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 pharmcophore 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 that used in mCSM.

Mutation-induced environmental pharmacophore changes. Fourteen environmental pharmacophore features are considered (see Figure S11). Here, the pharmacophore vector for *i*th residue of the query protein is written as $\overrightarrow{p_{ic}} = [p_{i1}, p_{i2}, \cdots, p_{il}]$, where l is the number of pharmacophore types and $c \epsilon [m, w]$, with m and w indicating mutant and wild structures respectively. The environmental pharmacophore changes due to the *i*th mutant can be calculated by $cos_{wm}(i) = cosin(\overrightarrow{p_{iw}}, \overrightarrow{p_{im}})$ and $rms_{wm}(i) = rmsd(\overrightarrow{p_{iw}}, \overrightarrow{p_{im}})$. The neighbor environment pharmacophore counts for the interactions of the target residue and all residues in contact (n_i^c), i.e., $\overrightarrow{pn_{ic}} = \sum_{k=1}^{n_i^c} \overrightarrow{p_{kw}}$. Thus, the neighbor environmental pharmacophore changes due to the *i*th mutation can be written as $cosN_{wm}(i) = cosin(\overrightarrow{p_{iw}}, \overrightarrow{p_{im}})$ and $rmsN_{wm}(i) = rmsd(\overrightarrow{pn_{iw}}, \overrightarrow{pn_{im}})$. 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 ($l=l_s+l_p$), 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 $\overrightarrow{pn_{ic\_ls}}$ (or $\overrightarrow{pn_{ic\_lp}}$), where the corresponding pharmacophore changes are then calculated by

$$\begin{cases} \cos\left(\overrightarrow{P_w}, \overrightarrow{P_m}\right) = \dfrac{\overrightarrow{P_w} \cdot \overrightarrow{P_m}}{\left\|\overrightarrow{P_w}\right\| \cdot \left\|\overrightarrow{P_m}\right\|} \\ rmsd\left(\overrightarrow{P_w}, \overrightarrow{P_m}\right) = \sqrt{\dfrac{\sum_{\mathrm{k}=1}^{\mathrm{l}} \left(P_w^{\mathrm{k}} - P_m^{\mathrm{k}}\right)^2}{L}} \end{cases} \quad (S1)$$

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

$$\begin{cases} PSIC(a,i) = ln\left[\frac{p_{ia}}{q_a}\right] \\ p_{ia} = \frac{n(a,i)_{eff}}{\sum_b n(b,i)_{eff}} \end{cases} \quad (S2)$$

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 statistical 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]

$$\begin{cases} \mathrm{JSD}_{ia} = \lambda p_{ia} log\frac{p_{ia}}{c_{ia}} + (1-\lambda) q_a log\frac{q_a}{c_{ia}} \\ \mathrm{JSD}_i = \sum_{a \in AA} JSD_{ia} \end{cases} \quad (S3)$$

where $c_{ia} = \lambda p_{ia} + (1-\lambda) q_a$ and λ=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, COMPBIAS, CROSSLNK, DISULFID, DNA_BIND, INIT_MET, INTRAMEM, LIPID, METAL, MOD_RES, MOTIF,

NON_STD, NP_BIND, PEPTIDE, PROPEP, REGION, PEPEAT, SIGNAL, SITE, TRANSIT, TRANSMEM, or ZN_FING. A contact is considered as functional, if any of the interacting residues is functional.

In addition, DAMpred considers an enlarged contact environment by counting the residues that are in contact with the contacting partners of the mutated residues, called indirect contacts. Similarly, if any of the indirect contact partners are functional, the mutation may result in an indirect effect on the residues and the contact is therefore counted as an indirectly functional contact. This group of features is listed as Feature 52-59 in Table S2.

I-TASSER modeling and structure-based feature extraction.

I-TASSER [15, 16] was used to construct 3D models for both wild-type and mutant sequences, where two groups of structure-based features, on protein surface and physics-based energy terms, are extracted from the I-TASSER models.

Surface-based features. This feature group describes the properties of target residues involved in protein surfaces. These include the likelihood score of the wild-type residue to bind with ligands, which is calculated by ConCavity [13] based on the surface geometry of the I-TASSER models. In addition, we count the depth score as the distance of atoms or residue to its closest molecule of bulk solvent, which is calculated by the Depth program [17].

Physics-based features. The second structure-based feature group is physics-based energy potentials evaluating the atomic interactions based on the I-TASSER models. Three kinds of energy terms were considered. The first is from EvoDesign [18], which combines the log-odds profile of the analogous structures of the I-TASSER models searched from the PDB and the fitness score of the mutant residues on the I-TASSER models that are built on secondary structure, backbone torsion angle, and solvent accessibility predictions. The second term is the free-energy changes ($\Delta\Delta G$) induced by the mutations, calculated by the SEF program [19], which counts for the probability distributions of rotamer types with specific secondary structure, solvent accessibility, and backbone Ranmachandran torsional angles. Here, if the SEF score is lower than zero, it means that this mutation increased the stability of this protein. The third term is the preferences of the side-chain conformers calculated by CIS-RR [20]. Here, the structural model for the mutant sequence is first reconstructed by CIS-RR from the wild-type I-TASSER models, while the empirical van der Waals potential and the rotamer preference score are then calculated for both wild-type and mutant structure.

Text S3. Derivation of the Bayes-guided artificial neural network (BANN) model

To determine the $n+1$ $\alpha_i(\mathrm{F})$ 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_i^l can be written as

$$\sigma_j = 1/\left(1 + e^{-\gamma_j(net_j - C_j)}\right) \qquad (S4)$$

where γ_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 $p_i = \log P(f_i|C_k)$ for S_D and S_N or $p_i = \log(P(f_i|C_D) - \log(P(f_i|C_N)$ for S_Δ between Layer-2 and the final output, which represents the possibility of the mutation feature f_i associated with the class C_k.

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 ($\vec{w} = \{w_{ij}\}, \vec{\gamma} = \{\gamma_j\}, \vec{C} = \{C_j\}$) by minimizing the training error of

$$E(\vec{w}, \vec{\gamma}, \vec{C}) \equiv \frac{1}{2N_d} \sum_{d=1}^{N_d} (t_d - o_d)^2 + \mu \sum_{i,j} {w_{ji}}^2 \quad (S5)$$

Here, o_d is the output of ANN which is the dependent variable and its corresponding independent variables are $\vec{w}$, $\vec{\gamma}$ 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 μ 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 $E(\vec{w}, \vec{\gamma}, \vec{C})$. Here, we use an example to illustrate how to update $\vec{w}$, $\vec{\gamma}$ 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 $\vec{\gamma}$ and $\vec{C}$ are independent of weights $\vec{w}$, so $\vec{\gamma}$ and $\vec{C}$ can be considered as constant when error function is derivative of $\vec{w}$. The updated weight is

$$\vec{w} \leftarrow \vec{w} + \Delta\vec{w} \quad (S6)$$

where the gradient training rule is

$$\begin{cases} \Delta\vec{w} = -\eta \nabla E(\vec{w}) \\ \nabla E(\vec{w}) \equiv \left[\frac{\partial E}{\partial {w_{11}}^1}, \frac{\partial E}{\partial {w_{12}}^1}, \cdots, \frac{\partial E}{\partial {w_{N_L(n+1)}}^1}, \frac{\partial E}{\partial {w_{11}}^2}, \cdots, \frac{\partial E}{\partial {w_{N_L(n+1)}}^2} \right] \end{cases} \quad (S7)$$

Here, η is the learning rate. We suppose $\mathrm{E}'(\vec{w}, \vec{\gamma}, \vec{C}) \equiv \frac{1}{2}(t_d - o_d)^2$, and have

$$\begin{cases} \Delta w_{ji} = -\eta \dfrac{\partial E}{\partial w_{ji}} = -\eta \left(\dfrac{\partial E'}{\partial w_{ji}} + 2\mu w_{ji} \right) \\ \dfrac{\partial E'}{\partial w_{ji}} = \dfrac{\partial E'}{\partial net_j} \times \dfrac{\partial net_j}{\partial w_{ji}} = \dfrac{\partial E'}{\partial net_j} \times x_{ji} \end{cases} \quad (S8)$$

The remaining task is to deduce $\partial E'/\partial net_j$,. 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 ∂net_j as $\partial \alpha_j$, x_{ji} as ${x_{ji}}^2$, and w_{ji} as ${w_{ji}}^2$. The $\partial E'/\partial net_j$ can be obtained by

$$\frac{\partial E'}{\partial net_j} = \frac{\partial E'}{\partial \alpha_j} = \frac{\partial E'}{\partial o_d} \times \frac{\partial o_d}{\partial \alpha_j} \quad (S9)$$

where

$$\begin{cases} \dfrac{\partial E'}{\partial o_d} = \dfrac{\partial (\frac{1}{2}(t_d - o_d)^2)}{\partial o_d} = -(t_d - o_d) \\ \dfrac{\partial o_d}{\partial \alpha_j} = \dfrac{\partial net_d}{\partial \alpha_j} = p_j \end{cases} \quad (S10)$$

If we define $\delta_{j,2}$ by

$$\frac{\partial E'}{\partial net_j} = \frac{\partial E'}{\partial \alpha_j} = -p_j(t_d - o_d) = -\delta_{j,2} \quad (S11)$$

we will have

$$\Delta w_{ji}{}^2 = -\eta\left(-p_j(t_d - o_d)x_{ji}{}^2 + 2\mu w_{ji}{}^2\right) = \eta\delta_{j,2}x_{ji}{}^2 - 2\eta\mu w_{ji}{}^2 \tag{S12}$$

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

$$\begin{aligned}
\frac{\partial E'}{\partial net_j} &= \sum_{k\in Down(j)} \frac{\partial E'}{\partial \alpha_k} \times \frac{\partial \alpha_k}{\partial net_j} = \sum_{k\in Down(j)} -\delta_{k,2}\frac{\partial \alpha_k}{\partial net_j} \\
&= \sum_{k\in Down(j)} -\delta_{k,2}\frac{\partial \alpha_k}{\partial \sigma_j} \times \frac{\partial \sigma_j}{\partial net_j} = \sum_{k\in Down(j)} -\delta_{k,2}w_{kj}{}^2 \times \frac{\partial \sigma_j}{\partial net_j} \\
&= \sum_{k\in Down(j)} -\delta_{k,2}w_{kj}{}^2 \times \frac{\partial \frac{1}{1+e^{-\gamma_j(net_j - C_j)}}}{\partial net_j} \\
&= \sum_{k\in Down(j)} -\delta_{k,2}w_{kj}{}^2\sigma_j(1-\sigma_j)\gamma_j
\end{aligned} \tag{S13}$$

If we define $\delta_{j,1}$ by

$$\frac{\partial E'}{\partial net_j} = -\gamma_j\sigma_j(1-\sigma_j)\sum_{k\in Down(j)} \delta_{k,2}w_{kj}{}^2 = -\gamma_j\delta_{j,1} \tag{S14}$$

we will have

$$\Delta w_{ji}{}^1 = \eta\gamma_j\delta_{j,1}x_{ji}{}^1 - 2\eta\mu w_{ji}{}^1 \tag{S15}$$

where $\delta_{j,1} = \sigma_j(1-\sigma_j)\sum_{k\in Down(j)}\delta_{k,2}w_{kj}{}^2$.

Step-2: calculation of $\Delta\gamma_j$ for layer 1. Because weights $\vec{w}$ and $\vec{C}$ are independent of $\vec{\gamma}$, so $\vec{w}$ and $\vec{C}$ can be considered as constant when error function is derivative of $\vec{\gamma}$. So we can update $\vec{\gamma}$ by

$$\vec{\gamma} \leftarrow \vec{\gamma} + \Delta\vec{\gamma} \tag{S16}$$

where the gradient training rule is

$$\begin{cases} \Delta\vec{\gamma} = -\eta\nabla E(\vec{\gamma}) \\ \nabla E(\vec{\gamma}) \equiv \left[\frac{\partial E}{\partial \gamma_1}, \frac{\partial E}{\partial \gamma_1}, \cdots, \frac{\partial E}{\partial \gamma_{N_L}}\right] \end{cases} \tag{S17}$$

$\Delta\gamma_j$ can be written as

$$\begin{cases} \Delta\gamma_j = -\eta\dfrac{\partial E}{\partial \gamma_j} \\ \dfrac{\partial E}{\partial \gamma_j} = \dfrac{\partial E}{\partial \sigma_j} \times \dfrac{\partial \sigma_j}{\partial \gamma_j} = \dfrac{\partial E}{\partial \sigma_j} \times \sigma_j(1-\sigma_j)(net_j - C_j) \end{cases} \tag{S18}$$

where $\frac{\partial E}{\partial \sigma_j} = \sum_{k\in Down(j)}\frac{\partial E}{\partial \alpha_k} \times \frac{\partial \alpha_k}{\partial \sigma_j} = \sum_{k\in Down(j)}\frac{\partial E}{\partial \alpha_k} \times w_{kj}{}^2 = -\sum_{k\in Down(j)}\eta_{k,2} \times w_{kj}{}^2$. Thus, we have

$$\Delta\gamma_j = -\eta\left((-\sum_{k\in Down(j)} \eta_{k,2}\times {w_{kj}}^2)\,\sigma_j(1-\sigma_j)(net_j - C_j)\right)$$
$$= \eta(net_j - C_j)\sigma_j(1-\sigma_j)\sum_{k\in Down(j)} \eta_{k,2}\times {w_{kj}}^2 = \eta(net_j - C_j)\,\delta_{j,1} \quad (S19)$$

Step-3: calculation of ΔC_j for layer 1. Because weights $\vec{w}$ and $\vec{\gamma}$ are independent of $\vec{C}$, so $\vec{w}$ and $\vec{\gamma}$ can be considered as constant when error function is derivative of $\vec{C}$. So we can update $\vec{C}$ by

$$\vec{C} \leftarrow \vec{C} + \Delta\vec{C} \quad (S20)$$

where the gradient training rule is

$$\begin{cases}\Delta\vec{C} = -\eta\nabla E(\vec{C}) \\ \nabla E(\vec{C}) \equiv \left[\frac{\partial E}{\partial C_1}, \frac{\partial E}{\partial C_1}, \cdots, \frac{\partial E}{\partial C_{N_L}}\right]\end{cases} \quad (S21)$$

ΔC_j can then be calculated by

$$\begin{aligned}\Delta C_j &= -\eta\frac{\partial E}{\partial C_j} \\ \frac{\partial E}{\partial C_j} &= \frac{\partial E}{\partial \sigma_j}\times\frac{\partial \sigma_j}{\partial C_j} = \frac{\partial E}{\partial \sigma_j}\sigma_j(1-\sigma_j)(-\gamma_j) \\ \frac{\partial E}{\partial \sigma_j} &= \sum_{k\in Down(j)}\frac{\partial E}{\partial \alpha_k}\times\frac{\partial \alpha_k}{\partial \sigma_j} = \sum_{k\in Down(j)}\frac{\partial E}{\partial \alpha_k}\times {w_{kj}}^2 \\ &= -\sum_{k\in Down(j)} \eta_{k,2}\times {w_{kj}}^2\end{aligned} \quad (S22)$$

Thus, we have

$$\Delta C_j = -\eta(-\gamma_j)\sigma_j(1-\sigma_j)\left(-\sum_{k\in Down(j)} \eta_{k,2}\times w_{kj}\right) = -\eta\gamma_j\delta_{j,1} \quad (S23)$$

In summary, at each step of minimization, the gradient descent training rules for second layer and first layer is described by

$$\begin{cases}\Delta {w_{ji}}^2 = \eta\delta_{j,2}{x_{ji}}^2 - 2\eta\mu{w_{ji}}^2 \\ \Delta {w_{ji}}^1 = \eta\gamma_j\delta_{j,1}{x_{ji}}^1 - 2\eta\mu{w_{ji}}^1 \\ \Delta\gamma_j = \eta\delta_{j,1}(net_j - C_j) \\ \Delta C_j = -\eta\delta_{j,1}\gamma_j\end{cases} \quad (S24)$$

where η is the learning rate and

$$\begin{cases}\delta_{j,2} = P_j(t_d - o_d) \\ \delta_{j,1} = \gamma_j\sigma_j(1-\sigma_j)\sum_{k\in Down(j)}\delta_{k,2}{w_{kj}}^2\end{cases} \quad (S25)$$

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 γ_j to 1, and the learning rate η and momentum μ 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(±):

$$\begin{cases} ACC = \dfrac{TP + TN}{TP + FN + TN + FP} \\ MCC = \dfrac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \\ Sensitivity(+) = \dfrac{TP}{TP + FN} \\ Sensitivity(-) = \dfrac{TN}{TN + FP} \\ Specificity(+) = \dfrac{TP}{TP + FP} \\ Specificity(-) = \dfrac{TN}{TN + FN} \end{cases} \quad (S26)$$

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 (PPC) 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.

	Total	Subsets									
		1	2	3	4	5	6	7	8	9	10
#Dm[a]	5355	536	536	535	534	533	534	536	536	538	537
#Nm[b]	5279	525	526	529	529	525	536	528	527	527	527
#Tm[c]	10634	1061	1062	1064	1063	1058	1070	1064	1063	1065	1064
#Dp[d]	617	70	50	84	51	45	87	68	65	49	48
#Np[e]	1836	133	144	154	199	198	240	109	230	221	208
#Tp[f]	2154	190	162	209	215	213	295	151	258	237	224

[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 S2. Summary of 70 features used in DAMpred and their distributions in D10634. 'DM' refers to disease-associated mutations, 'NM' to neutral mutations, and '*p*-value in M-W test' to *p*-value in the Mann-Whitney test for determining whether two datasets are drawn from the same distribution. If the *p*-value is lower than 0.05, the hypothesis that the distributions of the two datasets are the same can be rejected.

Feature Class	No	Feature	Mean		MCC		*p*-value in M-W test	Description
			DM	NM	Cutoff	Value		
Physicochemical Properties	*Pharmacophore for the wild-type residues*							
	1	HP_w	2.90	2.17	2	0.16	3.10E-078	Hydrophobic residue
	2	$noHP_w$	4.47	3.89	6	0.14	1.49E-044	Non-hydrophobic residue
	3	AR_w	1.28	0.92	2	0.14	1.31E-058	Aromatic rings
	4	$noAR_w$	6.09	5.13	6	0.17	3.84E-093	Non-aromatic rings
	5	PC_w	0.86	0.77	3	0.05	7.99E-006	Positive charge
	6	NC_w	0.72	0.75	4	0.02	2.07E-002	Negative charge
	7	noC_w	5.79	4.53	5	0.21	1.20E-132	Neutral charge
	8	BP_w	1.67	1.26	2	0.09	1.38E-013	Both wild-type and neighbor AA are polar
	9	OP_w	2.83	2.21	3	0.14	1.33E-065	Either of wild-type and neighbor AA is polar
	10	NP_w	1.87	1.59	5	0.10	3.63E-006	Both wild-type and neighbor are nonpolar t
	11	AC_w	9.12	8.05	12	0.12	1.19E-021	The count of reside being the hydrogen acceptor
	12	DO_w	5.59	4.83	8	0.14	1.94E-032	The count of reside being the hydrogen donor
	Pharmacophore for the mutant residues							
	13	HP_m	3.04	2.21	3	0.18	2.40E-098	Hydrophobic
	14	$noHP_m$	4.52	3.72	6	0.17	1.68E-081	Non-hydrophobic
	15	AR_m	1.33	0.95	1	0.15	3.22E-063	Aromatic rings
	16	$noAR_m$	6.23	4.98	7	0.21	4.00E-144	Non-aromatic rings
	17	PC_m	0.83	0.68	3	0.07	3.82E-015	Positive charge
	18	NC_m	0.69	0.69	4	0.03	1.45E-001	Negative charge
	19	noC_m	6.04	4.55	5	0.22	3.60E-161	Neutral charge
	20	BP_m	1.38	1.21	3	0.05	3.57E-002	Both wild-type and neighbor AA are polar
	21	OP_m	3.22	2.25	5	0.19	1.20E-121	Either of wild-type and neighbor AA is polar
	22	NP_m	1.95	1.47	5	0.13	4.82E-021	Both wild-type and neighbor are nonpolar t
	23	AC_m	9.31	7.84	15	0.15	1.18E-035	The count of reside being the hydrogen acceptor
	24	DO_m	5.69	4.75	7	0.16	9.94E-042	The count of reside being the hydrogen donor
	Mutation-induced environmental pharmacophore changes							
	25	cos_{WM}	0.74	0.82	0.9	0.22	2.00E-133	cosin for the pharmacophores of wild-type and mutant residues
	26	rms_{WM}	0.47	0.37	0.4	0.22	1.60E-125	RMSD for pharmacophores of wild-type and mutant residues
	27	$cosN_{WM}$	0.94	0.95	1.0	0.08	7.55E-019	cosin for neighbor pharmacophores of wild-type and mutant
	28	$rmsN_{WM}$	2.09	1.52	1.7	0.22	5.10E-154	RMSD for neighbor pharmacophores of wild-type and mutant
	29	$cosNS_{WM}$	0.92	0.93	1.0	0.11	2.02E-022	cosin for neighbor pharmacophores of wild-type and mutant residues related with single residue
	30	$rmsNS_{WM}$	1.77	1.24	1.8	0.24	1.30E-176	RMSD for neighbor pharmacophores of wild-type and mutant residues related with single residue
	31	$cosNP_{WM}$	0.92	0.95	0.8	0.06	2.53E-003	cosin for neighbor pharmacophores of wild-type and mutant residues related with residue paired
	32	$rmsNP_{WM}$	2.87	2.21	4.2	0.14	3.02E-042	RMSDfor neighbor pharmacophores of wild-type and mutant residues related with residue paired
	Other physicochemical properties							
	33	Volw	2.83	2.86	1.9	0.10	9.84E-002	The volume of wild-type residue
	34	Volm	2.91	2.88	3.2	0.11	3.86E-008	The volume of mutant residue
	35	dVol	0.08	0.02	0.7	0.15	1.23E-008	The volume difference
	36	Ww	132.0	130.81	75	0.10	5.92E-003	The weight of wild-type residue
	37	Wm	136.2	131.57	165	0.11	5.60E-012	The weight of mutant residue
	38	dW	4.23	0.76	42	0.17	5.46E-006	The molecular weight difference
Evolutionary Profiles	PSI-BLAST profile scores							
	39	$PSIC_w$	1.57	0.91	1.2	0.43	0.00E-000	The PSIC score for wild-type residue
	40	$PSIC_m$	-0.42	0.24	-0.1	0.45	0.00E-000	The PSIC score for mutant residue
	41	dPSIC	-1.99	-0.66	-1.1	0.54	0.00E-000	The PSIC score difference
	42	JSD_w	0.03	0.03	0.04	0.11	3.70E-001	The JSD score for wild-type residue
	43	JSD_m	0.02	0.03	0.03	0.11	1.22E-016	The JSD score for mutant residue
	44	dJSD	0.00	0.00	-0.01	0.10	1.99E-012	The JSD score difference
	45	JSD_i	0.47	0.32	0.5	0.32	2.20E-273	The JSD score at mutant position *i*
	LOMETS profile scores							
	46	$tPSIC_w$	0.78	0.45	0.8	0.22	1.00E-149	The PSIC score for wild-type residue
	47	$tPSIC_m$	-0.30	-0.02	0.04	0.19	1.50E-108	The PSIC score for mutant residue
	48	dtPSIC	-1.08	-0.47	-0.7	0.28	8.80E-243	The PSIC score difference
	Pfam profile scores							
	49	$Pfam_w$	1.83	2.40	1.5	0.32	9.40E-178	The Pfam score for wild-type residue
	50	$Pfam_m$	3.66	3.00	3.0	0.29	7.20E-119	The Pfam score for mutant residue

	51	dPfam	1.83	0.59	1.1	0.38	1.20E-298	The Pfam score difference
Contact Environments	Directly contacted residues							
	52	Intra	14.51	11.35	15	0.29	5.80E-245	The number of intramolecular contacts
	53	FunIntra	4.90	3.55	15	0.13	2.58E-026	The number of intramolecular functional contacts
	54	Inter	3.82	3.67	29	0.03	1.25E-001	The number of intermolecular contacts
	55	FunInter	0.54	0.43	25	0.02	6.29E-002	The number of intermolecular functional contacts
	Indirectly contacted residues							
	56	CIntra	65.98	51.58	55	0.25	4.10E-180	The number of intramolecular indirectly contacts
	57	CFunIntra	22.03	16.51	60	0.12	8.25E-025	The number of intramolecular functional indirectly contacts
	58	CInter	23.55	21.26	149	0.05	4.85E-002	The number of intermolecular indirectly contacts
	59	CFunInter	9.59	8.25	139	0.03	8.35E-004	The number of intermolecular functional indirectly contacts
I-TASSER model Based Properties	*Protein surface regions favorable for interactions*							
	60	CS	0.08	0.04	0.04	0.18	8.66E-093	the ConCavity score for the wild-type score
	61	Depth	6.72	5.38	5.6	0.27	2.60E-204	The average distance of atoms of wild-type residue to its closest molecule of bulk solvent,
	The energy function							
	62	ED	632.06	587.51	556	0.09	6.46E-011	The EvoDesign score
	63	ddG	1.62	0.52	3.0	0.22	8.97E-091	The stability changes upon mutation
	64	VDW_w	-343.4	-332.43	-358	0.07	1.63E-004	Van Der Waals potential of the wild-type residue from CISS-RR
	65	VDW_m	-331.3	-326.79	-1041	0.06	3.73E-002	Van Der Waals potential of the mutant residue from CISS-RR
	66	dVDW	12.11	5.65	2.3	0.24	1.20E-155	Van Der Waals potential difference
	67	RT_w	460.13	401.52	579	0.11	3.66E-027	rotamer term which measures the preferences of the wild-type side-chain conformers from CISS-RR.
	68	$CISRR_w$	116.69	69.09	-11	0.30	5.70E-119	CIS-RR score for the wild-type residue
	69	$CISRR_m$	129.06	74.96	-5	0.30	1.30E-149	CIS-RR score for the mutant residue
	70	dCISRR	12.37	5.87	4.5	0.21	1.40E-115	CIS-RR score difference

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.

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

[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

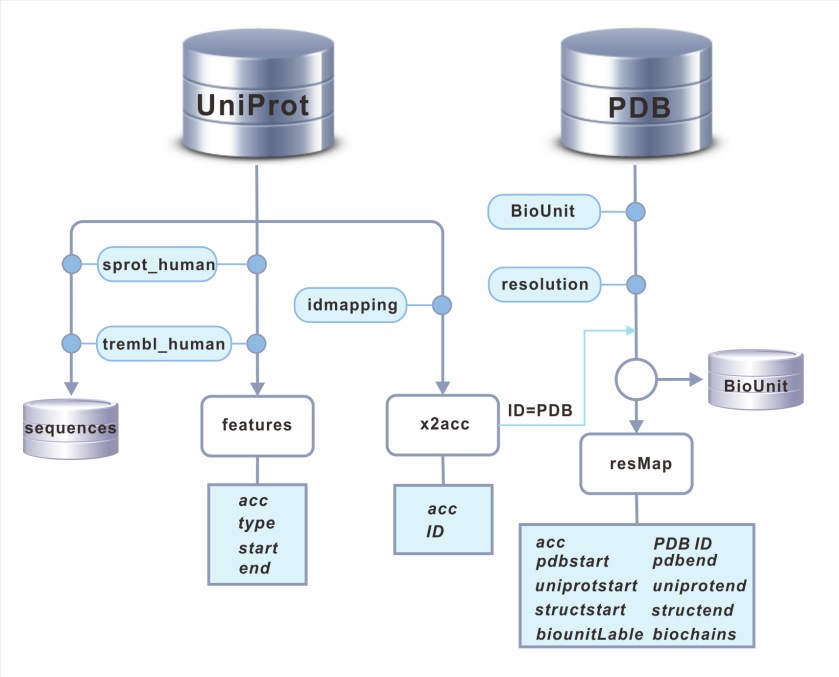

Figure S1. Pipeline for data mapping and dataset construction.

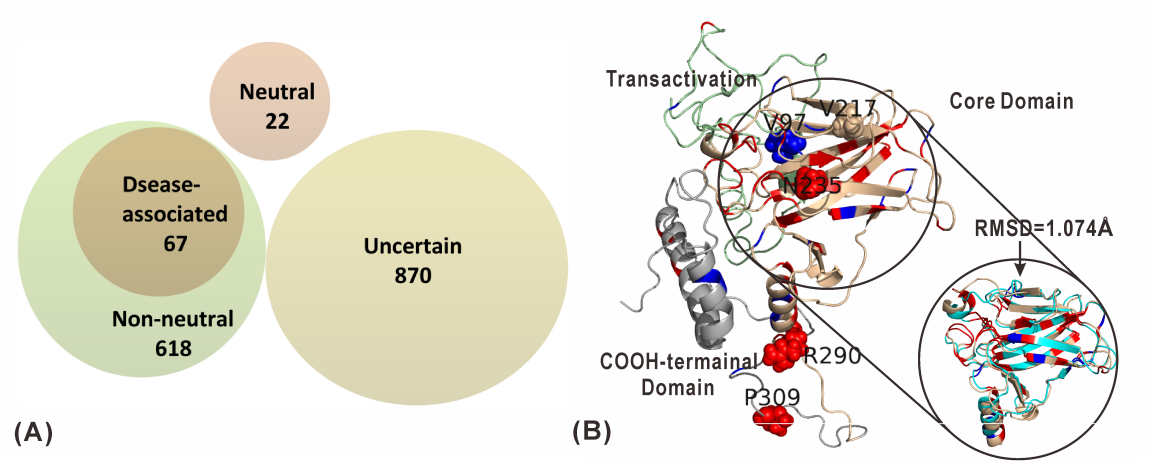

Figure S2. Disease-association mutation prediction on TP53 protein. (A) A Venn diagram showing the distribution of different mutations on the p53 protein. (B). I-TASSER model for the isoform P04637-1 of TP53 gene, where the transactivation, domain and COOH-terminal domains are marked in different colors. 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.

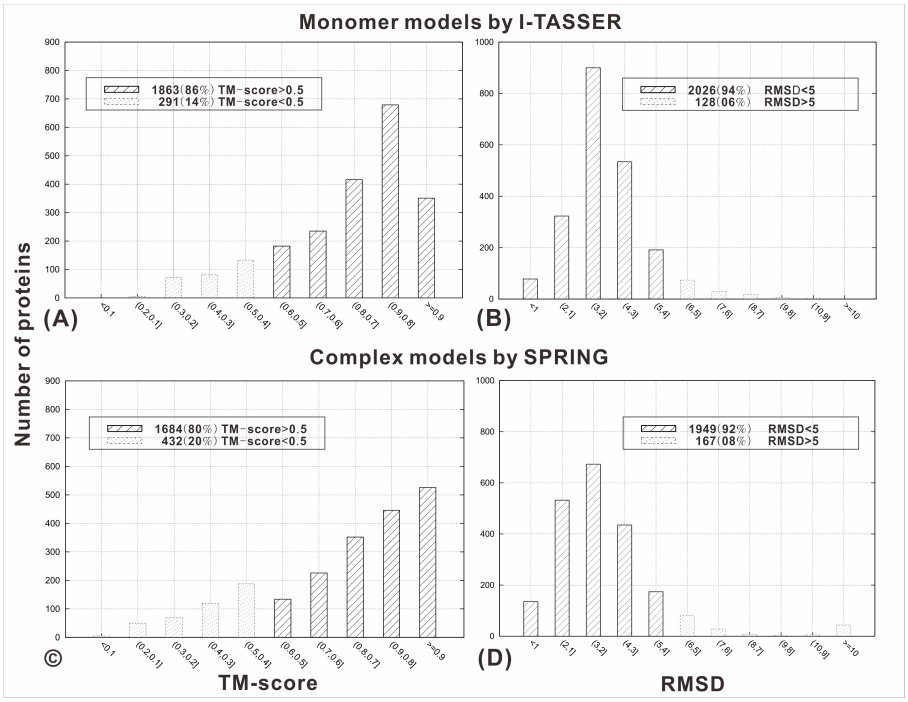

Figure S3. Quality of the protein structure predictions by I-TASSER and SPRING. (A, B) Histogram distribution of TM-score and RMSD of the I-TASSER models on 1974 proteins from the D10634 dataset. (C, D) Histogram distribution of TM-score and RMSD of the SPRING models on 2116 BioUnit complexes from the D10634 dataset.

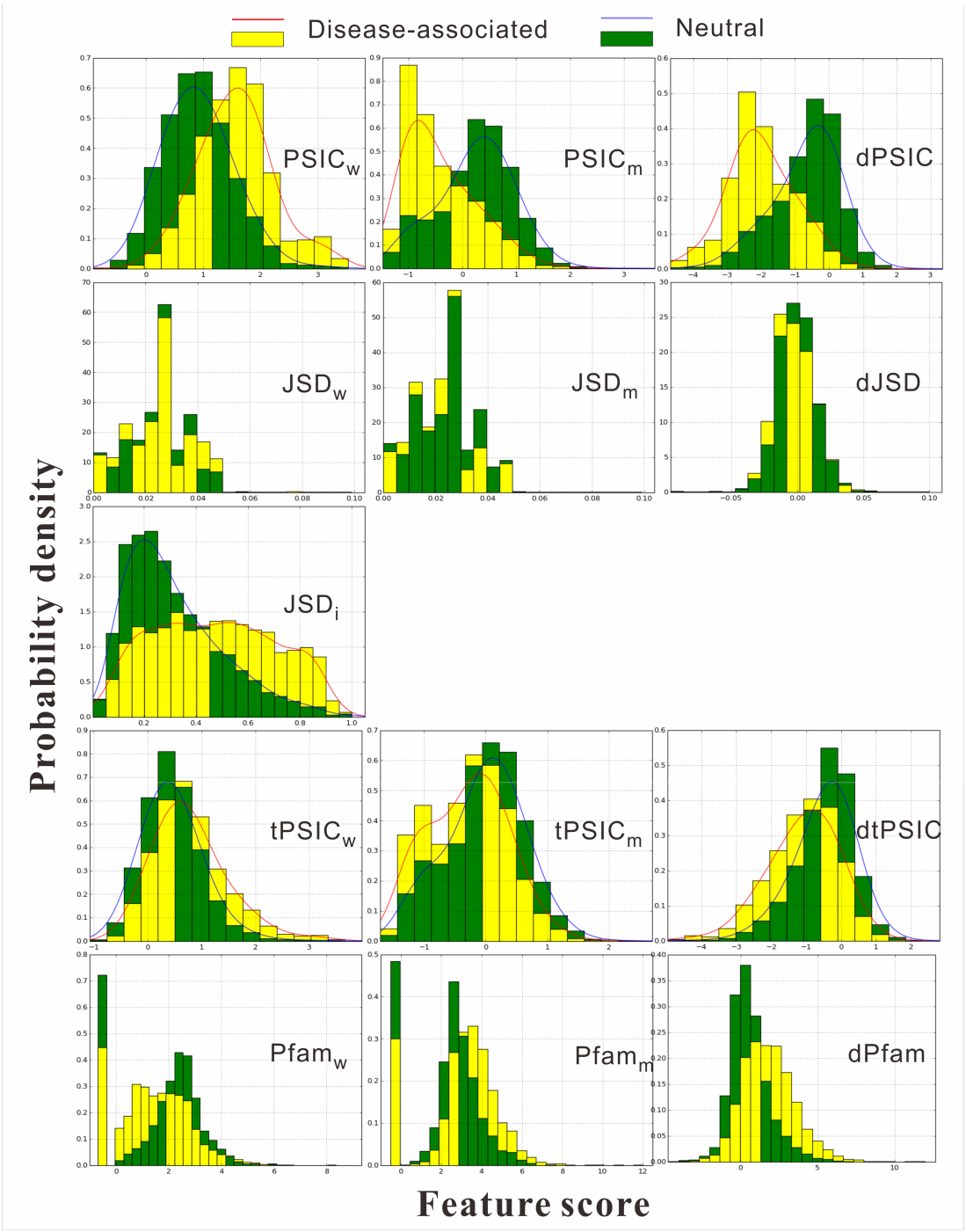

Figure S4. Histogram distribution of evolution features on disease-associated and neutral mutations from the D10634 dataset. The left-most bar laying in Pfam diagrams with feature score < 0 indicate the mutant positions not to be found in the Pfam families.

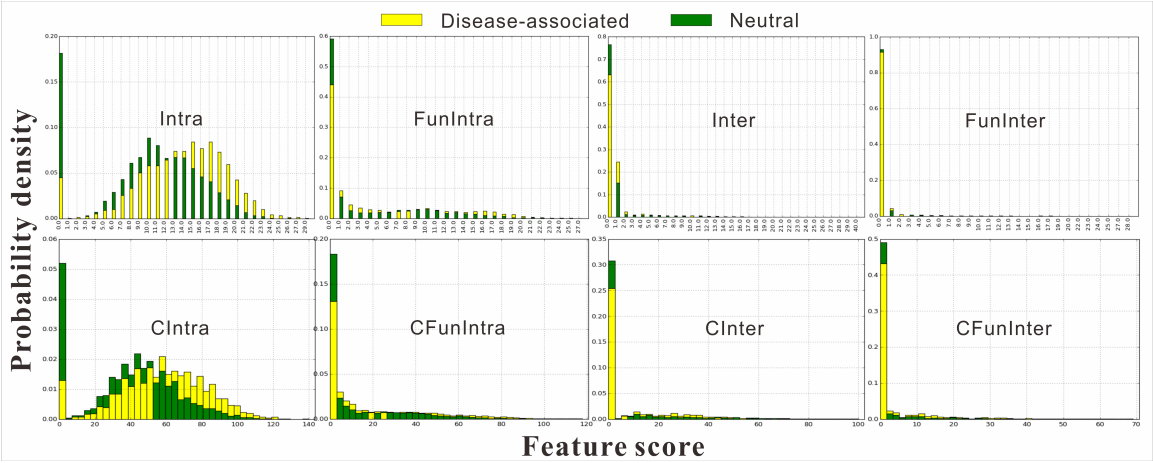

Figure S5. Histogram distribution of Contact features on SPRING biological assembly on disease-associated and neutral mutations from the D10634 dataset.

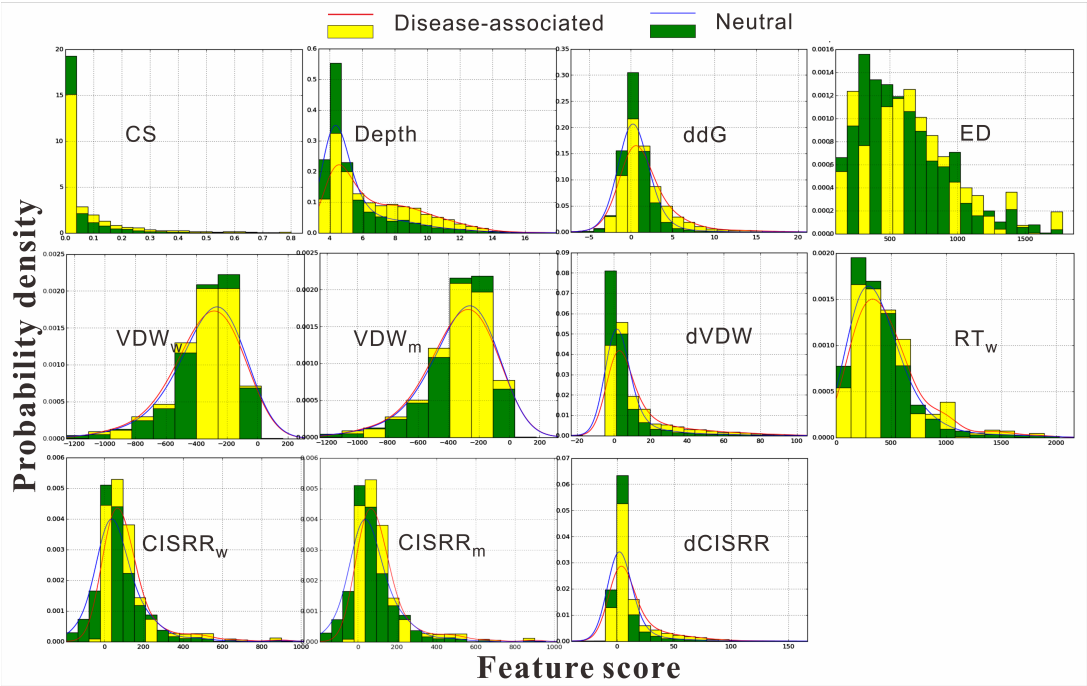

Figure S6. Histogram distribution of the I-TASSER model based features on disease-associated and neutral mutations from the D10634 dataset.

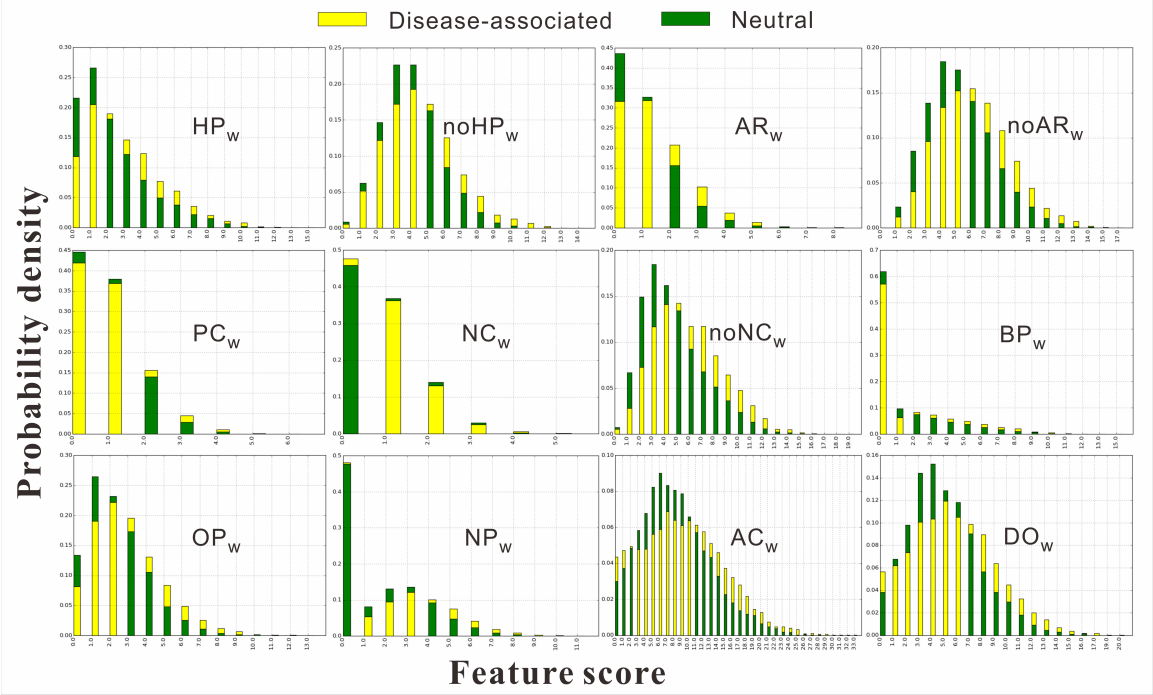

Figure S7. Histogram distribution of pharmacophore features for wild-type residue on disease-associated and neutral mutations from the D10634 dataset.

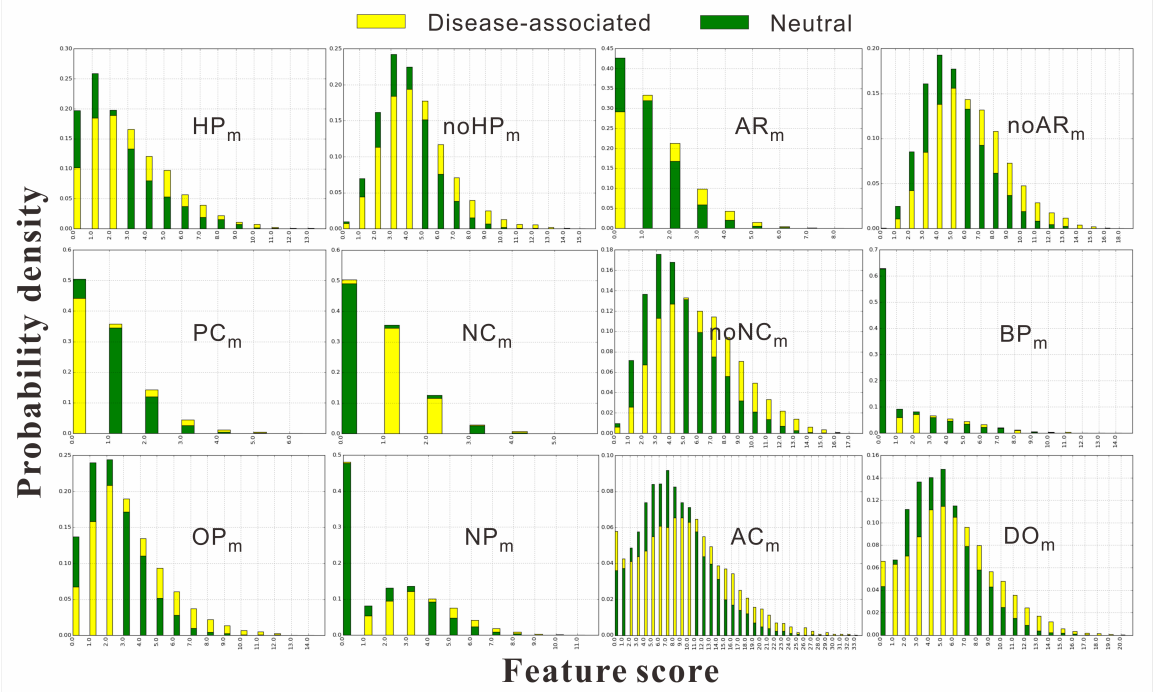

Figure S8. Histogram distribution of pharmacophore features for mutant residue on disease-associated and neutral mutations from the D10634 dataset.

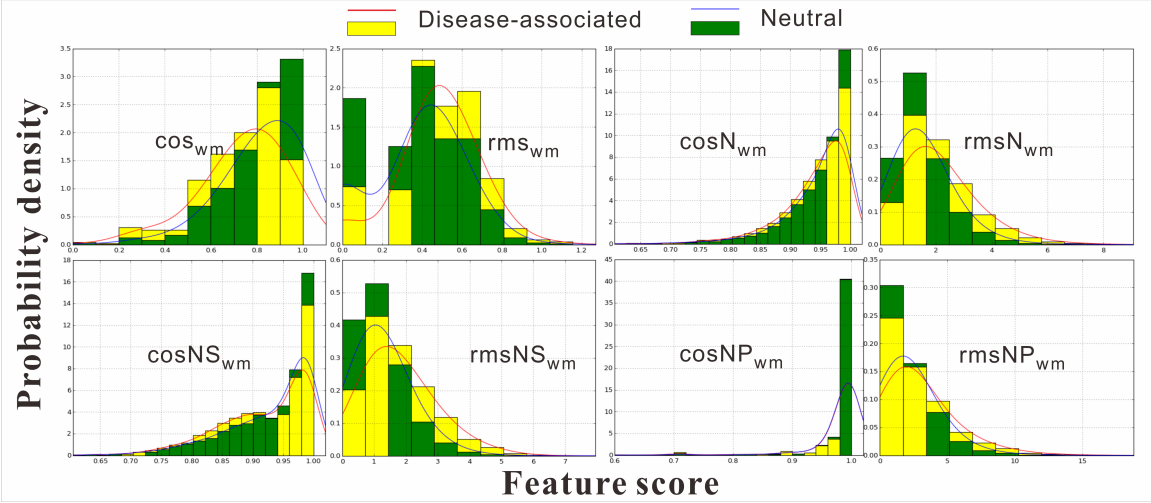

Figure S9. Histogram distribution of pharmacophore difference features on disease-associated and neutral mutations from the D10634 dataset.

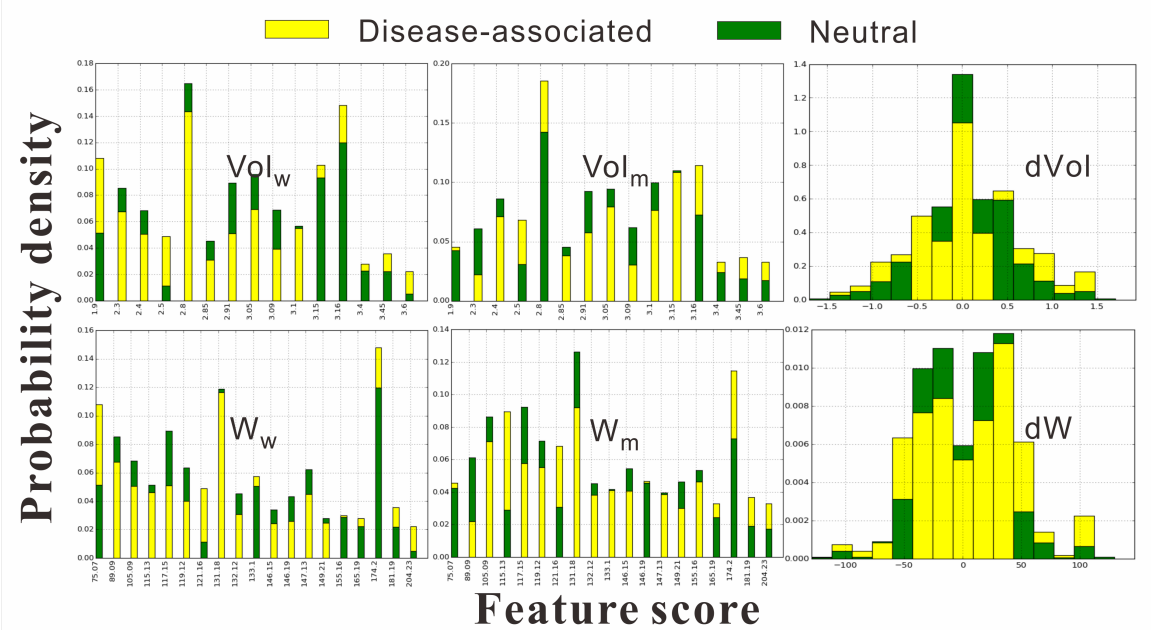

Figure S10. Histogram distribution of chemical features on disease-associated and neutral mutations from the D10634 dataset.

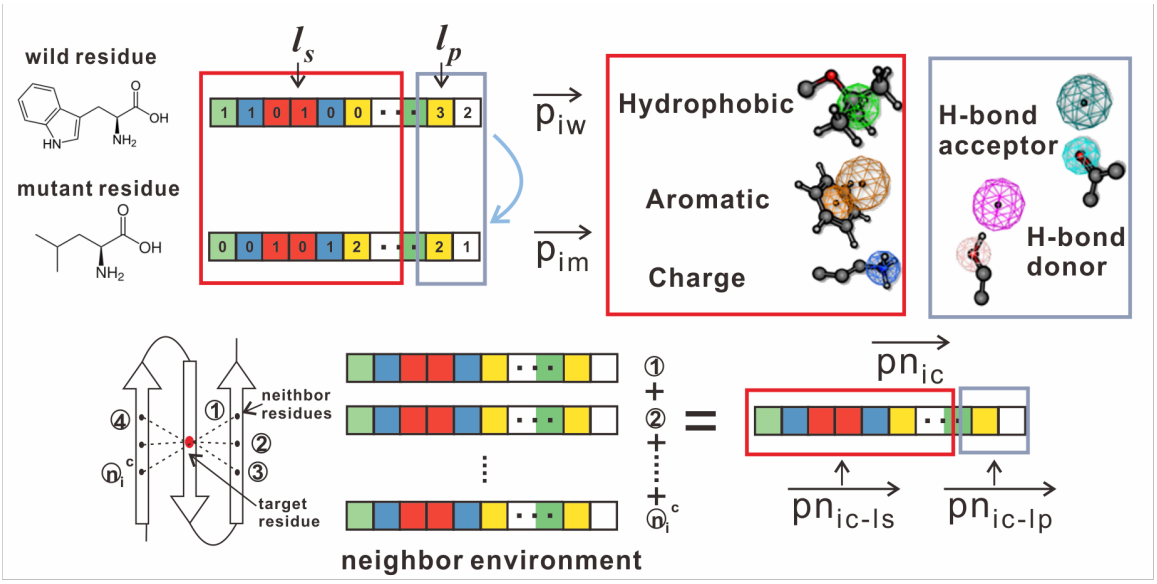

Figure S11. Illustration of the environmental pharmacophore for physicochemical properties.

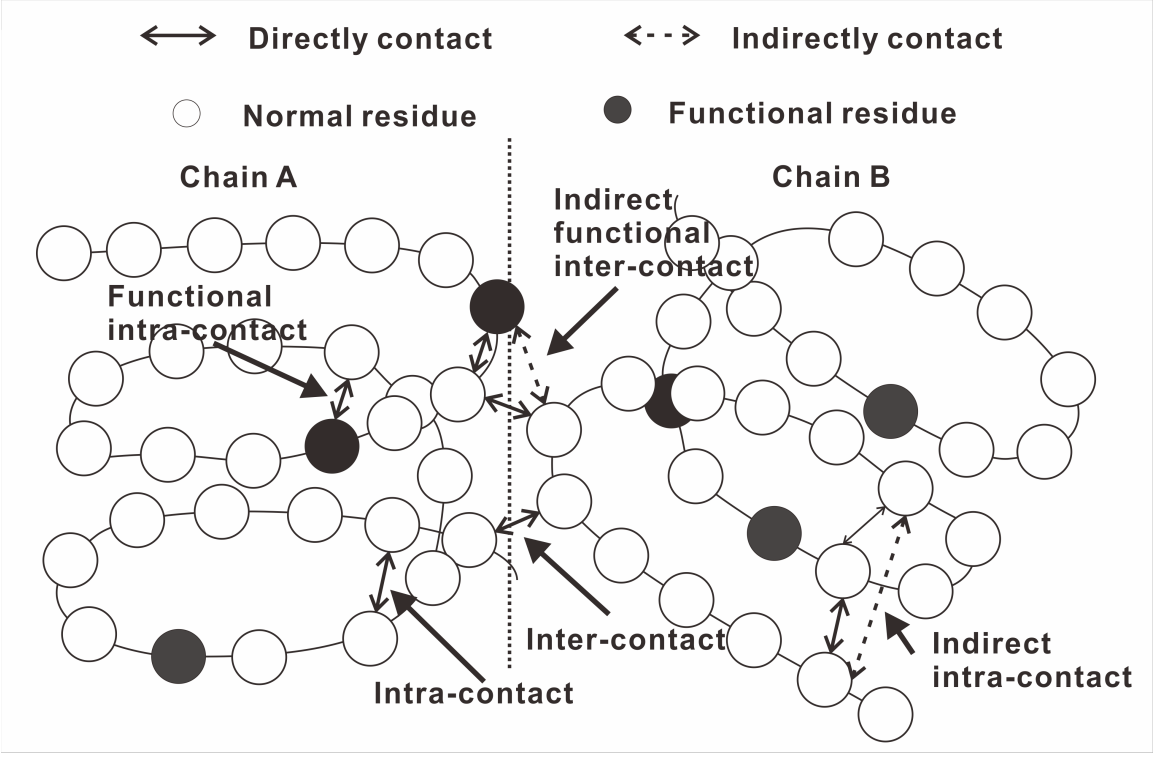

Figure S12. Illustration of contact environments in the SPRING biological assembly.

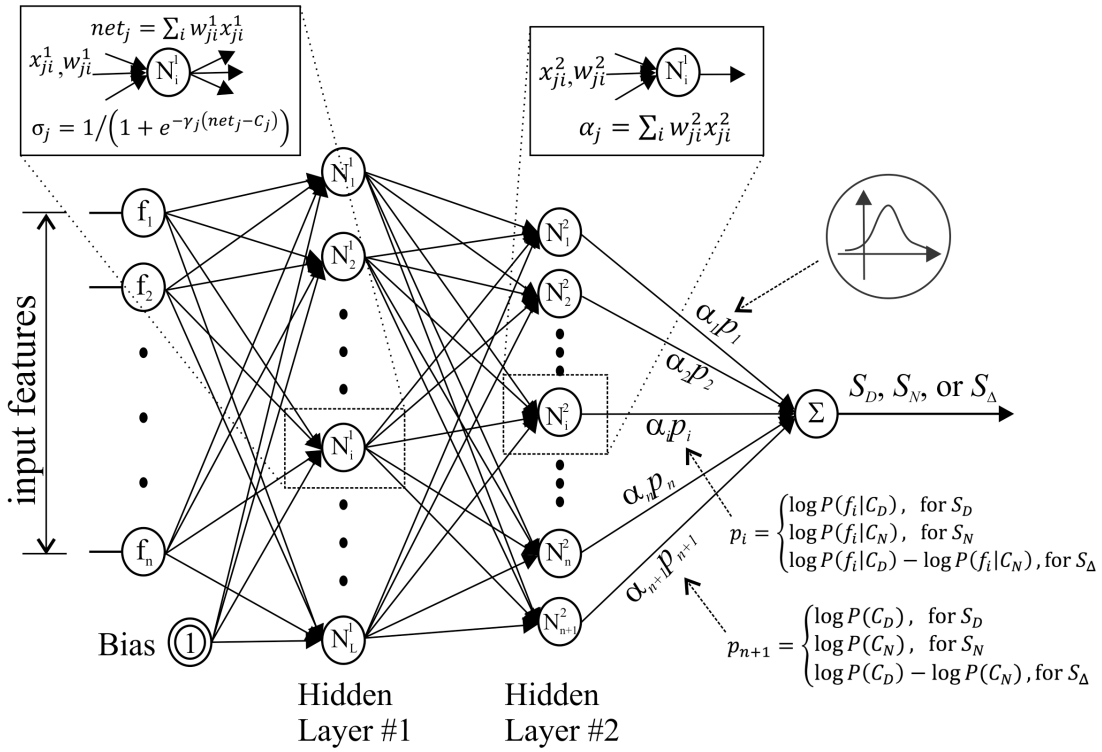

Figure S13. Illustration of the Bayes-guided Artificial Neural Network (BANN) learning model.

References cited in Supporting Information:

[1] Boutet E, Lieberherr D, Tognolli M, Schneider M, Bansal P, Bridge AJ, et al. UniProtKB/Swiss-Prot, the Manually Annotated Section of the UniProt KnowledgeBase: How to Use the Entry View. Methods Mol Biol. 2016;1374:23-54.
[2] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic acids research. 2000;28:235-42.
[3] Hipp R. SQLite, https://www.sqlite.org.
[4] Pires DEV, Ascher DB, Blundell TL. mCSM: predicting the effects of mutations in proteins using graph-based signatures. Bioinformatics. 2014;30:335-42.
[5] Pires DEV, de Melo-Minardi RC, da Silveira CH, Campos FF, Meira W. aCSM: noise-free graph-based signatures to large-scale receptor-based ligand prediction. Bioinformatics. 2013;29:855-61.
[6] Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research. 1997;25:3389-402.
[7] Wu S, Zhang Y. LOMETS: A local meta-threading-server for protein structure prediction. Nucl Acids Res. 2007;35:3375-82.
[8] Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: towards a more sustainable future. Nucleic acids research. 2016;44:D279-85.
[9] Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular systems biology. 2011;7:539.
[10] Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Sciences of the United States of America. 1992;89:10915-9.
[11] Sunyaev SR, Eisenhaber F, Rodchenkov IV, Eisenhaber B, Tumanyan VG, Kuznetsov EN. PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. Protein Eng. 1999;12:387-94.
[12] Yang J, Roy A, Zhang Y. Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. Bioinformatics. 2013;29:2588-95.
[13] Capra JA, Laskowski RA, Thornton JM, Singh M, Funkhouser TA. Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. PLoS computational biology. 2009;5:e1000585.
[14] Guerler A, Govindarajoo B, Zhang Y. Mapping Monomeric Threading to Protein-Protein Structure Prediction. Journal of chemical information and modeling. 2013;53:717-25.
[15] Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. Nature protocols. 2010;5:725-38.
[16] Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. The I-TASSER Suite: protein structure and function prediction. Nature methods. 2015;12:7-8.
[17] Tan KP, Nguyen TB, Patel S, Varadarajan R, Madhusudhan MS. Depth: a web server to compute depth, cavity sizes, detect potential small-molecule ligand-binding cavities and predict the pKa of ionizable residues in proteins. Nucleic Acids Res. 2013;41:W314-21.
[18] Mitra P, Shultis D, Zhang Y. EvoDesign: De novo protein design based on structural and evolutionary profiles. Nucleic acids research. 2013;41:W273-80.

[19] Xiong P, Wang M, Zhou X, Zhang T, Zhang J, Chen Q, et al. Protein design with a comprehensive statistical energy function and boosted by experimental selection for foldability. Nature communications. 2014;5:5330.
[20] Cao Y, Song L, Miao Z, Hu Y, Tian L, Jiang T. Improved side-chain modeling by coupling clash-detection guided iterative search with rotamer relaxation. Bioinformatics. 2011;27:785-90.