nature machine intelligence

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s42256-019-0119-z

In the format provided by the authors and unedited.

Predicting disease-associated mutation of metal-binding sites in proteins using a deep learning approach

Mohamad Koohi-Moghadam[1,2], Haibo Wang[1], Yuchuan Wang[1], Xinming Yang[1], Hongyan Li[1], Junwen Wang[2,3,4,5]* and Hongzhe Sun[1]*

[1]Department of Chemistry, The University of Hong Kong, Hong Kong, China. [2]Center for Individualized Medicine, Mayo Clinic, Scottsdale, AZ, USA. [3]Department of Health Sciences, Mayo Clinic, Scottsdale, AZ, USA. [4]Department of Molecular Pharmacology and Experimental Therapeutics, Mayo Clinic, Scottsdale, AZ, USA. [5]College of Health Solutions, Arizona State University, Scottsdale, AZ, USA. *e-mail: wang.junwen@mayo.edu; hsun@hku.hk

Predicting disease-associated mutation of metal-binding sites of proteins using a deep learning approach

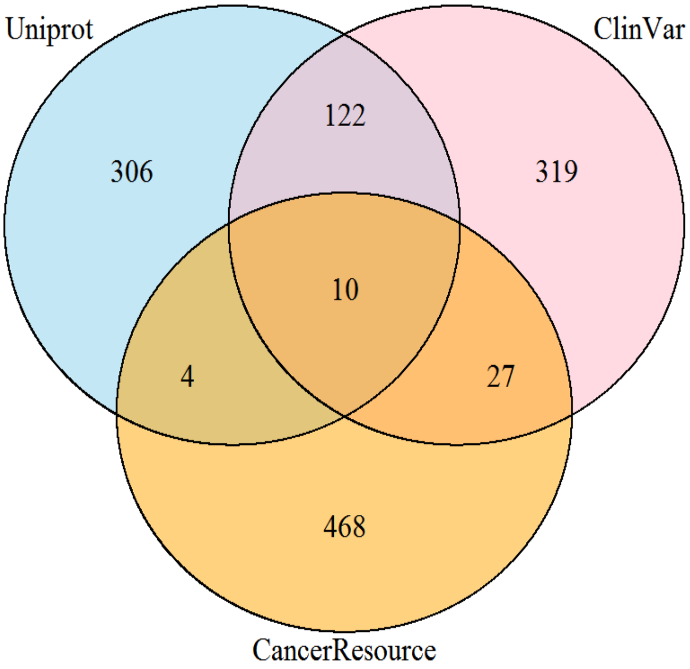
Mohamad Koohi-Moghadam[a,b], Haibo Wang[a], Yuchuan Wang[a], Xinming Yang[a], Hongyan Li[a], Junwen Wang[b,c,*]and Hongzhe Sun[a,*]

[a] Department of Chemistry, The University of Hong Kong, Hong Kong, P. R. China.

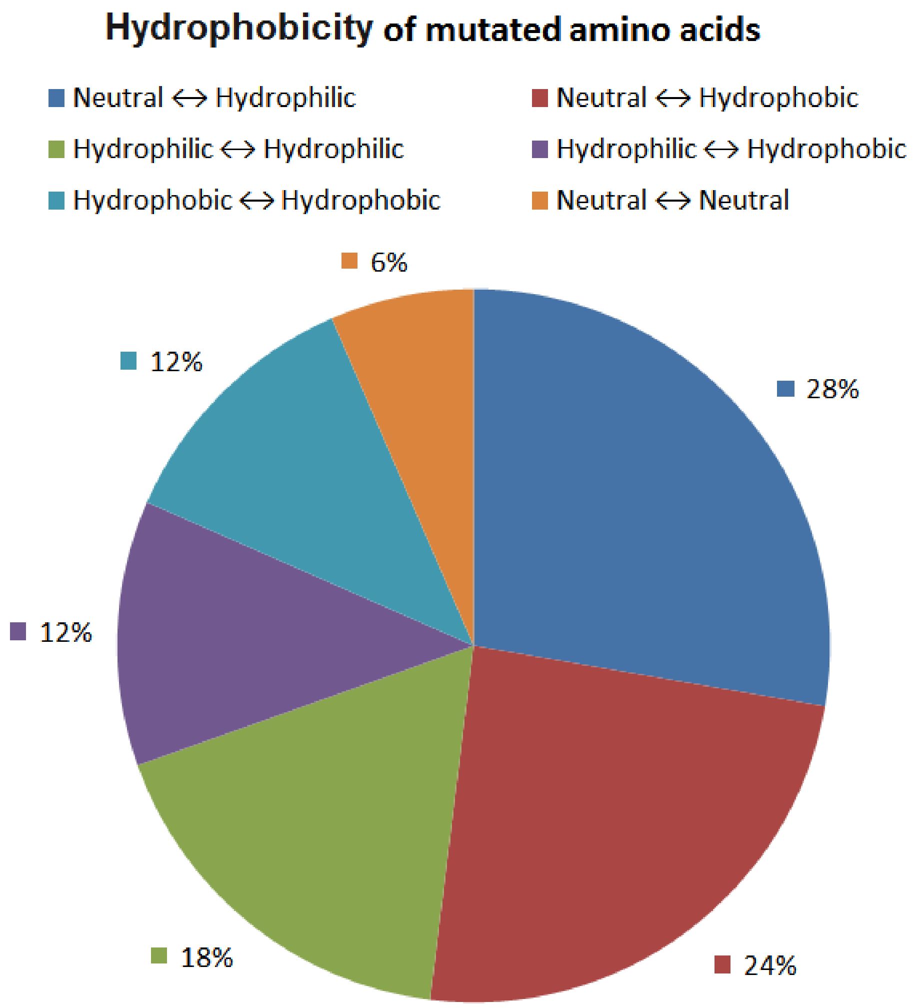
[b] Departments of Health Sciences Research and Molecular Pharmacology & Experimental Therapeutics, and Center for Individualized Medicine, Mayo Clinic, Scottsdale, AZ 85259, USA.

[c] College of Health Solutions, Arizona State University, Scottsdale, AZ 85259, USA.

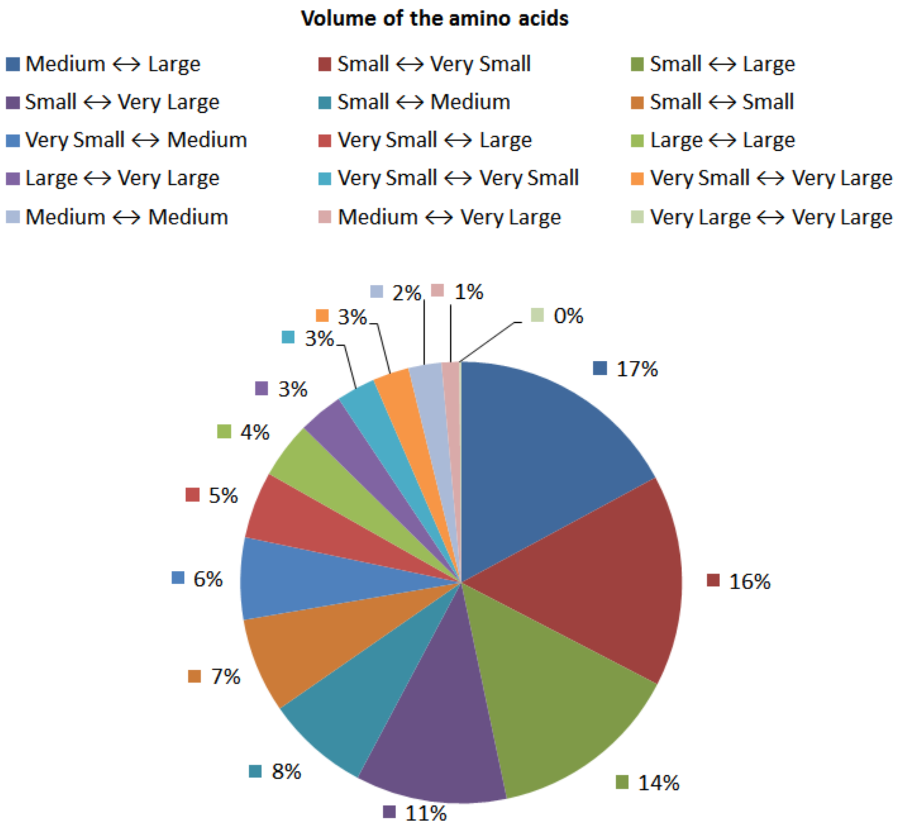
Supplementary Figures



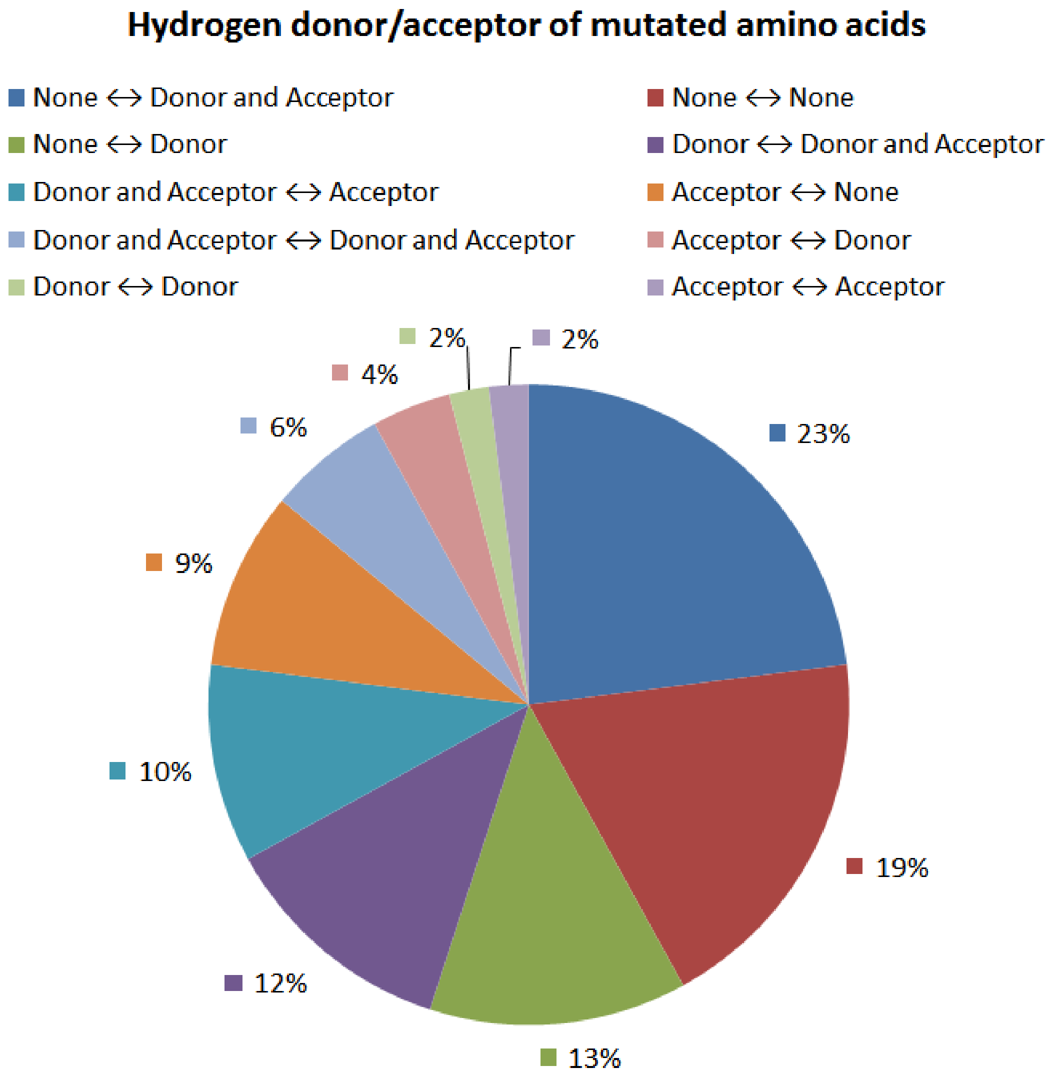
Supplementary Figure 1 | Venn diagram of the disease-associated mutations of the metal-binding sites obtained from different database. We collected the missense mutations that occur in the metal-binding sites from three different databases. Both germline and somatic mutations were considered. We found 478 missense mutations from ClinVar, 442 from Uniprot Hamsavar and 509 from CancerResource2. And these disease-associated mutations were used as positive set to train the proposed model.



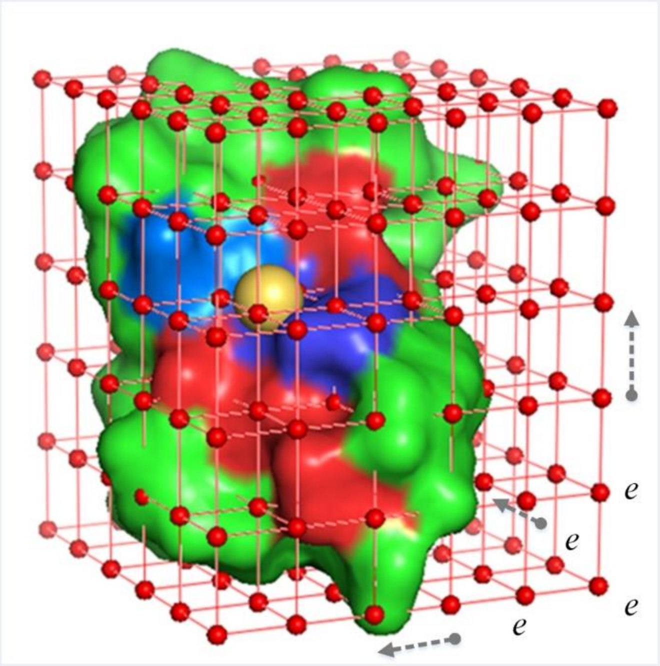
Supplementary Figure 2 | Summary of hydrophobicity of the mutated amino acids in metal-binding sites. The result shows most of the disease-associated amino acids in the metal-binding sites are mutated from neutral to hydrophilic, followed by a mutation from neutral to hydrophobic and hydrophilic to hydrophobic with proportions of 24% and 12%, respectively.



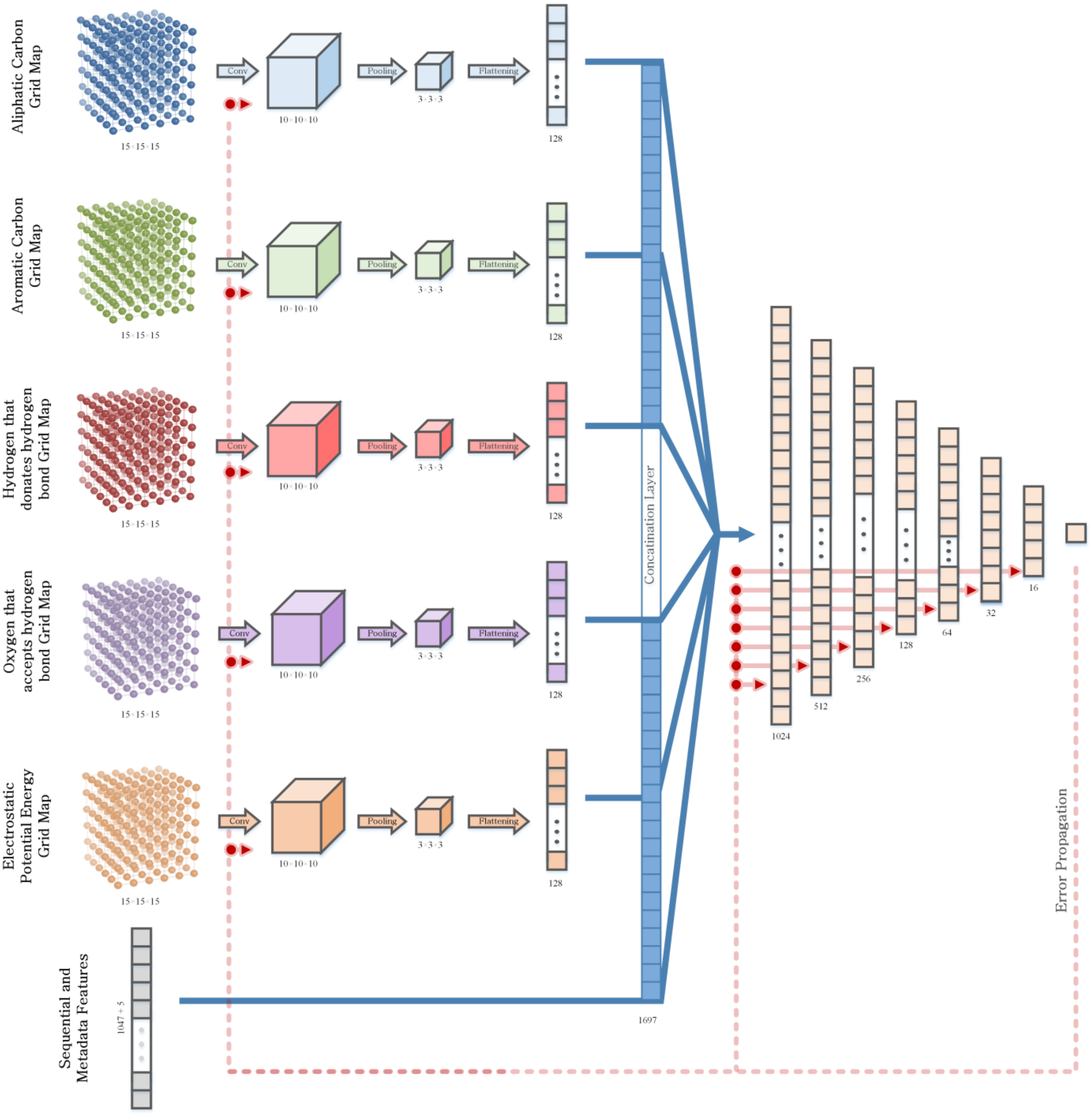
Supplementary Figure 3 | Summary of volume of the mutated amino acids in metal-binding sites. The volume study shows even a small change of the size of amino acids in the metal-binding site may change the function of the protein. The mutation from medium size to large has the first rank with 17%, followed by the second rank of small to very small amino acids (16%) and the third rank of small size to large size (14%).



Supplementary Figure 4 | Summary of hydrogen donor/acceptor of the mutated amino acid in metal-binding sites. The results show mutation from an amino acid with none donor or acceptor to a donor or acceptor amino acid have a great impact on the metal-binding site and are related to diseases.



Supplementary Figure 5 | Electrostatic affinity grid map of a metal-binding site. We generated five different affinity grid maps for each metal-binding site. Aliphatic carbon, aromatic carbon, hydrogen that donates hydrogen, oxygen that accepts hydrogen and electron are five probes that are used to build the spatial features. For example, here an electrostatic grid map was produced by putting an electron into each probe (red dots). We calculated the pairwise interaction energy between each probe and all the metal-binding sites atoms. We then stored the Coulombic interactions values as a three-dimensional lattice, and used it as input of our deep learning model. Here, the size of grid maps are 15×15×15Å^3 and grid spacing is 1Å.



Supplementary Figure 6 | Schematic of the MCCNN model. A multi-channel convolutional neural network was used to predict disease-associated mutations. We generated five three-dimensional lattices with the size of the 15×15×15 as spatial features. Each lattice was used in a separated channel. In each channel, we used convolution layer with the size of 10×10×10 to extract the features and a max-pooling layer with the size of 3×3×3 to reduce the number of features. We then flattened the max-pooling output to a vector with a size of 128. We also used the sequential features (1047 features) and metadata features (5 features) in a separated channel. We concatenated the output of these six channels. We finally used a fully connected neural network with 8 layers to integrate the information and generate the final probability. The error also propagates into a different layer of the network.

Supplementary Tables

Table S1. Disease associated mutations of metal-binding pocket

Attached Supplementary_Table_1.xlsx file

Table S2. Benign mutations of metal-binding pocket

Attached Supplementary_Table_2.xlsx file

Supplementary Table 3 | The number of diseases-associated and benign mutations in different metal-binding sites

Metal	Disease-associated dataset	Benign dataset
Zn	447	84
Ca	334	69
Mg	171	23
Fe	66	29
Na	59	12
Ni	44	6
Mn	40	3
K	31	2
Co	16	6
Cu	14	3
Sm	9	2
Hg	8	7
Cd	9	7
As	5	0
W	2	1
Au	1	0
Pb	0	1
Pt	0	1
Sr	0	1
Y	0	4
Total	1256	261

Supplementary Table 4 | 17 main heading names of the diseases. The name of diseases were mapped from different databases to the MeSH database using a python code.

Taxa	**Tree Number**	**Unique ID**
Aero-Digestive Tract disease	C04.588.443	D006258
Blood disease	C15.378	D006402
Breast disease	C17.800.090	D001941
Cardiovascular Disease	C14	D002318
Gastrointestinal disease	C06.405	D005767
Hereditary disease	C16.320	D030342
Immune system disease	C20	D007154
Kidney disease	C12.777.419	D007674
Liver disease	C06.552	D008107
Lung disease	C08.381	D008171
Metabolic disease	C18.452	D008659
Muscular disease	C05.651	D009135
Nervous System disease	C10	D009422
Ovarian disease	C19.391.630	D010049
Pancreatic disease	C06.689	D010182
Prostatic disease	C12.294.565	D011469
Skin and Connective Tissue disease	C17	D017437

Supplementary Table 5 | Hydrophobicity classification of amino acids.

Class	**Amino acids**
Hydrophobic	A, C, I, L, M, F, W, V
Hydrophilic	R, N, D, Q, E, K
Neutral	G, H, P, S, T, Y

Supplementary Table 6 | Volume classification of amino acids.

Class	**Amino acids**
Very small	A, G, S
Small	N, D, C, P, T
Medium	Q, E, H, V
Large	R, I, L, K, M
Very large	F, W, Y

Supplementary Table 7 | Hydrogen donor or acceptor classification of amino acids.

Class	**Amino acids**
Donor	R, K, W
Acceptor	D, E
Donor and acceptor	N, Q, H, S, T, Y
None	A, C, G, I, L, M, F, P, V

Supplementary Table 8 | Sequential features of metal-binding sites that are extracted using propy.

Feature groups	Features	No. of descriptors
Amino acid composition	Amino acid composition	20
Autocorrelation	Normalized Moreau–Broto autocorrelation	240
	Moran autocorrelation	240
	Geary autocorrelation	240
Composition, transition and distribution	Composition	21
	Transition	21
	Distribution	105
Quasi-sequence order	Sequence-order-coupling number	60
Pseudo-amino acid composition	Type I pseudo-amino acid composition	50

Supplementary Table 9 | The prediction result of the unseen data

Attached Supplementary_Table_9.xlsx file