 Functional regions of interest

We checked whether the AAS is located in the functional region described in the "FT" fields of uniprot_sprot.dat. We considered CA_BIND (calcium binding region), DNA_BIND (DNA-binding domain), NP_BIND (nucleotide phosphate binding region), TRANSMEM (membrane-spanning region) and ZN_FING (zinc fingers within the protein), and used five feature columns to denote them. When an AAS falls into a region above, the corresponding column was set to "1" and the others were set to "0". Another additional column was used to indicate whether an AAS falls into any of the regions above.

2.7 Sequence distance of nearby functional sites

An AAS is presumably more likely to be disease-associated when it is close to a

functional site. We considered the functional sites including ACT_SITE (enzyme's active sites), BINDING (binding sites for any chemical group), DISULFID (cysteine residues participating in disulfide bonds), and METAL (binding sites for a metal ion), which are annotated in the "FT" fields of UniProtKB/Swiss-Prot database. For each type of functional sites, we adopted the sequence distance between an AAS position and its closest functional site (22). Another column was used to represent the closest sequence distance between the AAS site and all four kinds of functional sites.

2.8 Structural distance of nearby functional sites

Similar to sequence distance, for each type of functional site (ACT_SITE, BINDING, DISULFID, METAL), we calculated the structural distance between an AAS position and its closest functional site. The spatial distance between two residues' C_β atoms (for glycine, C_α was used) was regarded as the structural distance (22). Another value was used to represent the closest structural distance between the AAS position and all four kinds of functional sites.

2.9 The composition of neighbor residues

Similar to the composition of nearby sequence, a vector of 20 components was used to encode the composition of 20 residue types for the structural contacting neighbors (or partners) whose C_α atoms are within a distance of 6.5 Å to the C_α atom of the wild-type amino acid of the AAS (23, 24).

2.10 Secondary structure states

The secondary structural information for each structure was derived from DSSP (version 2.2.1) (25, 26). Eight states (H, I, G, S, T, B, C, E) were denoted by eight feature columns. For a specified site, we assigned "1" to the column corresponding to the secondary structural state of this site, and "0" was assigned to all other columns.

2.11 Residue interaction network topological measures

The 3D structure of a protein can be coarsely represented as a residue-residue interaction network (or residue contact network), where vertices represent residues and edges represent contacts between residues (27). In such a network, four network topological features were calculated for each AAS by using the Python module NetworkX (28), including two measures of local interaction (degree and clustering coefficient) and two measures of global interaction (closeness and betweenness). We utilized the RING software (version 2.0) (7, 8) to construct the residue-residue interaction network, and the interaction cutoff of two C_α atoms was set to 6.5 Å. Another two features were derived to describe the topological importance of the contacting partners for an AAS position. One is NB_Max_Degree, which represents the maximum degree of the contacting partners, and the other is NB_Mean_Degree,

which represents the mean degree of all the contacting partners.

2.12 Physicochemical properties of the structural contacting neighbors

As mentioned in section 2.4, for each of the 20 amino acids, there are three biologically meaningful properties and five highly interpretable numerical factors (13, 18-20). Based on them, we designed five types of features to describe the physicochemical properties of the structural contacting neighbors of an AAS. NB_MeanX (X denotes one of the eight physicochemical properties or factors) is the average of the property X of all the contacting neighbors of an AAS, dX_Mean_Mutant is the difference between the NB_MeanX and the mutant amino acid, varX_Wild_NB is the variance of property X for the wild-type amino acid and its contacting neighbors, varX_Mutant_NB is the variance value of the property X for the mutant amino acid and its contacting neighbors, and dVarX is the difference between the varX_Mutant_NB and varX_Wild_NB. These calculations resulted in 40 features in total.

2.13 Contact energy

The contact energy can reflect the tendency that two residues contact with each other. For wild-type residue at the substitution position, the structural contacting neighbors were defined as described in section 2.9. Then we summed the contact energy between the wild-type residue and all its contacting neighbors, and the absolute value of this sum was also used as a feature. The contact energy between two residue types was retrieved from the matrices of MIYS990106 (hereinafter referred to as MJ1) and MIYS990107 (hereinafter referred to as MJ2) from the AAindex database (29, 30), respectively, resulting in four features. We further obtained the contact energy between the mutant residue and all the contacting neighbors of the wild-type residue, by simply querying each score between the mutant residue type and all the contacting neighbors from the matrices without structure model building of the mutant. Then the difference and absolute difference between mutant and wild-type amino acid were calculated as well, resulting in four additional features.

The contact energy can also be defined by using the Environment-dependent Residue Contact Energy (ERCE) matrices, which provide six matrices according to different secondary structure states of the two contacting residues, including Helix_Helix, Helix_Strand, Helix_Coil, Strand_Strand, Strand_Coil, and Coil_Coil (31). Eight secondary structural states, including H, I, G, S, T, B, C, and E, derived from DSSP software (version 2.2.1) (25, 26) were reduced to three states: only H was mapped to helix, E was mapped to strand, and all others were mapped to coil (31). For example, if a residue A in helix contacts with a residue B in strand, with a residue C in coil, and with a residue D in helix, the ERCE score for residue A is the sum of AB in Helix_Strand matrix, AC in Helix_Coil matrix, and AD in Helix_Helix matrix. For the mutant amino acid, the contacting neighbors and secondary structural states remained the same as for the wild-type residue. By using

ERCE, we obtained another four features.

2.14 Context-Dependent Substitution Score (CDSS)

Here, “context” means the solvent accessibility level and the secondary structure state of an AAS position. Eight substitution matrices, covering the contexts of Buried_Beta, Buried_Coil, Buried_Helix, Buried_Turn, Exposed_Beta, Exposed_Coil, Exposed_Helix, and Exposed_Turn, provide the structural and functional constraints in evolution within each specific context (32). According to their definition, residues with relative solvent accessibility (RSA) larger than 18% were considered exposed, and otherwise buried. Eight secondary structural states, namely H, I, G, S, T, B, C, and E, derived from DSSP software (version 2.2.1) (25, 26) were reduced to four states: H and I to helix, B and C to coil, E to beta, and all others to turn. Based on its “context”, the CDSS of an AAS can be retrieved from the eight matrices accordingly. For example, when an AAS occurs at a position with less than 18% RSA and secondary structural state H, the CDSS for this AAS is obtained from the Buried_Helix matrix.

2.15 MIcomulative

The reduction in uncertainty about one site due to another site can be measured by the mutual information (MI) of these two sites in a multiple sequence alignment. The sum of corrected MI values for a residue with all its structural contacting neighbors is defined as MIcomulative (33), which was directly calculated by using RING (version 2.0) (7, 8) after running PSI-BLAST (4) locally against UniRef90 database (Release 2018_06) (ncbi-blast-2.2.29+, three iterations, E-value 0.0001) (5) for each sequence segment parsed from the 3D structure. Intuitively, MIcomulative reflects the importance of a residue to its neighboring residues.

2.16 Relative solvent accessibility, Hydrogen bond, and ΔΔG

Solvent accessibility is a measure of residue exposure and has shown to be relevant to identifying disease-associated AAS (34, 35). Here, we calculated the relative solvent accessibility of the wild-type residue of the AAS in the 3D structure by using NACCESS (version 2.1.1) (36).

HBPLUS (version 3.15) was used to calculate the number of hydrogen bonds that the wild-type residue at the AAS site might form (37).

Decreased protein stability is a major molecular consequence of disease-associated AAS, so protein stability change will be a useful feature for the prediction of disease-associated AAS (38, 39). Free energy change (ΔΔG), a measure of protein stability change, was calculated by using FoldX (v.3) (40, 41). ΔΔG values were obtained by using the “BuildModel” command based on the optimized structures obtained from the “RepairPDB” command.

3. Performance evaluation and comparison with other tools

Six recommended performance metrics are defined as follows by regarding daAASs as positive samples and nAASs as negative samples (42):

$$\text{accuracy (ACC)} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

$$\text{sensitivity (Sen)} = \text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{specificity (Spe)} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

$$\text{PPV} = \text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}}$$

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

The receiver operating characteristic (ROC) curves for each testing dataset were generated by plotting sensitivity against 1-specificity (False Positive Rate, or FPR) at each cutoff for classifying the AAS into two categories. The area under ROC curve (AUC) was calculated to provide a cutoff-independent measure of the prediction performance as recommended (42, 43).

The prediction results on the testing datasets of other predictors were obtained according to the settings described below. SIFT (version 5.2.2) (44), HumDiv-trained PolyPhen-2 (PPH2_HD) and HumVar-trained PolyPhen-2 (PPH2_HV) (version 2.2.2r405c) (45), PROVEAN (version 1.1.5) (46), FATHMM's weighted method (FATHMM-W) and FATHMM's unweighted method (FATHMM-U) (version 2.3) (47), and PANTHER-PSEP (version 1.01) (48) were run locally with default settings for their parameters. Moreover, for those predictors relying on a locally installed BLAST+ package as the backend engine, including SIFT, PolyPhen-2, PROVEAN, and PANTHER-PSEP, ncbi-blast-2.2.29+ was adopted (5). For PROVEAN, which relies on CD-HIT, the cd-hit-v4.5.7-2011-12-16 (49, 50) was utilized as backend engine. In addition to these required packages,

several predictors also need to configure one or more specific databases accordingly. SIFT was run based on UniRef90 (Release 2018_06). PolyPhen-2 was run based on UniRef100 (Release 2018_06), and the required structural databases (PDB and DSSP) were downloaded on August 15, 2018. For PROVEAN, the NCBI nr database downloaded on May 12, 2017 was used.

For comparison, the prediction results of other predictors should also be converted to binary classification if its raw output is not. An AAS assigned by PANTHER-PSEP as “probably damaging” or “possibly damaging” was classified as “disease”, and “probably benign” as “neutral”. The binary classification of PolyPhen-2 was directly used. For PROVEAN, the author-recommended thresholds were adopted to output the binary classification. In detail, an AAS with a PROVEAN score $\leqslant$ -2.5 or $>$ -2.5 was assigned “disease” or “neutral”, respectively.

REFERENCES

[1] B. Webb and A. Sali: Comparative Protein Structure Modeling Using MODELLER. *Curr Protoc Bioinformatics*, 54, 5 6 1-5 6 37 (2016) doi:10.1002/cpbi.3

[2] U. Pieper, B. M. Webb, G. Q. Dong, D. Schneidman-Duhovny, H. Fan, S. J. Kim, N. Khuri, Y. G. Spill, P. Weinkam, M. Hammel, J. A. Tainer, M. Nilges and A. Sali: ModBase, a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Research*, 42(D1), D336-D346 (2014) doi:10.1093/nar/gkt1144

[3] P. J. A. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski and M. J. L. de Hoon: Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11), 1422-1423 (2009) doi:10.1093/bioinformatics/btp163

[4] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17), 3389-402 (1997) doi:10.1093/nar/25.17.3389

[5] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer and T. L. Madden: BLAST+: architecture and applications. *BMC Bioinformatics*, 10, 421 (2009) doi:10.1186/1471-2105-10-421

[6] V. Lopez-Ferrando, A. Gazzo, X. de la Cruz, M. Orozco and J. L. Gelpi: PMut: a web-based tool for the annotation of pathological variants on proteins, 2017 update. *Nucleic Acids Res*, 45(W1), W222-W228 (2017) doi:10.1093/nar/gkx313

[7] A. J. Martin, M. Vidotto, F. Boscariol, T. Di Domenico, I. Walsh and S. C. Tosatto: RING: networking interacting residues, evolutionary information and energetics in protein structures. *Bioinformatics*, 27(14), 2003-5 (2011) doi:10.1093/bioinformatics/btr191

[8] D. Piovesan, G. Minervini and S. C. Tosatto: The RING 2.0 web server for high quality residue interaction networks. *Nucleic Acids Res*, 44(W1), W367-74 (2016) doi:10.1093/nar/gkw315

[9] M. Gribskov, A. D. McLachlan and D. Eisenberg: Profile analysis: detection of distantly related

proteins. *Proc Natl Acad Sci U S A*, 84(13), 4355-8 (1987) doi:10.1073/pnas.84.13.4355
[10] Y. Itan, L. Shang, B. Boisson, E. Patin, A. Bolze, M. Moncada-Velez, E. Scott, M. J. Ciancanelli, F. G. Lafaille, J. G. Markle, R. Martinez-Barricarte, S. J. de Jong, X. F. Kong, P. Nitschke, A. Belkadi, J. Bustamante, A. Puel, S. Boisson-Dupuis, P. D. Stenson, J. G. Gleeson, D. N. Cooper, L. Quintana-Murci, J. M. Claverie, S. Y. Zhang, L. Abel and J. L. Casanova: The human gene damage index as a gene-level approach to prioritizing exome variants. *Proceedings of the National Academy of Sciences of the United States of America*, 112(44), 13615-13620 (2015) doi:10.1073/pnas.1518646112
[11] S. Petrovski, Q. Wang, E. L. Heinzen, A. S. Allen and D. B. Goldstein: Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet*, 9(8), e1003709 (2013) doi:10.1371/journal.pgen.1003709
[12] S. Henikoff and J. G. Henikoff: Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, 89(22), 10915-9 (1992) doi:10.1073/pnas.89.22.10915
[13] R. Grantham: Amino-Acid Difference Formula to Help Explain Protein Evolution. *Science*, 185(4154), 862-864 (1974) doi:10.1126/science.185.4154.862
[14] T. Miyata, S. Miyazawa and T. Yasunaga: Two types of amino acid substitutions in protein evolution. *J Mol Evol*, 12(3), 219-36 (1979) doi:10.1007/bf01732340
[15] C. J. Epstein: Non-randomness of amino-acid changes in the evolution of homologous proteins. *Nature*, 215(5099), 355-9 (1967) doi:10.1038/215355a0
[16] B. Clarke: Selective constraints on amino-acid substitutions during the evolution of proteins. *Nature*, 228(5267), 159-60 (1970) doi:10.1038/228159a0
[17] A. D. McLachlan: Tests for comparing related amino-acid sequences. Cytochrome c and cytochrome c 551. *J Mol Biol*, 61(2), 409-24 (1971) doi:10.1016/0022-2836(71)90390-1
[18] J. Kyte and R. F. Doolittle: A simple method for displaying the hydropathic character of a protein. *J Mol Biol*, 157(1), 105-32 (1982) doi:10.1016/0022-2836(82)90515-0
[19] J. M. Zimmerman, N. Eliezer and R. Simha: The characterization of amino acid sequences in proteins by statistical methods. *J Theor Biol*, 21(2), 170-201 (1968) doi:10.1016/0022-5193(68)90069-6
[20] W. R. Atchley, J. P. Zhao, A. D. Fernandes and T. Druke: Solving the protein sequence metric problem. *Proceedings of the National Academy of Sciences of the United States of America*, 102(18), 6395-6400 (2005) doi:10.1073/pnas.0408677102
[21] E. Capriotti, R. Calabrese and R. Casadio: Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics*, 22(22), 2729-34 (2006) doi:10.1093/bioinformatics/btl423
[22] Z. Q. Ye, S. Q. Zhao, G. Gao, X. Q. Liu, R. E. Langlois, H. Lu and L. Wei: Finding new structural and sequence attributes to predict possible disease association of single amino acid polymorphism (SAP). *Bioinformatics*, 23(12), 1444-50 (2007) doi:10.1093/bioinformatics/btm119
[23] H. Zhou, M. Gao and J. Skolnick: ENTPRISE: An Algorithm for Predicting Human Disease-Associated Amino Acid Substitutions from Sequence Entropy and Predicted Protein Structures. *PLoS One*, 11(3), e0150965 (2016) doi:10.1371/journal.pone.0150965
[24] E. Capriotti and R. B. Altman: Improving the prediction of disease-related variants using protein three-dimensional structure. *BMC Bioinformatics*, 12 Suppl 4, S3 (2011) doi:10.1186/1471-2105-12-S4-S3
[25] W. Kabsch and C. Sander: Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12), 2577-637 (1983)

doi:10.1002/bip.360221211
[26] W. G. Touw, C. Baakman, J. Black, T. A. te Beek, E. Krieger, R. P. Joosten and G. Vriend: A series of PDB-related databanks for everyday needs. *Nucleic Acids Res*, 43(Database issue), D364-8 (2015) doi:10.1093/nar/gku1028
[27] R. K. Grewal and S. Roy: Modeling proteins as residue interaction networks. *Protein Pept Lett*, 22(10), 923-33 (2015) doi:10.2174/0929866522666150728115552
[28] A. A. Hagberg, D. A. Schult and P. J. Swart: Exploring network structure, dynamics, and function using NetworkX. In: *Proceedings of the 7th Python in Science Conference (SciPy2008)*. Ed G. Varoquaux, T. Vaught&J. Millman. Pasadena, CA USA (2008)
[29] S. Miyazawa and R. L. Jernigan: Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues. *Proteins*, 34(1), 49-68 (1999) doi:10.1002/(sici)1097-0134(19990101)34:1<49::aid-prot5>3.0.co;2-l
[30] S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama and M. Kanehisa: AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res*, 36(Database issue), D202-5 (2008) doi:10.1093/nar/gkm998
[31] C. Zhang and S. H. Kim: Environment-dependent residue contact energies for proteins. *Proc Natl Acad Sci U S A*, 97(6), 2550-5 (2000) doi:10.1073/pnas.040573597
[32] J. M. Koshi and R. A. Goldstein: Context-Dependent Optimal Substitution Matrices. *Protein Engineering*, 8(7), 641-645 (1995) doi:10.1093/protein/8.7.641
[33] S. D. Dunn, L. M. Wahl and G. B. Gloor: Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, 24(3), 333-340 (2008) doi:10.1093/bioinformatics/btm604
[34] R. J. Dobson, P. B. Munroe, M. J. Caulfield and M. A. Saqi: Predicting deleterious nsSNPs: an analysis of sequence and structural attributes. *BMC Bioinformatics*, 7, 217 (2006) doi:10.1186/1471-2105-7-217
[35] C. T. Saunders and D. Baker: Evaluation of structural and evolutionary contributions to deleterious mutation prediction. *Journal of Molecular Biology*, 322(4), 891-901 (2002) doi:10.1016/S0022-2836(02)00813-6
[36] S. J. Hubbard and J. M. Thornton: 'NACCESS', Computer Program, Department of Biochemistry and Molecular Biology, University College London. (1993)
[37] I. K. McDonald and J. M. Thornton: Satisfying hydrogen bonding potential in proteins. *J Mol Biol*, 238(5), 777-93 (1994) doi:10.1006/jmbi.1994.1334
[38] P. Yue, Z. L. Li and J. Moult: Loss of protein structure stability as a major causative factor in monogenic disease. *Journal of Molecular Biology*, 353(2), 459-473 (2005) doi:10.1016/j.jmb.2005.08.020
[39] Z. Wang and J. Moult: SNPs, protein structure, and disease. *Hum Mutat*, 17(4), 263-70 (2001) doi:10.1002/humu.22
[40] J. Schymkowitz, J. Borg, F. Stricher, R. Nys, F. Rousseau and L. Serrano: The FoldX web server: an online force field. *Nucleic Acids Research*, 33, W382-W388 (2005) doi:10.1093/nar/gki387
[41] S. Khan and M. Vihinen: Performance of Protein Stability Predictors. *Human Mutation*, 31(6), 675-684 (2010) doi:10.1002/humu.21242
[42] M. Vihinen: How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. *BMC Genomics*, 13 Suppl 4, S2 (2012) doi:10.1186/1471-2164-13-s4-s2

[43] M. Vihinen: Guidelines for Reporting and Using Prediction Tools for Genetic Variation Analysis. *Human Mutation*, 34(2), 275-277 (2013) doi:10.1002/humu.22253
[44] P. C. Ng and S. Henikoff: Predicting deleterious amino acid substitutions. *Genome Res*, 11(5), 863-74 (2001) doi:10.1101/gr.176601
[45] I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov and S. R. Sunyaev: A method and server for predicting damaging missense mutations. *Nat Methods*, 7(4), 248-9 (2010) doi:10.1038/nmeth0410-248
[46] Y. Choi, G. E. Sims, S. Murphy, J. R. Miller and A. P. Chan: Predicting the functional effect of amino acid substitutions and indels. *PLoS One*, 7(10), e46688 (2012) doi:10.1371/journal.pone.0046688
[47] H. A. Shihab, J. Gough, D. N. Cooper, P. D. Stenson, G. L. A. Barker, K. J. Edwards, I. N. M. Day and T. R. Gaunt: Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models. *Human Mutation*, 34(1), 57-65 (2013) doi:10.1002/humu.22225
[48] H. M. Tang and P. D. Thomas: PANTHER-PSEP: predicting disease-causing genetic variants using position-specific evolutionary preservation. *Bioinformatics*, 32(14), 2230-2232 (2016) doi:10.1093/bioinformatics/btw222
[49] W. Li and A. Godzik: Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13), 1658-9 (2006) doi:10.1093/bioinformatics/btl158
[50] L. Fu, B. Niu, Z. Zhu, S. Wu and W. Li: CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23), 3150-2 (2012) doi:10.1093/bioinformatics/bts565

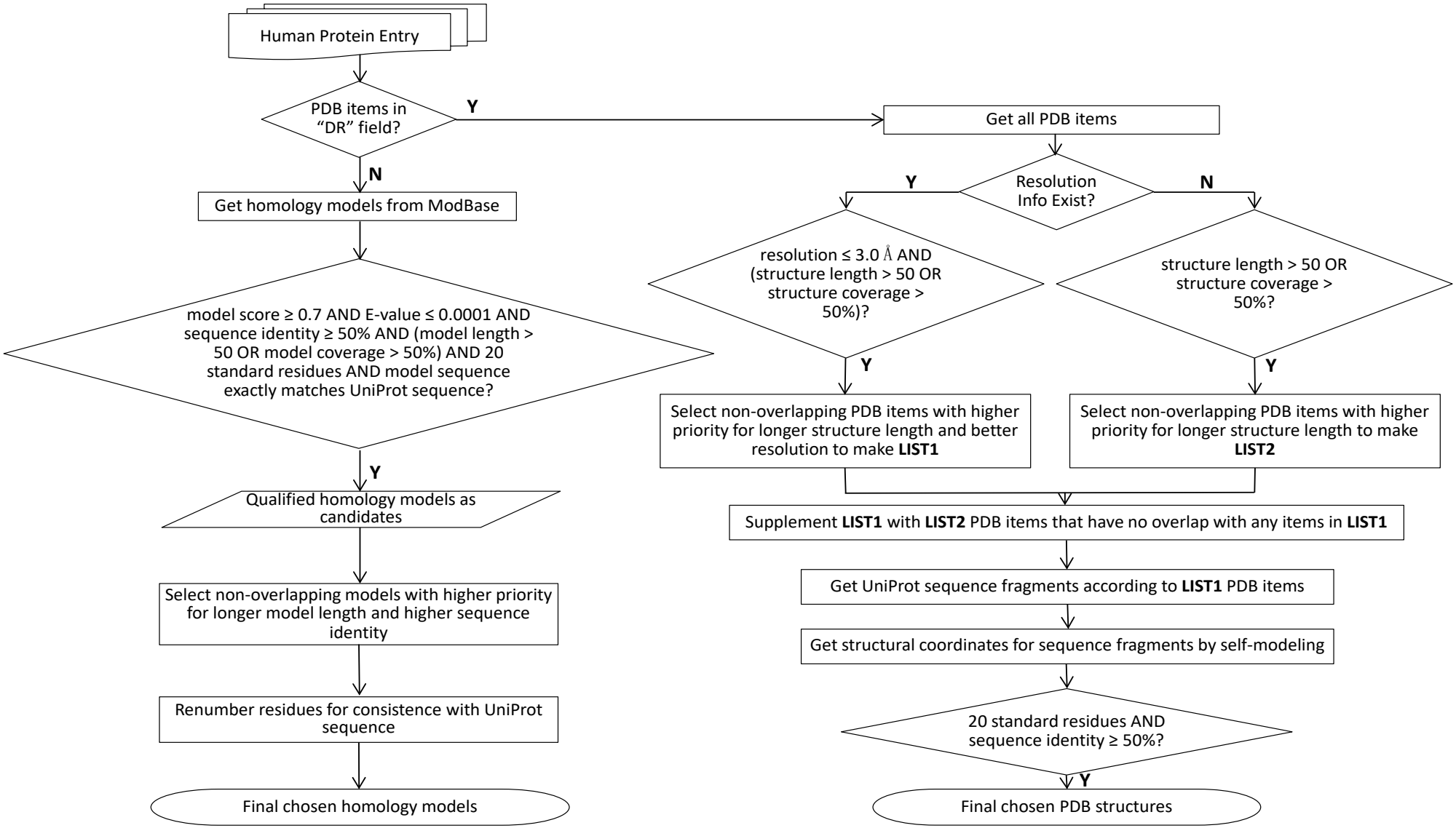

Fig. S1. Pipeline of structure acquiring and data cleaning. Model coverage or structure coverage are defined by the model length or structure length divided by the UniProt sequence length, respectively.

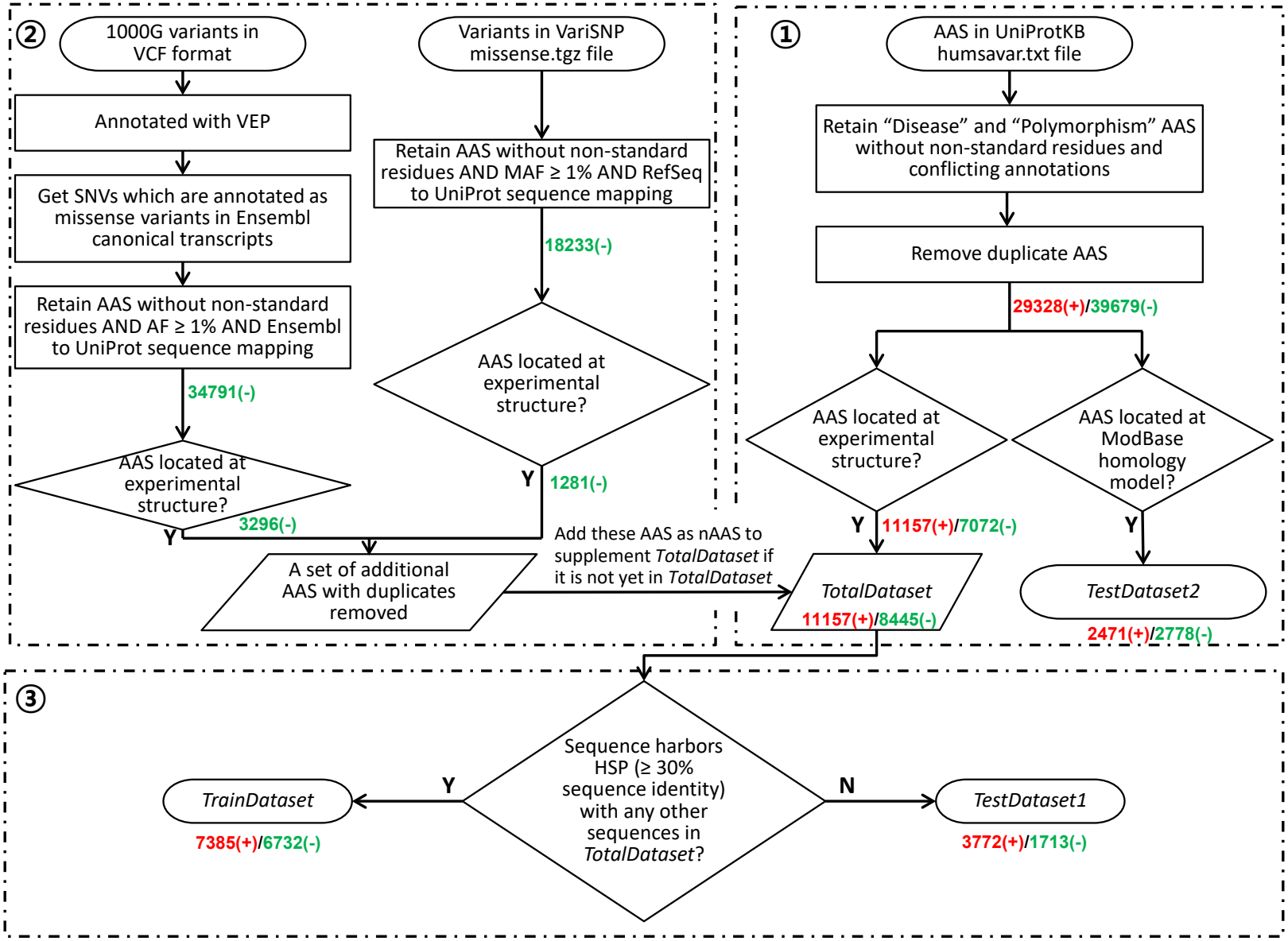

Fig. S2. Overall pipeline of AAS datasets preparation. The counts of AASs at each step are given in red (daAAS) and green (nAAS) numbers.

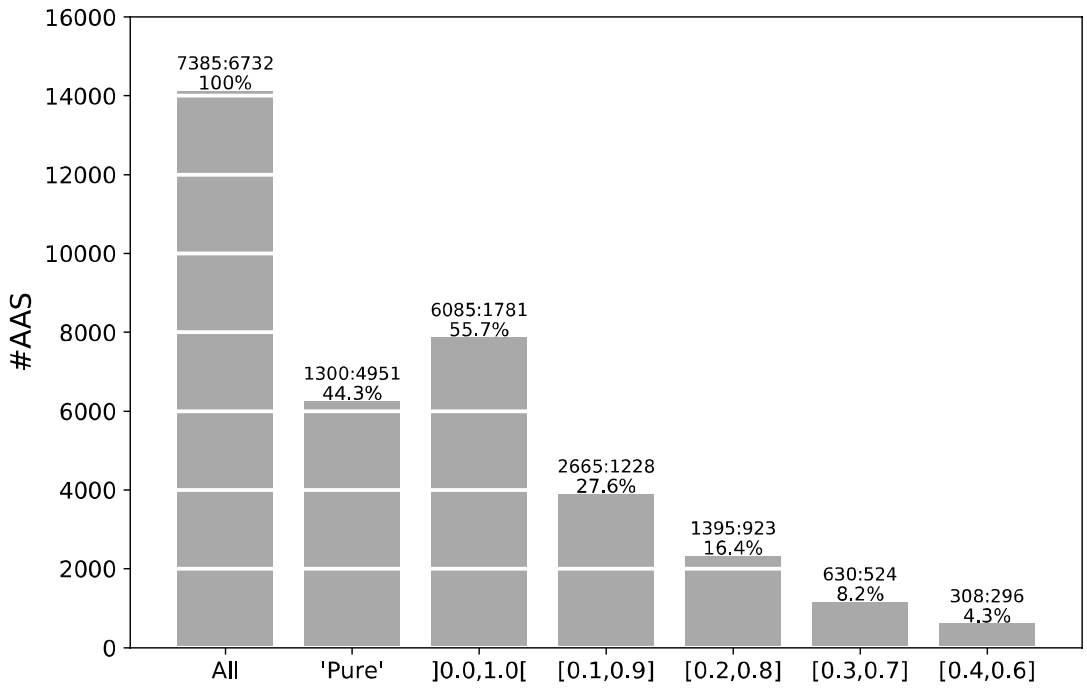

Fig. S3. The composition of different AAS subsets of *TrainDataset* defined by the proportion range of daAAS in each protein.

The 'All' denotes the whole *TrainDataset*, while 'Pure' contains only the AASs whose proteins carry only daAAS (proportion: 100%) or nAAS (proportion: 0%). Other labels of the X axis represent the intervals of the proportion of daAAS within its protein when constructing these subsets. The numbers of daAASs and nAASs of each subset are given on top of each bar. The percentage of AASs in each subset is also indicated.

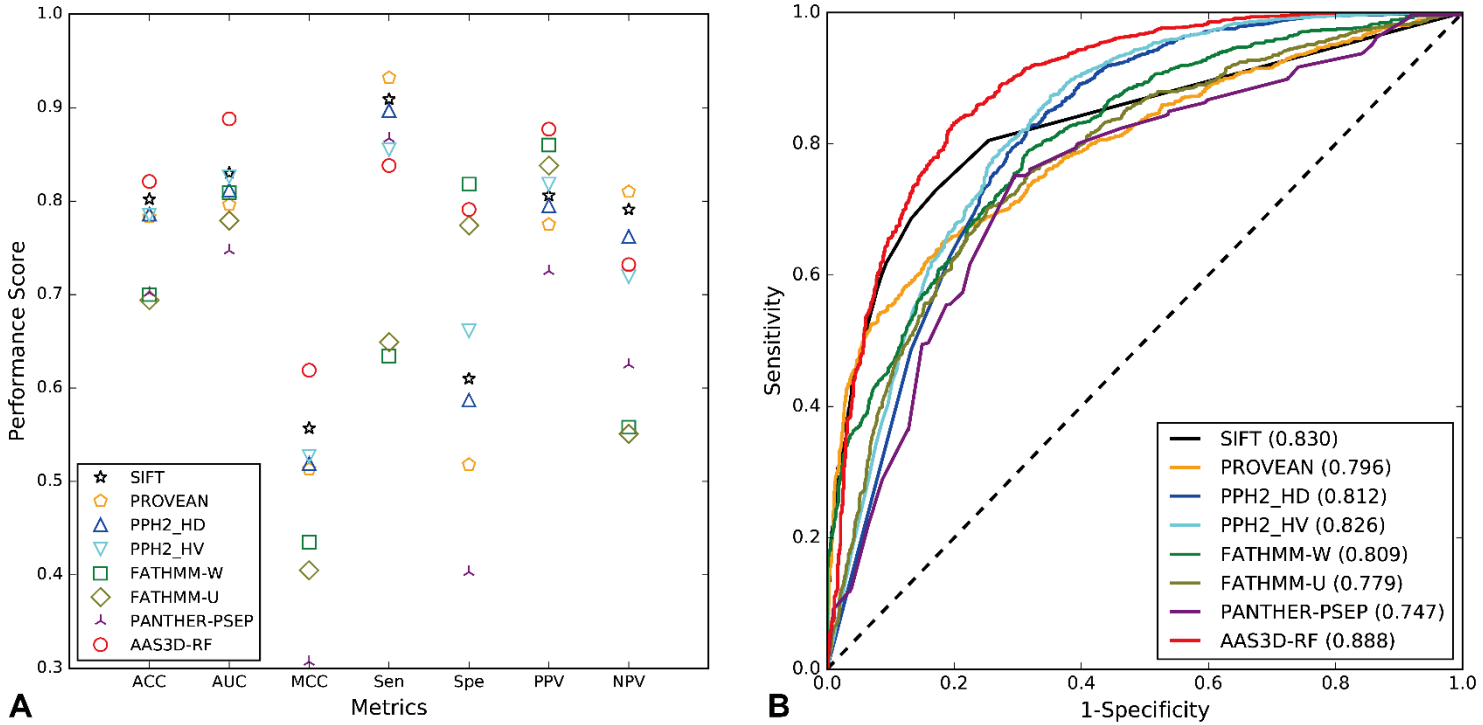

Fig. S4. Performance comparison of AAS3D-RF with seven other predictors on *TestDataset3*. (A) The performance in terms of seven metrics. (B) The ROC curves.

Table S1. The pre-specified ranges of hyperparameters for random search

Hyperparameter	Range
n_estimators	[50, 500]
max_depth	[3, 10]
max_features	[0.5, 1.0]

Table S2. The selected 21 features[1]

#	Feature Name	Description	Filter	Weighted
1	nwt	Number of wild-type amino acids in the aligned position	All	No
2	nal_no_human	Number of sequences in the alignment	No Human	No
3	nal_1e-45	Number of sequences in the alignment	E-value < 10^{-45}	No
4	naa_w	Number of amino acids in the aligned position	All	BLAST score
5	pwm_w	Position Weight Matrix score	All	BLAST score
6	naa_human_w	Number of amino acids in the aligned position	Human	BLAST score
7	nwt_1e-45_w	Number of wild-type amino acids in the aligned position	E-value < 10^{-45}	BLAST score
8	rmt_1e-45_w	Relative number of aligned mutant amino acids	E-value < 10^{-45}	BLAST score
9	Entropy_RING	Shannon's entropy of a given AAS position obtained from RING software		
10	RVIS	Residual Variation Intolerance Score assessing human genes' ability to tolerate common functional genetic variation in healthy individuals		
11	Miyata	Score of physicochemical dissimilarities between amino acid in Miyata matrix		
12	Abs_dF1	Absolute difference between factor Ⅰ solution score (bipolar) of mutant and wild-type amino acids		
13	Abs_dpI	Absolute difference between Isoelectric point of mutant and wild-type amino acids		
14	SeqDis_DISULFID	The least sequence distance between the AAS site and disulfide-bonded Cys		
15	dF1_Mean_Mutant	Difference between factor Ⅰ solution score of neighbors' mean value and mutant amino acid		
16	dVol_Mean_Mutant	Difference between Grantham molecular volume of neighbors' mean value and mutant amino acid		
17	varHP_Wild_NB	Variance of Kyte-Doolittle hydropathy index for the wild-type amino acid and its neighbors		
18	ERCE	The sum of Environment-dependent Residue Contact Energy between the wild-type residue and all its contacting neighbors derived from 6 matrices		
19	CDSS	Context-Dependent Substitution Score derived from 8 matrices		
20	RSA	3D structure-based residue Relative Solvent Accessibility computed by NACCESS		
21	ΔΔG	Mutational free energy change, a measure of protein stability change, calculated by FoldX		

[1]#1-9 are conservation-related features, #10 is gene-level metric, #11-13 reflect the physicochemical dissimilarities, #14 is the sequence distance between the AAS position and closest disulfide-bonded Cys, #15-21 are 3D structural features.

Table S3. Performance comparison of AAS3D-RF and structure-removed predictor on the two independent testing datasets

Predictors	Total	TP	FN	TN	FP	ACC	MCC	AUC	Sen	Spe	PPV	NPV
TestDataset1												
AAS3D-RF	5485	3068	704	1380	333	0.811	0.591	0.877	0.813	0.806	0.902	0.662
structure-removed	5485	2926	846	1368	345	0.783	0.542	0.864	0.776	0.799	0.895	0.618
TestDataset2												
AAS3D-RF	5249	2210	261	2192	586	0.839	0.684	0.913	0.894	0.789	0.790	0.894
structure-removed	5249	2204	267	2178	600	0.835	0.676	0.911	0.892	0.784	0.786	0.891
TestDataset2 **after removing sequence homologous to** ***TrainDataset*** **with $\geqslant$ 30% identity**												
AAS3D-RF	2003	1074	141	576	212	0.824	0.627	0.884	0.884	0.731	0.835	0.803

Table S4. Performance comparison of AAS3D-RF with seven popular predictors on the two independent testing datasets[1]

Predictors	Total[2]	TP	FN	TN	FP	ACC	MCC	AUC	Sen	Spe	PPV	NPV
TestDataset1												
SIFT	5371	3262	451	995	663	0.793	0.499	0.811	0.879	0.600	0.831	0.688
PPH2_HD	5485	3354	418	978	735	0.790	0.490	0.810	0.889	0.571	0.820	0.701
PPH2_HV	5485	3158	614	1149	564	0.785	0.504	0.829	0.837	0.671	0.848	0.652
PROVEAN	5485	3423	349	882	831	0.785	0.469	0.795	**0.907**	0.515	0.805	0.716
FATHMM-W	5394	3268	450	1278	398	**0.843**	**0.636**	**0.902**	0.879	0.763	0.891	**0.740**
FATHMM-U	5440	2466	1272	1308	394	0.694	0.398	0.774	0.660	0.769	0.862	0.507
PANTHER-PSEP	4864	3041	352	614	857	0.751	0.361	0.759	0.896	0.417	0.780	0.636
AAS3D-RF	5485	3068	704	1380	333	0.811	0.591	0.877	0.813	**0.806**	**0.902**	0.662
TestDataset2												
SIFT	4762	2021	154	1495	1092	0.738	0.531	0.834	0.929	0.578	0.649	**0.907**
PPH2_HD	5249	2255	216	1697	1081	0.753	0.543	0.837	0.913	0.611	0.676	0.887
PPH2_HV	5249	2177	294	1898	880	0.776	0.571	0.855	0.881	0.683	0.712	0.866
PROVEAN	5249	2301	170	1497	1281	0.724	0.504	0.808	**0.931**	0.539	0.642	0.898
FATHMM-W	5195	1945	510	2215	525	0.801	0.600	0.889	0.792	**0.808**	0.787	0.813
FATHMM-U	5202	1680	782	2185	555	0.743	0.484	0.813	0.682	0.797	0.752	0.736
PANTHER-PSEP	4232	1856	268	984	1124	0.671	0.373	0.765	0.874	0.467	0.623	0.786
AAS3D-RF	5249	2210	261	2192	586	**0.839**	**0.684**	**0.913**	0.894	0.789	**0.790**	0.894

[1]The best value of each performance metric is in bold and underlined.

[2]The "Total" numbers are different because not all AASs have received prediction result for some predictors.

Table S5. Performance comparison of AAS3D-RF and structure-removed predictor on AAS data with and without reliable conservation features

	TP	TN	FP	FN	MCC	ACC
***TestDataset1* with nal_1e-45<200**						
AAS3D-RF	75	61	8	35	0.552	0.760
structure-removed	40	63	6	70	0.308	0.575
***TestDataset1* with nal_1e-45>=200**						
AAS3D-RF	2993	1319	325	669	0.592	0.813
structure-removed	2886	1305	339	776	0.551	0.790
***TestDataset2* with nal_1e-45<200**						
AAS3D-RF	7	26	0	3	0.792	0.917
structure-removed	7	24	2	3	0.645	0.861
***TestDataset2* with nal_1e-45>=200**						
AAS3D-RF	2203	2166	586	258	0.683	0.838
structure-removed	2197	2154	598	264	0.676	0.835

Table S6. Overview of *TestDataset3*

Class	# of AAS	# of Proteins[1]	# of Structures[1]
Disease	1,675	339	359
Neutral	939	334	347
Total	2,614	582	616

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

Table S7. Performance comparison on *TestDataset3*[1]

Predictors	Total[2]	TP	FN	TN	FP	ACC	MCC	AUC	Sen	Spe	PPV	NPV
SIFT	2353	1369	137	517	330	0.802	0.557	0.830	0.909	0.610	0.806	0.791
PPH2_HD	2614	1503	172	551	388	0.786	0.519	0.812	0.897	0.587	0.795	0.762
PPH2_HV	2614	1432	243	621	318	0.785	0.527	0.826	0.855	0.661	0.818	0.719
PROVEAN	2614	1561	114	486	453	0.783	0.513	0.796	**0.932**	0.518	0.775	**0.810**
FATHMM-W	2552	1032	597	755	168	0.700	0.435	0.809	0.634	**0.818**	0.860	0.558
FATHMM-U	2585	1078	583	715	209	0.694	0.405	0.779	0.649	0.774	0.838	0.551
PANTHER-PSEP	1979	1106	170	283	420	0.702	0.307	0.747	0.867	0.403	0.725	0.625
AAS3D-RF	2614	1403	272	743	196	**0.821**	**0.619**	**0.888**	0.838	0.791	**0.877**	0.732

[1]The best value of each performance metric is in bold and underlined.

[2]The "Total" numbers are different because not all AASs have received prediction result for some predictors.