Supporting Information

Supporting Texts

Text S1. Proteomic mapping database construction
To have a quick access to the structural proteomic data in the DAMpred training process, a set of derivative libraries are constructed from two primary sequence and structure databases, UniProtKB [1] and PDB [2], with a mapping pipeline depicted in Figure S1. The UniProtKB is the central access point with extensively curated protein information, containing function, classification and cross-references of protein sequences. It includes two components of Swiss-Prot containing manually annotated records and TrEMBL with computationally analyzed records awaiting manual annotation. We extract all human entries from Swiss-Prot and TrEMBL, and construct two derivative libraries. The first is a feature table that lists the binding sites, enzyme active sites, posttranslational modifications, and other characteristics reported in the cited references; the second is a sequence library that is built by the makeblastdb program from the FASTA files of all human sequences in UniProt, prepared for DAMpred homology search. In addition, a lookup table of ‘x2acc’ is built from the UniProt idmapping data, which records the mapping information of the UniProt accession and other cross-reference IDs (Gene ID, RefSeq, PDB GO etc).

The biological assembly (also referred as BioUnit) is the macromolecular assembly that has either been shown (or is believed) to be the minimum functional form of the molecule. In order to quickly retrieve the right biological information of the target proteins, a BioUnit database is constructed by collecting all the UniProt structures in the table x2acc that have a cross-reference ID to the PDB from the PDB ftp site (ftp://ftp.wwpdb.org/pub/pdb/data/biounit/PDB/all/). For each entry, the PDB residue position is mapped to the UniProt sequence residue position based on the DBREF record. Accordingly, a new table resMap is collected, which contains records of UniProt accession (acc), PDB ID, the initial and last sequence number of the PDB sequence segment (pdbSeqStart, pdbSeqEnd), the initial sequence number of the UniProt sequence segment (UniProtStart), the start and last residue number of the PDB structure part (structStart, structEnd), resolution, and BioUnit chains of each protein entry (Figure S1).

The architecture of the derivative databases are constructed using SQLite, a relational database management system [3], where the entire package of the databases can be downloaded at http://zhanglab.ccmb.med.umich.edu/DAMpred/ download/human.sqlite.

Test S2. Feature collections of DAMpred
In total, we collected 70 features that are extracted from physicochemical properties, biological assembly, and I-TASSER structural prediction. A detailed list of the features is given in Table S2, with the feature extraction process depicted in Figure 1A. The 70 features are categorized into four groups based on their properties.

Physicochemical properties
The physicochemical property features in DAMpred include the pharmacophore of the target residues and the mutation-induced environmental pharmacophore changes.

Pharmacophore of residues. A pharmacophore is an abstract description of the structural and chemical properties of the amino acids, which can be represented by a set of numerical values, known as the pharmacophore vector [4, 5]. Pharmacophore properties of each residue considered by DAMpred include hydrophobicity (HP or noHP), aromatic rings (AR or noAR), and charge (positive: PC, negative: NC and neutral: noC). In addition, DAMpred considers the pair-wise polarity and hydrogen-bonding interactions. The polarity is specified as BP (both are polar), OP (either is polar), or NP (both are nonpolar) for the target residue pairs. The hydrogen bond is specified by four different atom types of O-H⋯:N, O-H⋯:O, N-H⋯:N, and N-H⋯:O, where the
hydrogen-donor-acceptor angle and hydrogen-acceptor distance cutoffs are set as 30° and 3.5 Å respectively. The counts of the acceptor (AC) and donor (DO) residues are also considered for each protein. This group of features is listed as Features 1-24 at the top of Table S2.

We note that the concept of pharmacophore vector has been previously used by Pires et al in mCSM [4, 5]. However, there are several essential differences between the implementations of the pharmacophore vector in mCSM and DAMpred. First, DAMpred measures the items in pharmacophore in the unit of residues instead of atoms by mCSM; Second, DAMpred computes the physicochemical properties based on the amino acid types and the I-TASSER structural models, which is different from mCSM that uses a patent protected program PMapper; Third, DAMpred considers the contact interactions of residues from both wild and mutant structures, while mCSM used graph-based atom distance patterns considering only wild structure. Thus, although the concept is quite similar, the actual implementation and content of the pharmacophore vector in DAMpred are different from mCSM.

Mutation-induced environmental pharmacophore changes. Fourteen environmental pharmacophore features are considered (see Figure S11). Here, the pharmacophore vector for ith residue of the query protein is written as $p_{ic} = [p_{i1}, p_{i2}, \ldots, p_{il}]$, where l is the number of pharmacophore types and $e[m, w]$, with m and w indicating mutant and wild structures respectively. The environmental pharmacophore changes due to the ith mutant can be calculated by $\cos_{wm}(i) = \cos(\bar{p}_w, \bar{p}_m)$ and $\text{rmsd}_{wm}(i) = \text{rmsd}(\bar{p}_w, \bar{p}_m)$. The neighbor environment pharmacophore counts for the interactions of the target residue and all residues in contact ($n_i$), i.e., $\bar{p}_{ic} = \sum_{k=1}^{n_i} \bar{p}_{ik}$. Thus, the neighbor environmental pharmacophore changes due to the ith mutation can be written as $\cos_{nwm}(i) = \cos(\bar{p}_{icw}, \bar{p}_{icmp})$ and $\text{rmsd}_{nwm}(i) = \text{rmsd}(\bar{p}_{icw}, \bar{p}_{icmp})$. Here, a contact is defined for two residues if the distance of any heavy atoms from them is below 4.2 Å in the protein structure. We consider two types of properties ($l = l_1 + l_2$), where one is related with single residue (Hydrophobic, aromatic rings and charge, volume and weight) and another with paired residues (polarity and hydrogen bond). The neighbor pharmacophore vector corresponding to single residue (or paired residues) is denoted by $\bar{p}_{ic, ts}$ (or $\bar{p}_{ic, fp}$), where the corresponding pharmacophore changes are then calculated by

$$\cos(\bar{p}_w, \bar{p}_m) = \frac{\bar{p}_w \cdot \bar{p}_m}{\|\bar{p}_w\| \cdot \|\bar{p}_m\|}$$

$$\text{rmsd}(\bar{p}_w, \bar{p}_m) = \sqrt{\frac{\sum_{k=1}^{l_1} (p_{ik}^w - p_{ik}^m)^2}{l}}$$

This group of features is listed as Features 25-32 in Table S2.

In addition to the local and environmental pharmacophores, we consider the common physicochemical properties, including the volume and weight from the wild-type and mutant residues, which are listed as Feature 33-38 in Table S2.

**Evolutionary profiles.**

Evolution is a major driven force for protein structure and function determination, where sequence profiles from multiple sequence alignments contain information on how the protein families evolve. To identify distant-homology relations between sequences, three sequence profiles are collected in DAMpred by PSI-BLAST [6], LOMETS [7] and Pfam [8] separately.

**PSI-BLAST profile.** The wild-type and mutant sequences are searched through the Uniref90 non-redundant sequence database by PSI-BLAST [6] with three iterations and an E-value cutoff 0.001. The resultant sequences are then passed to Clustal Omega [9] to obtain multiple sequence alignments (MSAs), and the position-specific independent count (PSIC) scores are then calculated from both MSAs from wildtype and mutant sequences. At the mutant position $i$, the PSIC score for both wild-type ($S_{iw}$) and mutant ($S_{im}$) amino acids are defined by
$\begin{align*}
PSIC(a, i) &= \ln \left( \frac{p_{ia}}{q_a} \right) \\
p_{ia} &= \frac{n(a, i)_{eff}}{\sum_b n(b, i)_{eff}}
\end{align*}$

(S2)

where $q_a$ is the background frequency of amino acid $a$ in the MSA; $p_{ia}$ is the probability of the amino acid $a$ at the $i$th alignment position $i$, with $n(a, i)_{eff}$ being the number of counts of $a$ at the $i$th position. Compared to the widely used position-specific scoring matrix (PSSM) score [10], the major advantage of the PSIC is that $n(a, i)_{eff}$ is calculated from the overall similarity of the sequences that share the amino acid type at this position with the help of the structural features, which allows the fast computation of the true position-specific sequence weights [11].

In addition, the Jensen-Shannon divergence (JSD) score is an index measuring the extent of the evolutionary conservation of each residue position along the protein chain, which has been previously shown to provide state-of-the-art performance in identifying catalytic sites and ligand binding sites [12, 13]. The JSD score in DAMpred is calculated by [12]

$$\begin{align*}
\text{JSD}_{ia} &= \lambda p_{ia} \log \frac{p_{ia}}{c_{ia}} + (1 - \lambda) q_a \log \frac{q_a}{c_{ia}} \\
\text{JSD}_i &= \sum_{a \in A} \text{JSD}_{ia}
\end{align*}$$

(S3)

where $c_{ia} = \lambda p_{ia} + (1 - \lambda) q_a$ and $\lambda=0.5$. All the PSI-BLAST based features are listed as Features 39-45 in Table S2.

**LOMETS profile.** Multiple threading programs, LOMETS [7], are used to thread the wild-type and mutant sequences through a non-redundant PDB library to identify both homologous and analogous proteins. These sequences have often a low sequence identity to the query but usually adopt similar structural folds due to the integration of structural elements in the threading alignments. A threading based MSA is constructed by mapping the structural templates to the query based on the pair-wise threading alignments. The PSIC scores are calculated based on the Eq. (10), where these features are listed as Features 46-48 in Table S2.

**Pfam profile.** The Pfam database [8] contains a large collection of protein domain families, each represented by a hidden Markov model (HMM). The query sequence is matched by Pfam-scan through the Pfam-A library to identify homologous Pfam families. If a Pfam family is identified with the mutant positions included in the MSA, DAMpred will read the profile-HMM to obtain the match or insert emission probability of wild type and mutant amino acid types, based on the mutant position matching with the domain residue or ‘-’ shown in result file of Pfam-scan, respectively. These features are listed as Features 49-51 in Table S2.

**The contact environments in SPRING biological assembly**

DAMpred considers four types of the contact-environment based mutation features deduced from the complex structural models built by the dimeric threader, SPRING [14]; these include the number of intramolecular contacts (*Intra*), the number of intramolecular contacts involving functional residues (FunIntra), the number of intermolecular contacts (*Inter*), and the number of intermolecular contacts involving functional residues (FunInter). Figure S13 diagrammatized above contact groups.

The contacts are defined based on the structure in the target proteins, where two residues from the same chain (or from two BioUnit chains) are considered as in contact if the minimum distance of any heavy atoms from the two residues is below 5 (or 6) Å. To examine if the mutation involves a functional residue, the query sequence is searched by BLASTp through the UniProt human sequence database with one iteration and an E-value cutoff 0.001, where up to 5 homologous sequences with a sequence identity in [0.3, 0.8] are selected. A residue is defined as functional, if the aligned residue from any of the 5 homologies is annotated, based on the UniProt accession, as ACT_SITE, BINDING, CARBOHYD, CA_BIND, COILED, COMPIBIA, CROSSLINK, DISULFID, DNA_BIND, INIT_MET, INTRAMEM, LIPID, METAL, MOD_RES, MOTIF,
mutations for which we know the target label
feature
log
don't need to learn the weight of
only need to learn
where
the sigmoid nodes as the output of
network.

To determine the \( n+1 \) \( \alpha_l(P) \) functions in Eq. (3), Artificial Neural Network (ANN) is chosen, where the architecture of the special network structure is depicted in Fig. S10. This network has two hidden layers which consist of \( N_l \) sigmoid nodes (Layer #1) and \( n+1 \) linear nodes (Layer #2), respectively. A bias neuron is invented in the input of a network to increase the capacity of the network. Let \( x_{ji} \) be the \( i \)th input of the node \( j \), \( w_{ji} \) be the weight for \( i \)th input of the node \( j \), and \( net_j = \sum_i w_{ji}x_{ji} \) be the weighted sum of all inputs of the node \( j \). If one node in Layer-2 has an input from node \( j \) from Layer-1, we set these nodes from Layer-2 as Down\((j)\). The squashing function in the sigmoid nodes as the output of \( N_l \) can be written as

\[
\sigma_j = 1/(1 + e^{-\gamma_{j}(net_{j}-C_{j})})
\]

where \( \gamma_{j} \) and \( C_{j} \) are the slope and the inflection point, respectively. In this BANN structure, we only need to learn \( w_{ji}^1 \) between input and Layer-1 and \( w_{ji}^2 \) between Layer-1 and Layer-2; but we don’t need to learn the weight of \( p_l = \log P(f_l|C_{lp}) \) for \( S_D \) and \( S_N \) or \( p_l = \log P(f_l|C_{lp}) - \log (P(f_l|C_{lp})) \) for \( S_g \) between Layer-2 and the final output, which represents the possibility of the mutation feature \( f_l \) associated with the class \( C_{lp} \).

We choose back propagation training approach to build the network structure. If we have \( N_d \) mutations for which we know the target label \( t_d \) (disease-associated or neural), we should be able
to estimate the three sets of parameters \( \{w_{ij}\}, \{y_j\}, \{c_j\} \) by minimizing the training error of

\[
E(\vec{w}, \vec{y}, \vec{c}) = \frac{1}{2N_d} \sum_{d=1}^{N_d} (t_d - o_d)^2 + \mu \sum_{i,j} w_{ji}^2
\]  

(S5)

Here, \( o_d \) is the output of ANN which is the dependent variable and its corresponding independent variables are \( \vec{w}, \vec{y} \) and \( \vec{c} \). The penalty term \( (\mu \sum_{i,j} w_{ji}^2) \) is used to guide the gradient descent to search for smaller weight vectors, for decreasing the risk of over fitting, where \( \mu \) is the momentum.

The idea of stochastic gradient descent is to calculate the weight update according to the increment of each individual sample error. It starts with an arbitrary initial weight vector, and then modifies the vector at a small rate. At each step, the modified weight vector is generated in the steepest direction along the error surface, until the global minimum error point is obtained. To calculate the steepest direction along the error surface, this direction can be obtained by calculating the derivative of each component of \( E(\vec{w}, \vec{y}, \vec{c}) \). Here, we use an example to illustrate how to update \( \vec{w}, \vec{y} \) and \( \vec{c} \) with one sample \( (N_s = 1) \), which include three steps.

**Step-1: calculation of the weight of \( w_{ji}^2 \) and \( w_{ji}^1 \).** Because \( \vec{y} \) and \( \vec{c} \) are independent of weights \( \vec{w} \), so \( \vec{y} \) and \( \vec{c} \) can be considered as constant when error function is derivative of \( \vec{w} \). The updated weight is

\[
\vec{w} \leftarrow \vec{w} + \Delta \vec{w}
\]  

(S6)

where the gradient training rule is

\[
\begin{align*}
\Delta \vec{w} &= -\eta \nabla E(\vec{w}) \\
\nabla E(\vec{w}) &\equiv \left[ \frac{\partial E}{\partial w_{11}}, \frac{\partial E}{\partial w_{12}}, \ldots, \frac{\partial E}{\partial w_{N_L(n+1)}}, \frac{\partial E}{\partial w_{N_L(n+1)^2}}, \frac{\partial E}{\partial w_{N_L(n+1)^3}} \right]
\end{align*}
\]  

(S7)

Here, \( \eta \) is the learning rate. We suppose \( E'(\vec{w}, \vec{y}, \vec{c}) = \frac{1}{2} (t_d - o_d)^2 \), and have

\[
\begin{align*}
\Delta w_{ji} &= -\eta \frac{\partial E}{\partial w_{ji}} = -\eta \left( \frac{\partial E'}{\partial w_{ji}} + 2\mu w_{ji} \right) \\
\frac{\partial E'}{\partial w_{ji}} &= \frac{\partial E'}{\partial \text{net}_j} \times \frac{\partial \text{net}_j}{\partial w_{ji}} = \frac{\partial E'}{\partial \text{net}_j} \times x_{ji} \\
\end{align*}
\]  

(S8)

The remaining task is to deduce \( \partial E'/\partial \text{net}_j \). We can consider two situations in turn: the weight \( w_{ji}^2 \) for Layer-2 and the weight \( w_{ji}^1 \) for Layer-1.

In Layer 2, we mark \( \partial \text{net}_j \) as \( \partial \alpha_j \) as \( x_{ji}^1 \), and \( w_{ji} \) as \( w_{ji}^2 \). The \( \partial E'/\partial \text{net}_j \) can be obtained by

\[
\frac{\partial E'}{\partial \text{net}_j} = \frac{\partial E'}{\partial \alpha_j} = \frac{\partial E'}{\partial o_d} \times \frac{\partial o_d}{\partial \alpha_j}
\]  

(S9)

where

\[
\begin{align*}
\frac{\partial E'}{\partial o_d} &= \frac{\partial (\frac{1}{2} (t_d - o_d)^2)}{\partial o_d} = -(t_d - o_d) \\
\frac{\partial o_d}{\partial \alpha_j} &= \frac{\partial \text{net}_d}{\partial \alpha_j} = p_j
\end{align*}
\]  

(S10)

If we define \( \delta_{j,2} \) by

\[
\frac{\partial E'}{\partial \text{net}_j} = \frac{\partial E'}{\partial \alpha_j} = -p_j(t_d - o_d) = -\delta_{j,2}
\]  

(S11)
we will have
\[ \Delta w_{ji}^2 = -\eta \left(-p_j (t_d - o_d)x_{ji}^2 + 2\mu w_{ji}^2\right) = \eta \delta_{j,2} x_{ji}^2 - 2\eta \mu w_{ji}^2 \]  
(S12)
where \( \delta_{j,2} = p_j (t_d - o_d) \).

In layer 1, we mark \( x_{ji} \) as \( x_{ji}^1 \) and \( w_{ji} \) as \( w_{ji}^1 \). The \( \partial E'/\partial net_j \) can be obtained by
\[
\frac{\partial E'}{\partial net_j} = \sum_{k \in \text{Down}(j)} \frac{\partial E'}{\partial a_k} \frac{\partial a_k}{\partial net_j} = \sum_{k \in \text{Down}(j)} -\delta_{k,2} \frac{\partial a_k}{\partial net_j} \\
= \sum_{k \in \text{Down}(j)} -\delta_{k,2} \frac{\partial a_k}{\partial net_j} \frac{\partial \sigma_j}{\partial net_j} = \sum_{k \in \text{Down}(j)} -\delta_{k,2} w_{kj}^2 \frac{\partial \sigma_j}{\partial net_j} \\
= \sum_{k \in \text{Down}(j)} -\delta_{k,2} w_{kj}^2 \sigma_j (1 - \sigma_j) y_j 
\]
(S13)
If we define \( \delta_{j,1} \) by
\[
\frac{\partial E}{\partial net_j} = -\gamma_j \sigma_j (1 - \sigma_j) \sum_{k \in \text{Down}(j)} \delta_{k,2} w_{kj}^2 = -\gamma_j \delta_{j,1} 
\]
we will have
\[ \Delta w_{ji}^1 = \eta \gamma_j \delta_{j,1} x_{ji}^1 - 2\eta \mu w_{ji}^1 \]  
(S15)
where \( \delta_{j,1} = \sigma_j (1 - \sigma_j) \sum_{k \in \text{Down}(j)} \delta_{k,2} w_{kj}^2 \).

**Step-2: calculation of \( \Delta \gamma_j \) for layer 1.** Because weights \( \bar{w} \) and \( \bar{C} \) are independent of \( \gamma \), so \( \bar{w} \) and \( \bar{C} \) can be considered as constant when error function is derivative of \( \gamma \). So we can update \( \gamma \) by
\[ \gamma \leftarrow \gamma + \Delta \gamma \]  
(S16)
where the gradient training rule is
\[
\begin{bmatrix} \Delta \gamma \nabla E(\gamma) \end{bmatrix} = \begin{bmatrix} \frac{\partial E}{\partial \gamma_1}, \frac{\partial E}{\partial \gamma_2}, \ldots, \frac{\partial E}{\partial \gamma_N_L} \end{bmatrix} 
\]
(S17)
\( \Delta \gamma_j \) can be written as
\[
\begin{cases} 
\Delta \gamma_j = -\eta \frac{\partial E}{\partial \gamma_j} \\
\frac{\partial E}{\partial \gamma_j} = \sum_{k \in \text{Down}(j)} \frac{\partial E}{\partial a_k} \frac{\partial a_k}{\partial \gamma_j} = \sum_{k \in \text{Down}(j)} w_{kj}^2 \frac{\partial \sigma_j}{\partial \gamma_j} = \sum_{k \in \text{Down}(j)} \eta_{k,2} \frac{\partial \sigma_j}{\partial \gamma_j} (net_j - C_j) 
\end{cases} 
\]
(S18)
where \( \sum_{k \in \text{Down}(j)} \frac{\partial E}{\partial a_k} \frac{\partial a_k}{\partial \gamma_j} = \sum_{k \in \text{Down}(j)} w_{kj}^2 = -\sum_{k \in \text{Down}(j)} \eta_{k,2} \frac{\partial \sigma_j}{\partial \gamma_j} (net_j - C_j) \). Thus, we have
\[
\Delta y_j = -\eta \left( - \sum_{k \in \text{down}(j)} \eta_{k,2} \times w_{kj}^2 \sigma_j (1 - \sigma_j) (\text{net}_j - C_j) \right)
\]
\[
= \eta (\text{net}_j - C_j) \sigma_j (1 - \sigma_j) \sum_{k \in \text{down}(j)} \eta_{k,2} \times w_{kj}^2 = \eta (\text{net}_j - C_j) \delta_{j,1}
\]

**Step-3: calculation of \(\Delta C_j\) for layer 1.** Because weights \(\vec{w}\) and \(\vec{y}\) are independent of \(\vec{C}\), so \(\vec{w}\) and \(\vec{y}\) can be considered as constant when error function is derivative of \(\vec{C}\). So we can update \(\vec{C}\) by
\[
\vec{C} \leftarrow \vec{C} + \Delta \vec{C}
\]
where the gradient training rule is
\[
\begin{align*}
\begin{cases}
\Delta \vec{C} = -\eta \nabla E(\vec{C}) \\
\nabla E(\vec{C}) \equiv \left[ \frac{\partial E}{\partial \vec{C}_1}, \frac{\partial E}{\partial \vec{C}_2}, \ldots, \frac{\partial E}{\partial \vec{C}_{N_L}} \right]
\end{cases}
\end{align*}
\]
\(\Delta C_j\) can then be calculated by
\[
\Delta C_j = -\eta \frac{\partial E}{\partial C_j}
\]
\[
\frac{\partial E}{\partial C_j} = \frac{\partial E}{\partial \sigma_j} \times \frac{\partial \sigma_j}{\partial C_j} = \frac{\partial E}{\partial \sigma_j} \sigma_j (1 - \sigma_j) (-\gamma_j)
\]
\[
\frac{\partial E}{\partial \sigma_j} = - \sum_{k \in \text{down}(j)} \frac{\partial E}{\partial \alpha_k} \times \frac{\partial \alpha_k}{\partial \sigma_j} = \sum_{k \in \text{down}(j)} \frac{\partial E}{\partial \alpha_k} \times w_{kj}^2
\]
Thus, we have
\[
\Delta C_j = -\eta (-\gamma_j) \sigma_j (1 - \sigma_j) \left( - \sum_{k \in \text{down}(j)} \eta_{k,2} \times w_{kj}^2 \right) = -\eta \gamma_j \delta_{j,1}
\]

In summary, at each step of minimization, the gradient descent training rules for second layer and first layer is described by
\[
\begin{align*}
\begin{cases}
\Delta w_{ji}^2 = \eta \delta_{j,2} x_{ji}^2 - 2\eta \mu w_{ji}^2 \\
\Delta w_{ji}^1 = \eta \gamma \delta_{j,1} x_{ji}^1 - 2\eta \mu w_{ji}^1 \\
\Delta y_j = \eta \delta_{j,1} (\text{net}_j - C_j) \\
\Delta C_j = -\eta \delta_{j,1} \gamma_y
\end{cases}
\end{align*}
\]
where \(\eta\) is the learning rate and
\[
\begin{align*}
\begin{cases}
\delta_{j,2} = P_j (t_d - o_d) \\
\delta_{j,1} = \gamma_y \sigma_j (1 - \sigma_j) \sum_{k \in \text{down}(j)} \delta_{k,2} w_{kj}^2
\end{cases}
\end{align*}
\]

The initial values were tested to give a nice influence on the results of the parameter estimation. The inflection point \(C_j\) was initially set to 0.5, the slope \(\gamma_j\) to 1, and the learning rate \(\eta\) and momentum \(\mu\) to 0.001, respectively. To avoid the local optimization trap, we randomly divided the
training dataset into 10 parts and then updated the weights based on the minimization of the increment of the learning errors in a loop, in a similar way as in stochastic gradient descent.

Here, the main difference between the proposed BANN scheme and the traditional ANN is on the output layer. The weight of output layer from the traditional ANN needs to be learned based on training rules, but the one from BANN is calculated as the logarithm of the posterior probabilities of the features obtained from the disease-associated and neutral datasets. One advantage of the new training protocol is the integration of the inherent probability of the different features in the network, which can help avoid over-fitting and also increases the convergent speed and generalizing ability.

**Test S4. Method evaluation**

*Evaluation criterions.* The prediction results are mainly evaluated by the accuracy (ACC), Matthews correlation coefficient (MCC), sensitivity(±) and specificity(±):

\[
\begin{align*}
    \text{ACC} &= \frac{TP + TN}{TP + FN + TN + FP} \\
    \text{MCC} &= \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \\
    \text{Sensitivity}(+) &= \frac{TP}{TP + FN} \\
    \text{Sensitivity}(-) &= \frac{TN}{TN + FP} \\
    \text{Specificity}(+) &= \frac{TP}{TP + FP} \\
    \text{Specificity}(-) &= \frac{TN}{TN + FN}
\end{align*}
\]

(S26)

Here, TP (FP) is the number of correctly (incorrectly) identified disease-associated mutations and TN (FN) is the number of correctly (incorrectly) identified neutral mutations. While ACC is one of the most common evaluation criterions when assessing the performance of a predictor, the MCC considers both true and false positives and negatives which is generally regarded as a more balanced measure that can be used even if the classes are of very different sizes. Sensitivity(+) refers to the true positive rate (TPR) or recall, Specificity(+) refers to positive predictive value (PPV) or precision, Sensitivity(-) is the true negative rate (TNR), and Specificity(-) the negative predictive value (NPV). All these criterions are used to examine the performance of our method in different aspects.
Supporting Tables

**Table S1.** The distribution of proteins and mutations in the ten folds of D10634 for protein-level cross-validation experiment.

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>Subsets</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>#Dm(^a)</td>
<td>5355</td>
<td>536</td>
</tr>
<tr>
<td>#Nm(^b)</td>
<td>5279</td>
<td>525</td>
</tr>
<tr>
<td>#Tm(^c)</td>
<td>10634</td>
<td>1061</td>
</tr>
<tr>
<td>#Dp(^d)</td>
<td>617</td>
<td>70</td>
</tr>
<tr>
<td>#Np(^e)</td>
<td>1836</td>
<td>133</td>
</tr>
<tr>
<td>#Tp(^f)</td>
<td>2154</td>
<td>190</td>
</tr>
</tbody>
</table>

\(^a\)#Dm: the number of disease-associated mutations.  
\(^b\)#Nm: the number of neutral mutations  
\(^c\)#Tm: the total number of all mutations  
\(^d\)#Dp: the number of proteins with disease-associated mutations  
\(^e\)#Np: the number of proteins with neutral mutations  
\(^f\)#Tp: the total number of all proteins
<table>
<thead>
<tr>
<th>Feature Class</th>
<th>Feature</th>
<th>Mean</th>
<th>MCC</th>
<th>Cutoff Value</th>
<th>p-value in M-W test</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pharmaphore for the wild-type residues</td>
<td>HP&lt;sub&gt;W&lt;/sub&gt;</td>
<td>2.90</td>
<td>2.17</td>
<td>0.16</td>
<td>3.10E-078</td>
<td>Hydrophobic residue</td>
</tr>
<tr>
<td></td>
<td>noHP&lt;sub&gt;W&lt;/sub&gt;</td>
<td>4.47</td>
<td>3.89</td>
<td>0.14</td>
<td>1.49E-044</td>
<td>Non-hydrophobic residue</td>
</tr>
<tr>
<td></td>
<td>AR&lt;sub&gt;W&lt;/sub&gt;</td>
<td>1.28</td>
<td>0.92</td>
<td>0.14</td>
<td>1.31E-058</td>
<td>Aromatic rings</td>
</tr>
<tr>
<td></td>
<td>noAR&lt;sub&gt;W&lt;/sub&gt;</td>
<td>6.09</td>
<td>5.13</td>
<td>0.17</td>
<td>3.84E-093</td>
<td>Non-aromatic rings</td>
</tr>
<tr>
<td></td>
<td>PC&lt;sub&gt;W&lt;/sub&gt;</td>
<td>0.86</td>
<td>0.77</td>
<td>0.05</td>
<td>7.99E-006</td>
<td>Positive charge</td>
</tr>
<tr>
<td></td>
<td>NC&lt;sub&gt;W&lt;/sub&gt;</td>
<td>0.72</td>
<td>0.75</td>
<td>0.02</td>
<td>2.07E-002</td>
<td>Negative charge</td>
</tr>
<tr>
<td></td>
<td>noCAM&lt;sub&gt;W&lt;/sub&gt;</td>
<td>5.79</td>
<td>4.53</td>
<td>0.21</td>
<td>1.20E-132</td>
<td>Neutral charge</td>
</tr>
<tr>
<td></td>
<td>BP&lt;sub&gt;W&lt;/sub&gt;</td>
<td>1.67</td>
<td>1.26</td>
<td>0.09</td>
<td>1.38E-013</td>
<td>Both wild-type and neighbor AA are polar</td>
</tr>
<tr>
<td></td>
<td>OP&lt;sub&gt;W&lt;/sub&gt;</td>
<td>2.83</td>
<td>2.21</td>
<td>0.14</td>
<td>1.33E-065</td>
<td>Either of wild-type and neighbor AA is polar</td>
</tr>
<tr>
<td></td>
<td>NP&lt;sub&gt;W&lt;/sub&gt;</td>
<td>1.87</td>
<td>1.59</td>
<td>0.10</td>
<td>3.63E-006</td>
<td>Both wild-type and neighbor are nonpolar</td>
</tr>
<tr>
<td></td>
<td>AC&lt;sub&gt;W&lt;/sub&gt;</td>
<td>9.12</td>
<td>8.05</td>
<td>0.12</td>
<td>1.19E-021</td>
<td>The count of reside being the hydrogen acceptor</td>
</tr>
<tr>
<td></td>
<td>DO&lt;sub&gt;W&lt;/sub&gt;</td>
<td>5.59</td>
<td>4.83</td>
<td>0.14</td>
<td>1.94E-032</td>
<td>The count of reside being the hydrogen donor</td>
</tr>
<tr>
<td>Pharmaphore for the mutant residues</td>
<td>HP&lt;sub&gt;M&lt;/sub&gt;</td>
<td>3.04</td>
<td>2.21</td>
<td>0.18</td>
<td>2.40E-098</td>
<td>Hydrophobic</td>
</tr>
<tr>
<td></td>
<td>noHP&lt;sub&gt;M&lt;/sub&gt;</td>
<td>4.52</td>
<td>3.72</td>
<td>0.17</td>
<td>1.68E-081</td>
<td>Non-hydrophobic</td>
</tr>
<tr>
<td></td>
<td>AR&lt;sub&gt;M&lt;/sub&gt;</td>
<td>1.33</td>
<td>0.95</td>
<td>0.15</td>
<td>3.22E-063</td>
<td>Aromatic rings</td>
</tr>
<tr>
<td></td>
<td>noAR&lt;sub&gt;M&lt;/sub&gt;</td>
<td>6.23</td>
<td>4.98</td>
<td>0.21</td>
<td>4.00E-144</td>
<td>Non-aromatic rings</td>
</tr>
<tr>
<td></td>
<td>PC&lt;sub&gt;M&lt;/sub&gt;</td>
<td>0.83</td>
<td>0.68</td>
<td>0.07</td>
<td>3.82E-015</td>
<td>Positive charge</td>
</tr>
<tr>
<td></td>
<td>NC&lt;sub&gt;M&lt;/sub&gt;</td>
<td>0.69</td>
<td>0.69</td>
<td>0.03</td>
<td>1.45E-001</td>
<td>Negative charge</td>
</tr>
<tr>
<td></td>
<td>noCAM&lt;sub&gt;M&lt;/sub&gt;</td>
<td>6.04</td>
<td>4.55</td>
<td>0.22</td>
<td>3.60E-161</td>
<td>Neutral charge</td>
</tr>
<tr>
<td></td>
<td>BP&lt;sub&gt;M&lt;/sub&gt;</td>
<td>1.38</td>
<td>1.21</td>
<td>0.05</td>
<td>3.57E-002</td>
<td>Both wild-type and neighbor AA are polar</td>
</tr>
<tr>
<td></td>
<td>OP&lt;sub&gt;M&lt;/sub&gt;</td>
<td>3.22</td>
<td>2.25</td>
<td>0.19</td>
<td>1.20E-121</td>
<td>Either of wild-type and neighbor AA is polar</td>
</tr>
<tr>
<td></td>
<td>NP&lt;sub&gt;M&lt;/sub&gt;</td>
<td>1.95</td>
<td>1.47</td>
<td>0.13</td>
<td>4.82E-021</td>
<td>Both wild-type and neighbor are nonpolar</td>
</tr>
<tr>
<td></td>
<td>AC&lt;sub&gt;M&lt;/sub&gt;</td>
<td>9.31</td>
<td>7.84</td>
<td>0.15</td>
<td>1.18E-035</td>
<td>The count of reside being the hydrogen acceptor</td>
</tr>
<tr>
<td></td>
<td>DO&lt;sub&gt;M&lt;/sub&gt;</td>
<td>5.69</td>
<td>4.75</td>
<td>0.16</td>
<td>9.94E-042</td>
<td>The count of reside being the hydrogen donor</td>
</tr>
<tr>
<td>Mutation-induced environmental pharmaphore changes</td>
<td>cos&lt;sub&gt;NNWM&lt;/sub&gt;</td>
<td>0.74</td>
<td>0.82</td>
<td>0.22</td>
<td>2.00E-133</td>
<td>cos for the pharmacophores of wild-type and mutant residues</td>
</tr>
<tr>
<td></td>
<td>rms&lt;sub&gt;NNWM&lt;/sub&gt;</td>
<td>0.47</td>
<td>0.37</td>
<td>0.22</td>
<td>1.60E-125</td>
<td>RMSD for pharmacophores of wild-type and mutant residues</td>
</tr>
<tr>
<td></td>
<td>cos&lt;sub&gt;NNNM&lt;/sub&gt;</td>
<td>0.94</td>
<td>0.95</td>
<td>1.00</td>
<td>7.55E-019</td>
<td>cos for neighbor pharmacophores of wild-type and mutant</td>
</tr>
<tr>
<td></td>
<td>rms&lt;sub&gt;NNNM&lt;/sub&gt;</td>
<td>2.09</td>
<td>1.52</td>
<td>1.72</td>
<td>5.10E-154</td>
<td>RMSD for neighbor pharmacophores of wild-type and mutant</td>
</tr>
<tr>
<td></td>
<td>cos&lt;sub&gt;NNWM&lt;/sub&gt;</td>
<td>0.92</td>
<td>0.93</td>
<td>1.00</td>
<td>2.02E-022</td>
<td>cos for neighbor pharmacophores of wild-type and mutant residues</td>
</tr>
<tr>
<td></td>
<td>rms&lt;sub&gt;NNWM&lt;/sub&gt;</td>
<td>1.77</td>
<td>1.24</td>
<td>0.24</td>
<td>1.30E-176</td>
<td>RMSD for neighbor pharmacophores of wild-type and mutant residues</td>
</tr>
<tr>
<td></td>
<td>cos&lt;sub&gt;NNP&lt;/sub&gt;</td>
<td>0.92</td>
<td>0.95</td>
<td>0.06</td>
<td>2.53E-003</td>
<td>cos for neighbor pharmacophores of wild-type and mutant residues</td>
</tr>
<tr>
<td></td>
<td>rms&lt;sub&gt;NNP&lt;/sub&gt;</td>
<td>2.87</td>
<td>2.21</td>
<td>0.24</td>
<td>3.02E-042</td>
<td>RMSD for neighbor pharmacophores of wild-type and mutant residues</td>
</tr>
<tr>
<td>Other physicochemical properties</td>
<td>Vol&lt;sub&gt;W&lt;/sub&gt;</td>
<td>2.83</td>
<td>2.86</td>
<td>1.9</td>
<td>9.84E-002</td>
<td>The volume of wild-type residue</td>
</tr>
<tr>
<td></td>
<td>Vol&lt;sub&gt;M&lt;/sub&gt;</td>
<td>2.91</td>
<td>2.88</td>
<td>3.2</td>
<td>3.86E-008</td>
<td>The volume of mutant residue</td>
</tr>
<tr>
<td></td>
<td>dVol&lt;sub&gt;I&lt;/sub&gt;</td>
<td>0.08</td>
<td>0.02</td>
<td>0.7</td>
<td>1.23E-008</td>
<td>The volume difference</td>
</tr>
<tr>
<td></td>
<td>W&lt;sub&gt;W&lt;/sub&gt;</td>
<td>132.0</td>
<td>130.81</td>
<td>75</td>
<td>5.92E-003</td>
<td>The weight of wild-type residue</td>
</tr>
<tr>
<td></td>
<td>W&lt;sub&gt;M&lt;/sub&gt;</td>
<td>136.2</td>
<td>131.57</td>
<td>165</td>
<td>5.60E-012</td>
<td>The weight of mutant residue</td>
</tr>
<tr>
<td></td>
<td>dW&lt;sub&gt;I&lt;/sub&gt;</td>
<td>4.23</td>
<td>0.76</td>
<td>42</td>
<td>5.46E-006</td>
<td>The molecular weight difference</td>
</tr>
<tr>
<td>PSI-BLAST profile scores</td>
<td>PSIC&lt;sub&gt;W&lt;/sub&gt;</td>
<td>1.57</td>
<td>0.91</td>
<td>1.2</td>
<td>0.00E-000</td>
<td>The PSIC score for wild-type residue</td>
</tr>
<tr>
<td></td>
<td>PSIC&lt;sub&gt;M&lt;/sub&gt;</td>
<td>-0.42</td>
<td>0.24</td>
<td>-0.1</td>
<td>0.00E-000</td>
<td>The PSIC score for mutant residue</td>
</tr>
<tr>
<td></td>
<td>dPSIC&lt;sub&gt;i&lt;/sub&gt;</td>
<td>-1.99</td>
<td>-0.66</td>
<td>-1.1</td>
<td>0.00E-000</td>
<td>The PSIC score difference</td>
</tr>
<tr>
<td></td>
<td>JSD&lt;sub&gt;W&lt;/sub&gt;</td>
<td>0.03</td>
<td>0.03</td>
<td>0.04</td>
<td>3.70E-001</td>
<td>The JSD score for wild-type residue</td>
</tr>
<tr>
<td></td>
<td>JSD&lt;sub&gt;M&lt;/sub&gt;</td>
<td>0.02</td>
<td>0.03</td>
<td>0.03</td>
<td>1.22E-016</td>
<td>The JSD score for mutant residue</td>
</tr>
<tr>
<td></td>
<td>dJSD&lt;sub&gt;i&lt;/sub&gt;</td>
<td>0.00</td>
<td>0.00</td>
<td>-0.01</td>
<td>1.99E-012</td>
<td>The JSD score difference</td>
</tr>
<tr>
<td></td>
<td>JSD&lt;sub&gt;i&lt;/sub&gt;</td>
<td>0.47</td>
<td>0.32</td>
<td>0.5</td>
<td>2.20E-273</td>
<td>The JSD score at mutant position i</td>
</tr>
<tr>
<td>LOMETS profile scores</td>
<td>IPSIC&lt;sub&gt;W&lt;/sub&gt;</td>
<td>0.78</td>
<td>0.45</td>
<td>0.8</td>
<td>1.00E-149</td>
<td>The PSIC score for wild-type residue</td>
</tr>
<tr>
<td></td>
<td>IPSIC&lt;sub&gt;M&lt;/sub&gt;</td>
<td>-0.30</td>
<td>-0.02</td>
<td>0.04</td>
<td>1.50E-108</td>
<td>The PSIC score for mutant residue</td>
</tr>
<tr>
<td></td>
<td>dIPSIC&lt;sub&gt;i&lt;/sub&gt;</td>
<td>-1.08</td>
<td>-0.47</td>
<td>-0.7</td>
<td>8.80E-243</td>
<td>The PSIC score difference</td>
</tr>
<tr>
<td>Pfam profile scores</td>
<td>Pfam&lt;sub&gt;W&lt;/sub&gt;</td>
<td>1.83</td>
<td>2.40</td>
<td>1.5</td>
<td>9.40E-178</td>
<td>The Pfam score for wild-type residue</td>
</tr>
<tr>
<td></td>
<td>Pfam&lt;sub&gt;M&lt;/sub&gt;</td>
<td>3.66</td>
<td>3.00</td>
<td>3.0</td>
<td>7.20E-119</td>
<td>The Pfam score for mutant residue</td>
</tr>
</tbody>
</table>
Table S3. Comparison of different machine-learning methods used for training the classification model of disease and neutral mutations. The data are generated by the protein-level 10-fold cross validation on the D10634 dataset. Bold fonts highlight the best predictor in each category.

<table>
<thead>
<tr>
<th>Methods(^a)</th>
<th>GBC</th>
<th>KNC</th>
<th>SVC</th>
<th>ANN</th>
<th>BANN</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Model trained by top 20 features</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MCC</td>
<td>0.559</td>
<td>0.552</td>
<td>0.561</td>
<td>0.575</td>
<td><strong>0.589</strong></td>
</tr>
<tr>
<td>ACC</td>
<td>0.779</td>
<td>0.776</td>
<td>0.780</td>
<td>0.787</td>
<td><strong>0.795</strong></td>
</tr>
<tr>
<td>SEN(^+)</td>
<td>0.784</td>
<td>0.756</td>
<td>0.779</td>
<td>0.795</td>
<td><strong>0.817</strong></td>
</tr>
<tr>
<td>SPE(^+)</td>
<td>0.779</td>
<td><strong>0.789</strong></td>
<td>0.784</td>
<td>0.785</td>
<td>0.784</td>
</tr>
<tr>
<td>SEN(^-)</td>
<td>0.775</td>
<td><strong>0.795</strong></td>
<td>0.782</td>
<td>0.780</td>
<td>0.772</td>
</tr>
<tr>
<td>SPE(^-)</td>
<td>0.779</td>
<td>0.763</td>
<td>0.777</td>
<td>0.789</td>
<td><strong>0.806</strong></td>
</tr>
<tr>
<td><strong>p-value(^c)</strong></td>
<td>2.60E-14</td>
<td>2.80E-49</td>
<td>7.22E-29</td>
<td>2.25E-10</td>
<td>*7.83E-6</td>
</tr>
<tr>
<td><strong>Model trained by all 70 features</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MCC</td>
<td>0.566</td>
<td>0.504</td>
<td>0.570</td>
<td>0.580</td>
<td><strong>0.601</strong></td>
</tr>
<tr>
<td>ACC</td>
<td>0.783</td>
<td>0.748</td>
<td>0.785</td>
<td>0.790</td>
<td><strong>0.800</strong></td>
</tr>
<tr>
<td>SEN(^+)</td>
<td>0.783</td>
<td>0.669</td>
<td>0.777</td>
<td>0.795</td>
<td><strong>0.812</strong></td>
</tr>
<tr>
<td>SPE(^+)</td>
<td>0.786</td>
<td>0.799</td>
<td>0.782</td>
<td>0.790</td>
<td><strong>0.796</strong></td>
</tr>
<tr>
<td>SEN(^-)</td>
<td>0.783</td>
<td><strong>0.830</strong></td>
<td>0.793</td>
<td>0.785</td>
<td>0.788</td>
</tr>
<tr>
<td>SPE(^-)</td>
<td>0.781</td>
<td>0.716</td>
<td>0.778</td>
<td>0.791</td>
<td><strong>0.805</strong></td>
</tr>
<tr>
<td><strong>p-value(^c)</strong></td>
<td>5.39E-6</td>
<td>1.85E-138</td>
<td>1.47E-16</td>
<td>3.99E-3</td>
<td></td>
</tr>
</tbody>
</table>

\(^a\)Training methods: GBC: gradient boosting classifier; KNC: k-nearest neighbor classifier; SVC: support vector classifier; ANN: artificial neural network; BANN: Bayes-classifier guided ANN.

\(^b\)Models are trained on the top-20 features ranked by the p-values in the Table S2.

\(^c\)p-value in McNemar’s Test is calculated for each comparison related to corresponding BANN, where ‘*’ denote the p-value between BANN with top 20 features and BANN with all 70 features.
Supporting Figures

**Figure S1.** Pipeline for data mapping and dataset construction.

**Figure S2.** Disease-association mutation prediction on TP53 protein. (A) A Venn diagram showing the distribution of different mutations on the p53 protein. (B) I-TASSER model for the isoform P04637-1 of TP53 gene, where the transactivation, domain and COOH-terminal domains are marked in different colors. The residues in red are disease-associated mutations and those in blue are neutral mutations. The inset is a superposition of the I-TASSER model with the X-ray structure (PDBID: 1tsrA) in the core domain with a RMSD=1.07 Å, where homologous templates including 1tsrA has been excluded in the I-TASSER modeling simulations.
Figure S3. Quality of the protein structure predictions by I-TASSER and SPRING. (A, B) Histogram distribution of TM-score and RMSD of the I-TASSER models on 1974 proteins from the D10634 dataset. (C, D) Histogram distribution of TM-score and RMSD of the SPRING models on 2116 BioUnit complexes from the D10634 dataset.
**Figure S4.** Histogram distribution of evolution features on disease-associated and neutral mutations from the D10634 dataset. The left-most bar laying in Pfam diagrams with feature score < 0 indicate the mutant positions not to be found in the Pfam families.
**Figure S5.** Histogram distribution of Contact features on SPRING biological assembly on disease-associated and neutral mutations from the D10634 dataset.

**Figure S6.** Histogram distribution of the I-TASSER model based features on disease-associated and neutral mutations from the D10634 dataset.
Figure S7. Histogram distribution of pharmacophore features for wild-type residue on disease-associated and neutral mutations from the D10634 dataset.

Figure S8. Histogram distribution of pharmacophore features for mutant residue on disease-associated and neutral mutations from the D10634 dataset.
Figure S9. Histogram distribution of pharmacophore difference features on disease-associated and neutral mutations from the D10634 dataset.

Figure S10. Histogram distribution of chemical features on disease-associated and neutral mutations from the D10634 dataset.
Figure S11. Illustration of the environmental pharmacophore for physicochemical properties.

Figure S12. Illustration of contact environments in the SPRING biological assembly.
Figure S13. Illustration of the Bayes-guided Artificial Neural Network (BANN) learning model.
References cited in Supporting Information:


