## Supplementary Notes

### Extended introduction

Missense variants are a type of genetic variation, which causes an amino acid substitution with a single nucleotide change in the protein-coding region of the genome. Missense variants are a major class of genetic risk across a broad range of common and rare diseases, such as cancer[1], autism[2], congenital heart disease[3], and epilepsy[4]. However, the functional effects of most missense variants reported in clinical genetic testing are unknown and are classified as variants of uncertain significance (VUS). For example, in the ClinVar database of human variations and phenotypes, ~75% of missense variants are VUS and ~5% are classified with ambiguity, for which several research groups give conflicting interpretations[5,6]. Variants with an allele frequency less than 1e-4 in population are considered as rare variants otherwise common variants[7]. Based on population genetics, common variants are extremely unlikely to be under strong selection, therefore they usually do not have a large genetic impact. A small fraction of common variants may still have functional impact, but a better way to study them is by genetic association, for which there is enough statistical power for common variants given reasonable sample size. Statistical association analysis of individual rare variants requires prohibitively large sample sizes to reach sufficient statistical power[2,7]. Therefore, prediction methods are needed to help identify those rare variants that are most likely to cause disease.

Numerous methods have been developed to address the problem. These methods differ in several aspects, including the prediction features, the model architecture, how the features are represented in the model, the training data sets, and how the model is trained.

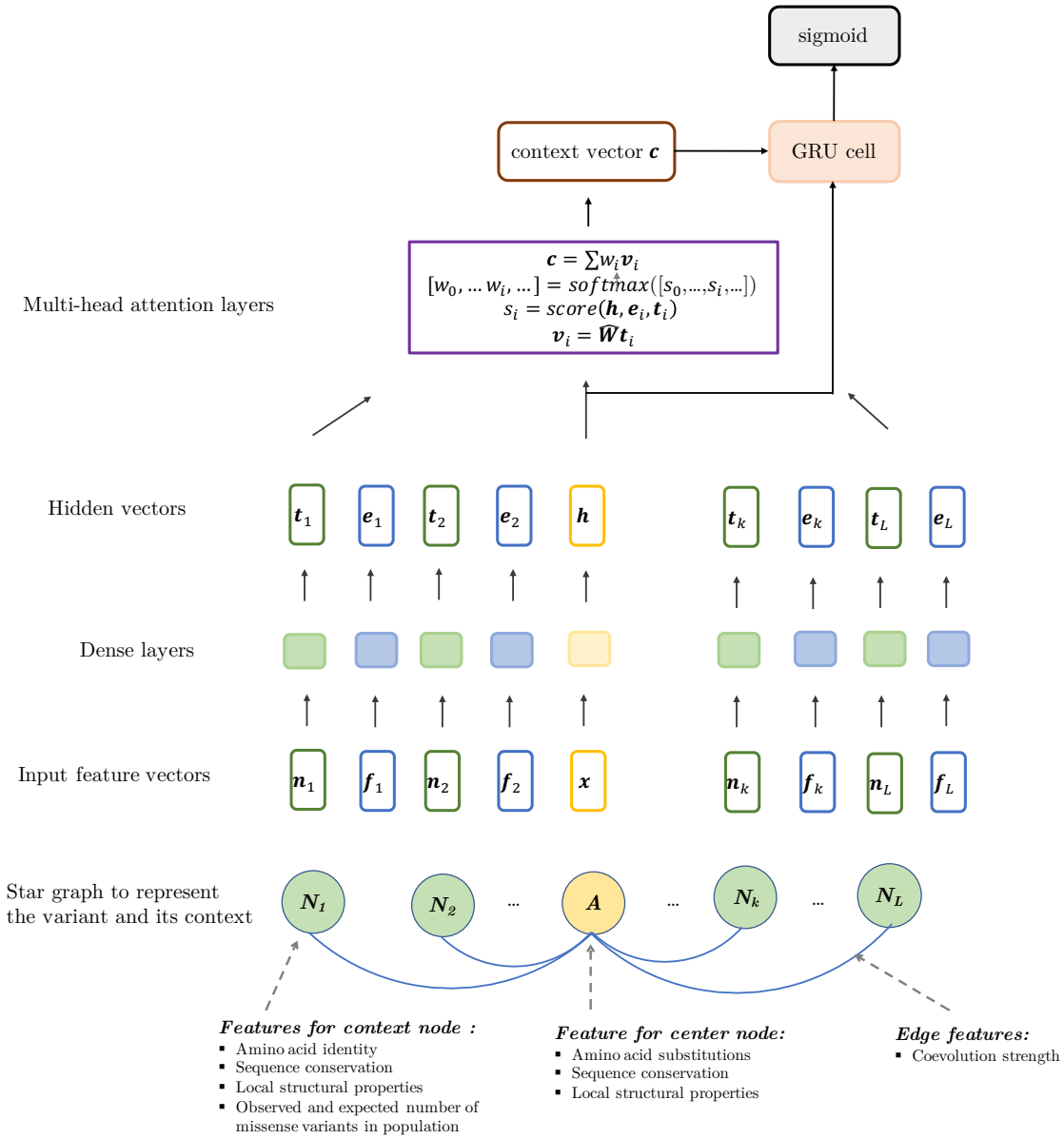
Sequence conservation is the main prediction features for early computational methods such as GERP[8], SIFT[9], and phastCons[10]. PolyPhen2[11] and MVP[12] also include protein local structural properties such as protein secondary structures. MPC[13] and CCRs[14] estimate sub-genic coding constraints from large human population sequencing data which provide additional information not captured by previous methods.

Several machine learning-based methods have been developed to ensemble these features or existing scores. CADD[15] is a meta method based on a support vector machine model, while REVEL[16], ClinPred[17], and M-CAP[18] used decision tree-based machine learning models such as random forest and gradient boosting tree model. MVP is based on convolutional neural networks.

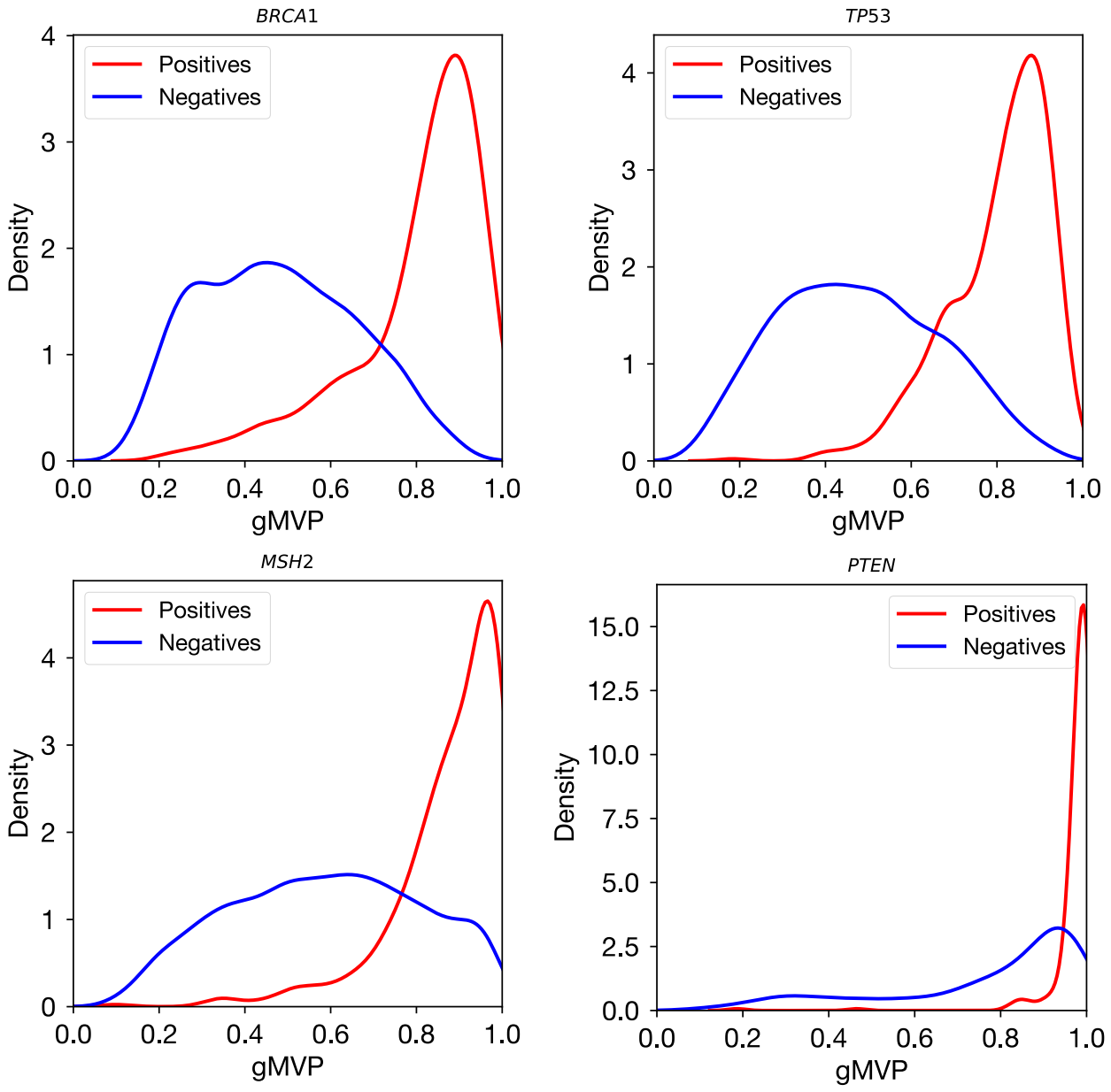
Learning feature representations from raw data instead of engineered features or exiting scores is another trend in literature. For example, PrimateA[19]I learns protein context from sequences and local structural properties using deep representation learning instead of using the existing conservation scores such as SIFT and GERP. Representation learning can avoid using any previous prediction tools. On the one hand, the predicted scores from representation learning

can be used as another independent feature source in the future ensemble predictors. On the other hand, the learned representations from raw data are more optimal for the machine learning model than the engineered features.

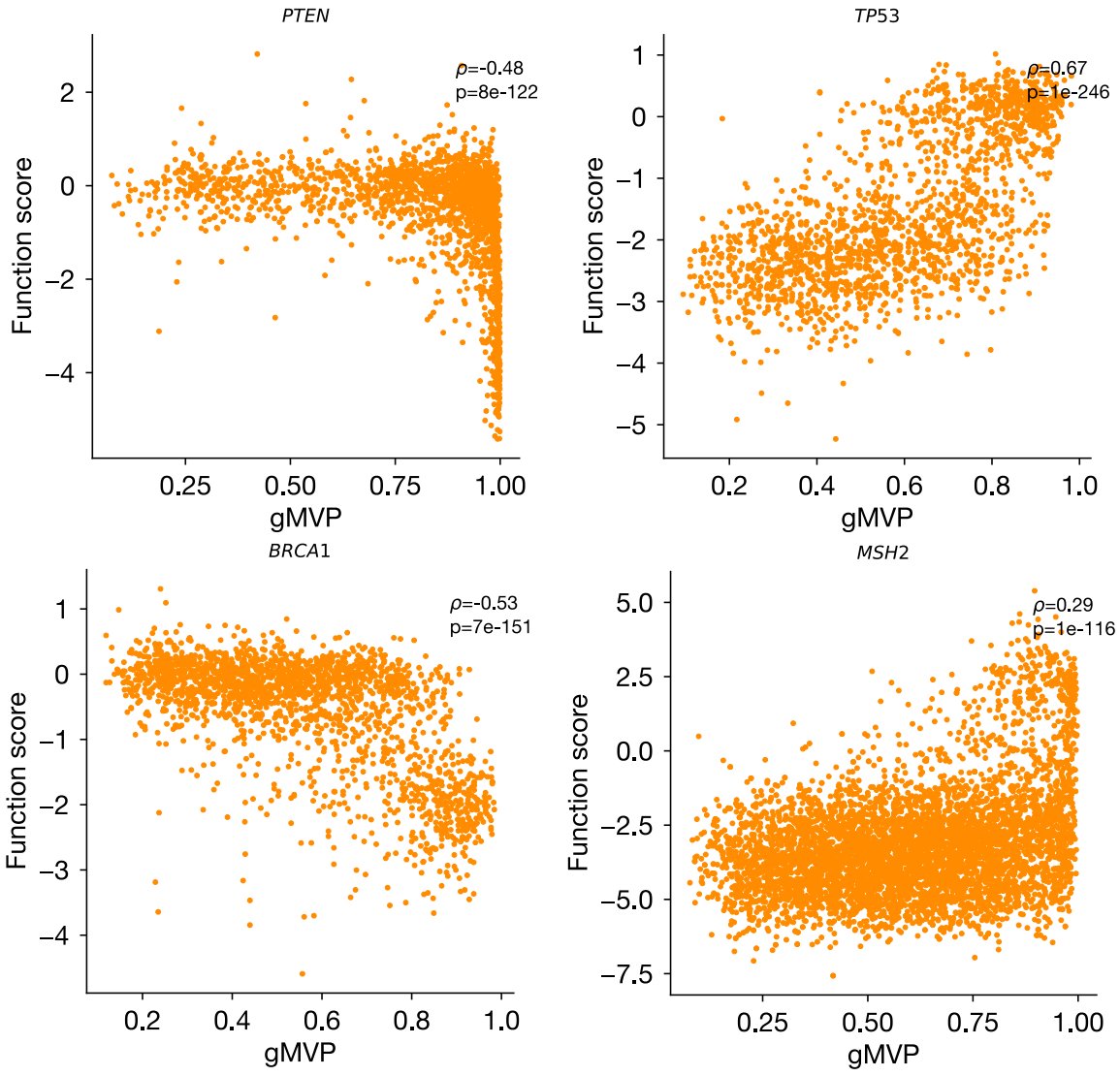
A number of studies have reported evidence that functionally damaging missense variants are clustered in 3-dimensional protein structures. Coevolution can explicitly capture residue dependencies between positions and recent published methods EVmutation[20] and PIVOTAL[21] have shown that coevolution helps to improve the prediction accuracy. PIVOTAL is a supervised ensemble predictor which combines coevolution between positions with prediction scores from other existing methods such as SIFT, and M-CAP, while EVmutation is an unsupervised model which learns coevolution and conservation using Markov Random Fields from MSAs.



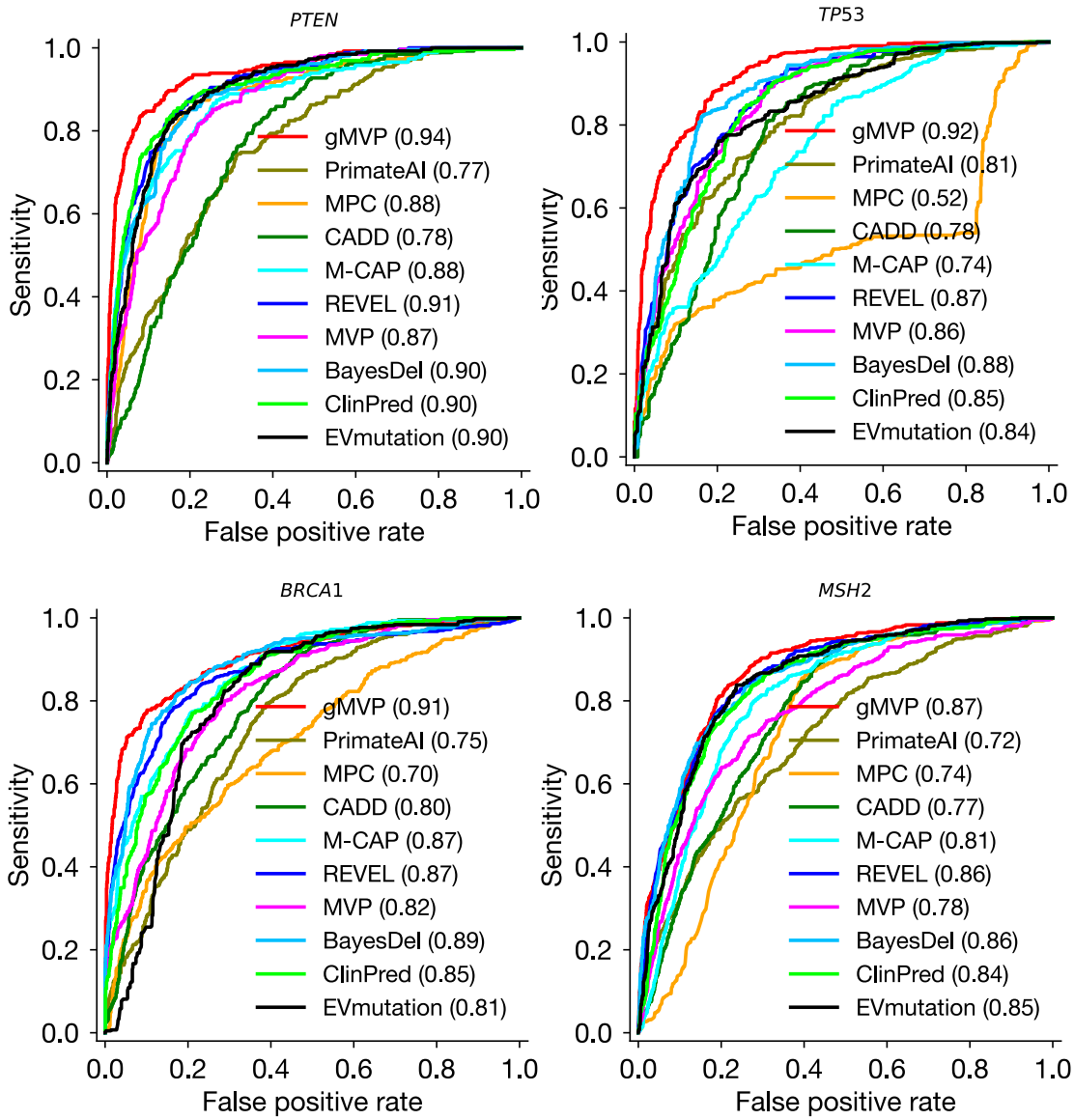

**Supplementary Figure 1. The model architecture of gMVP.** We first build a graph to represent a missense variant and its protein context defined as 128 amino acids flanking the amino acid of interest. The amino acid of interest is the center node (colored orange) and the flanking amino acids are the context nodes (colored light green). All context nodes are connected with the center node but not with each other. We use coevolution strength as edge features and used conservation and structural properties as features for both center node and context nodes. We additionally include amino acid substitution as features for center node and primary sequence and the expected and observed number of rare missense variants in the general population for context nodes. The input feature vectors for edges, center node, and context nodes are denoted as $f_k$, $x$, and $n_k$, respectively. We apply three 1-depth dense layers to encode the input feature vectors $f_k$, $x$, and $n_k$ to latent vectors $e_k$, $h$, and $t_k$, respectively. We next use a multi-head attention layer to learn a context vector c. We then use a gated recurrent neural layer to leverage the context vector c and the latent vector of the variant node $h$. We finally used a sigmoid layer to perform classification.



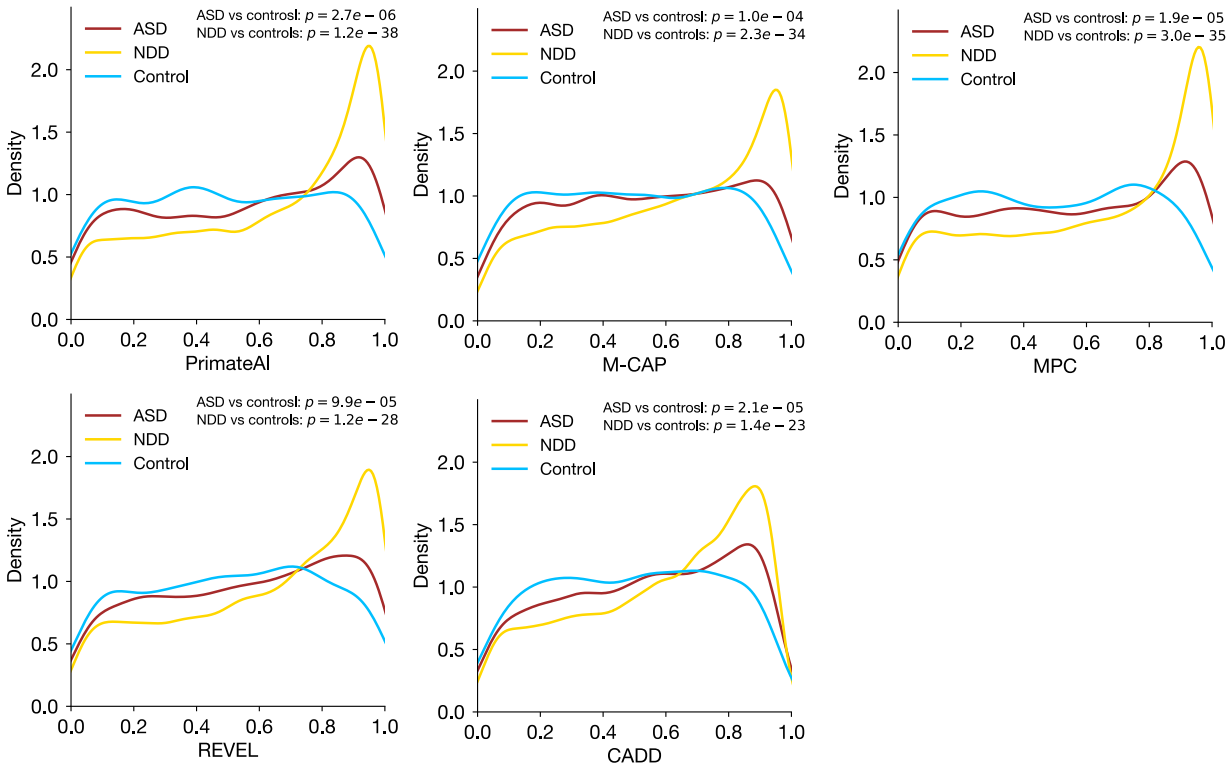

**Supplementary Figure 2. The distributions of gMVP scores of the damaging *(labeled positives)* and neutral variants *(labeled negatives)* on known disease genes, including *TP53*, *PTEN*, *BRCA1*, and *MSH2*.** The labels are determined by functional readout data of deep mutational scan experiments.



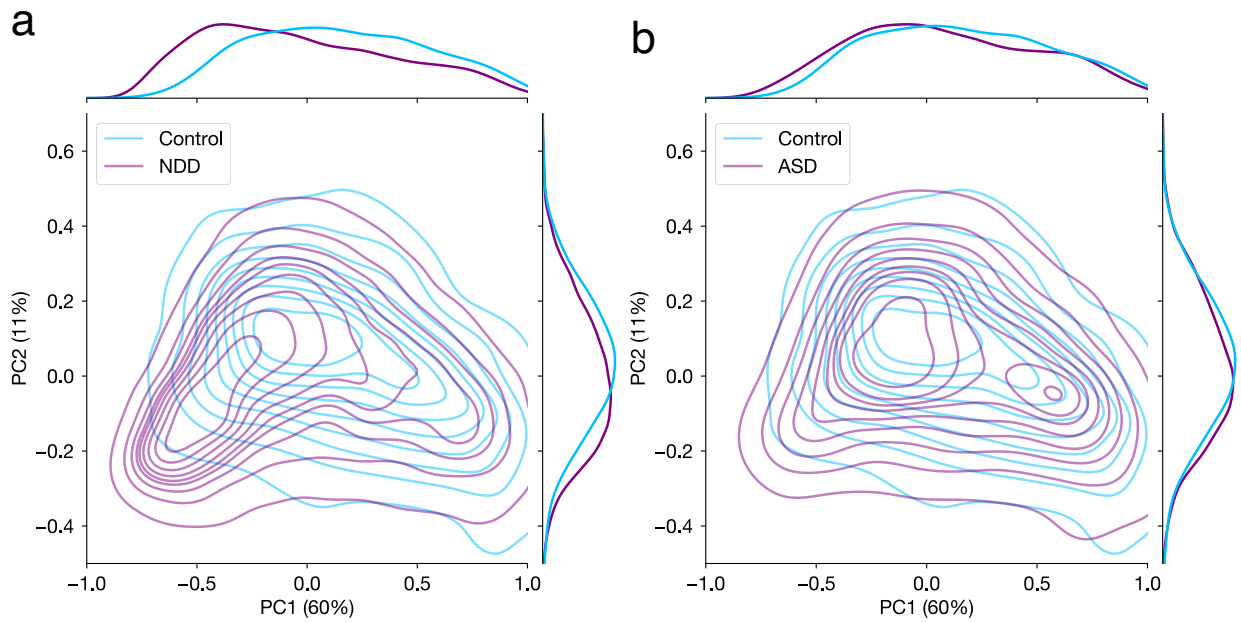

**Supplementary Figure 3. gMVP scores correlate with functional readout data from deep mutational scan experiments of known disease genes, including *PTEN*, *TP53*, *BRCA1*, and *MSH2*.** The correlation depends on the assay performed. The functional scores for *PTEN* and *BRCA1* correlate negatively, and the scores for *TP53* and *MSH2* correlate positively with the pathogenicity of the variants, respectively.



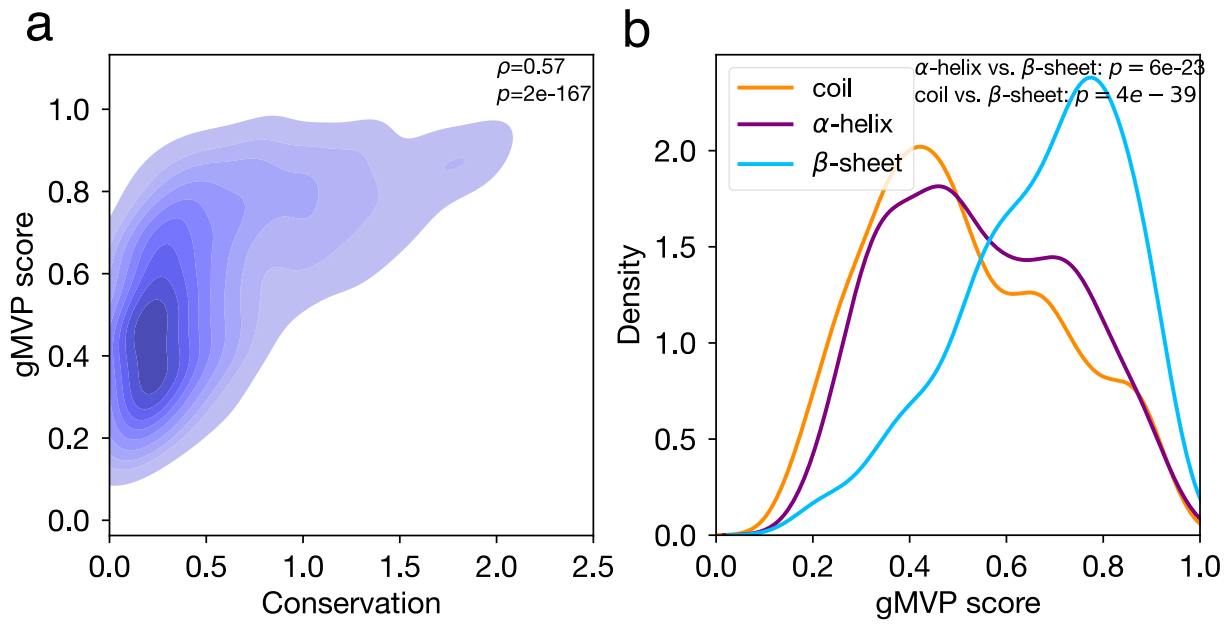

**Supplementary Figure 4. Evaluating gMVP and published methods in identifying damaging variants on known disease genes, including *TP53*, *PTEN*, *BRCA1*, and *MSH2*.** The receiver operating characteristic curves (ROC) of gMVP and published methods are shown for each gene using functional readout data as ground truth. There are 432 positives and 1476 negatives in *BRCA1*, 258 positives and 1601 negatives in *PTEN*, 540 positives and 1108 negatives in *TP53*, and 414 positives and 5439 negatives in *MSH2*.



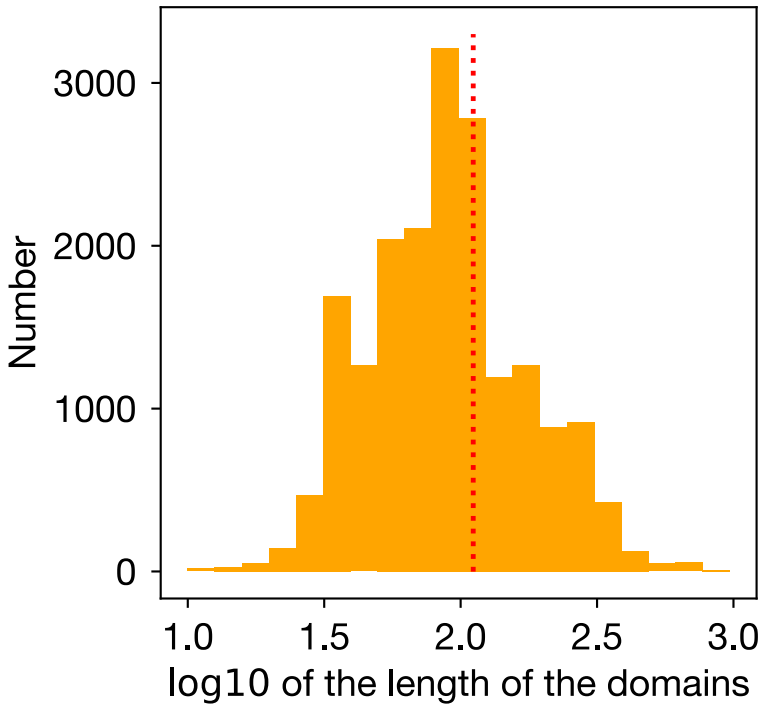

**Supplementary Figure 5. Distributions of predicted scores of published methods of rare *de novo* missense variants from ASD and NDD cases and controls**. We used two-sided Mann–Whitney U test to assess the statistical significance of the difference between cases and controls. NDD: neural developmental disorders; ASD: autism spectrum disorder; controls: unaffected siblings from the ASD study. Number of *de novo* missense variants compared: ASD: 2,913; NDD: 17,964; controls: 927. Number of trios in the data: ASD: 5,924; NDD: 31,058; controls: 31058.



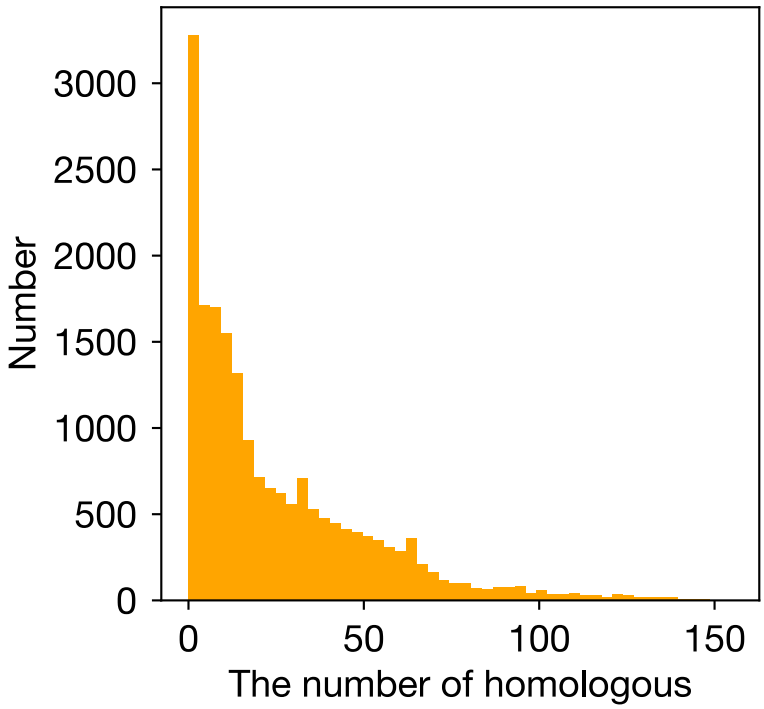

**Supplementary Figure 6. Principal component analysis of *de novo* missense variants in ASD and controls.** We performed principal component analysis (PCA) on the *de novo* variants from cases and controls. The input of the PCA is a score matrix where each row represents a variant and each column represents the predicted score of gMVP or other methods. The contours show the distribution of PC1/2 scores of the variants in cases and controls. The density curves along the axes show the distribution of PC1 or PC2 scores of cases and controls. **(a)** PC1 versus PC2 of *de novo* variants from NDD cases and controls. **(b)** PC1 versus PC2 of *de novo* variants from ASD cases and controls.



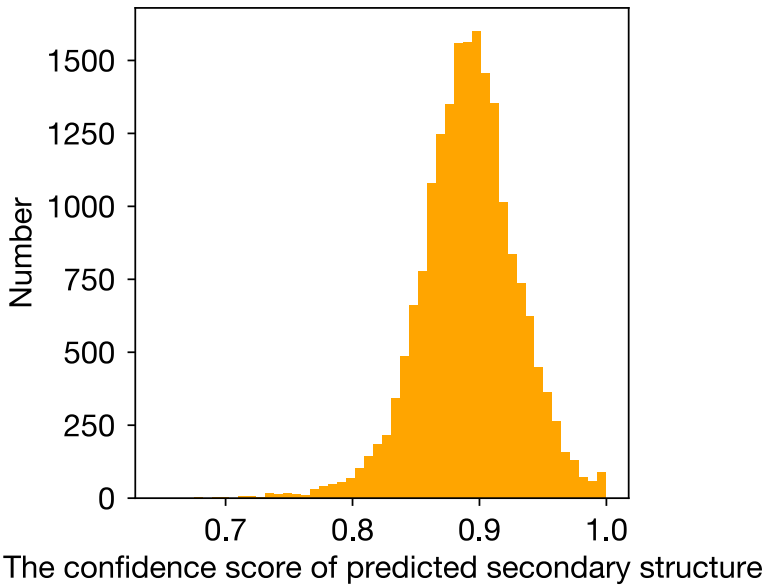

**Supplementary Figure 7. gMVP scores correlate with evolutionary conservation and protein secondary structure**. We show gMVP scores of all possible missense variants in *BRCT2* domain of *BRCA1*. We measured the evolutionary conservation for each protein position with the Kullback–Leibler divergence between amino acid distribution among homologous sequences and amino acid distribution in nature. We obtained the secondary structures using the solved protein structure of *BRCT2* domain. (**a**) gMVP scores versus evolutionary conservation. (b) Distributions of gMVP scores of variants located on the coils, $\alpha$-helices, and $\beta$-sheets, respectively. We used two-sided Mann–Whitney U test to assess the statistical significance of the difference between the gMVP score of variants on the $\beta$-sheet and on the coils and $\alpha$-helix regions.



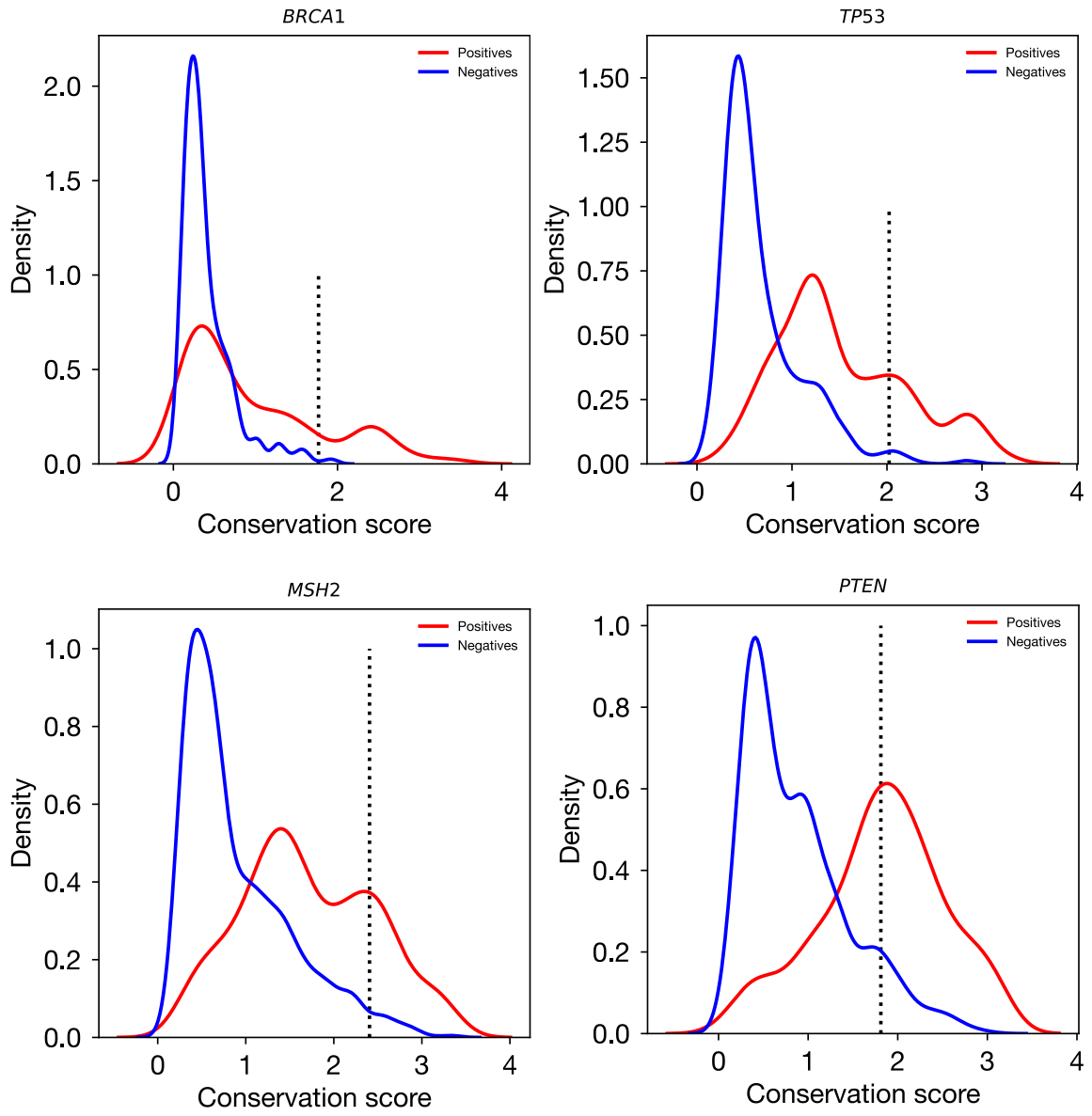

**Supplementary Figure 8. The distribution of the length of the domains of human proteins.** Here, we used 18,738 domains annotated with UniProtKB database. The average length of the domains is ~111 amino acids.



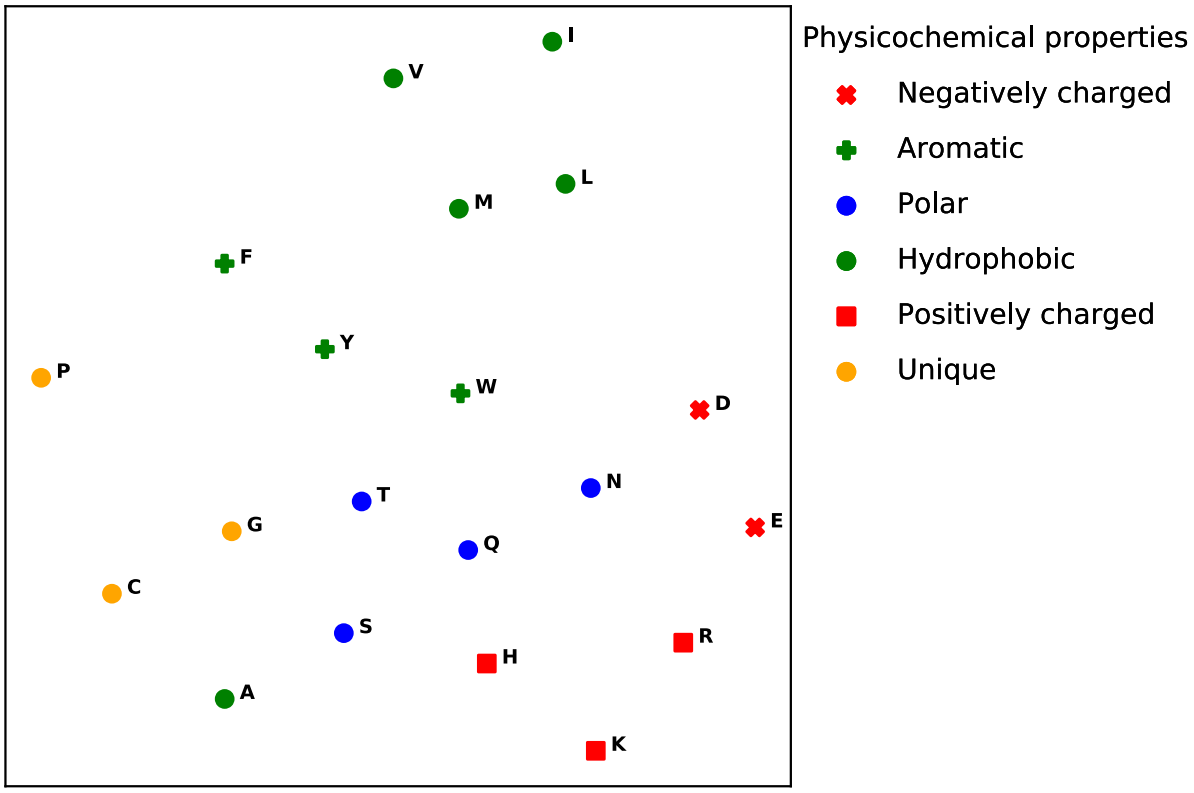

**Supplementary Figure 9. The distribution of the number of homologous of human proteins in Ensembl database.**



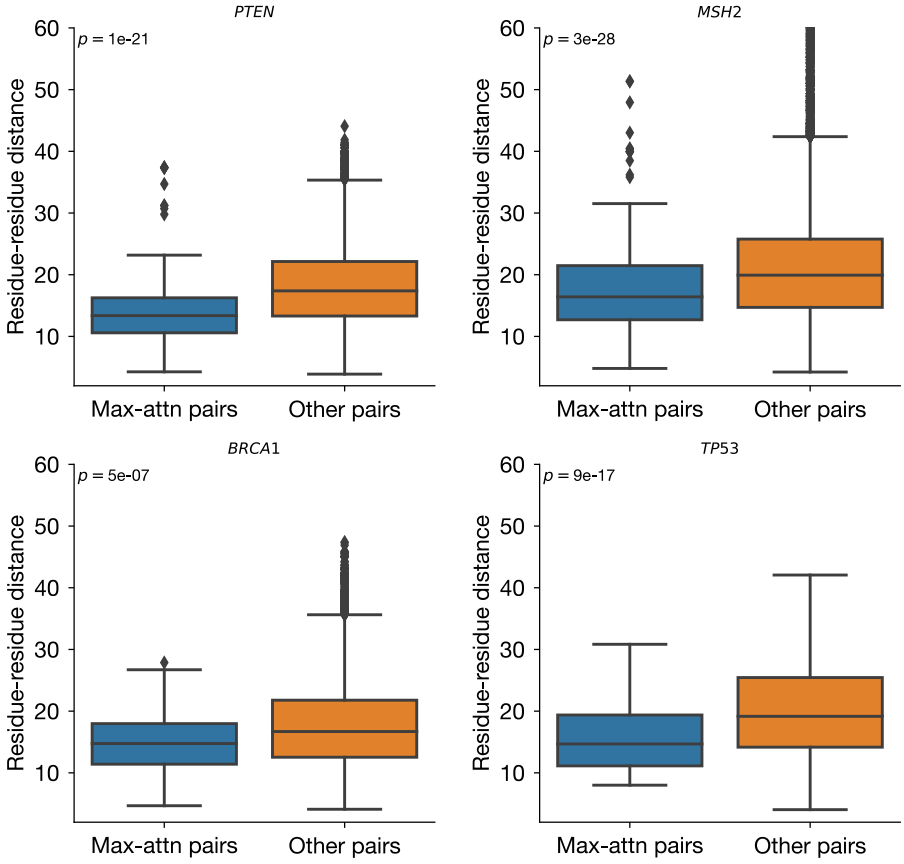

**Supplementary Figure 10. The distribution of confidence scores of the predicted secondary structures.** We predicted the secondary structures with NetSurfP2.



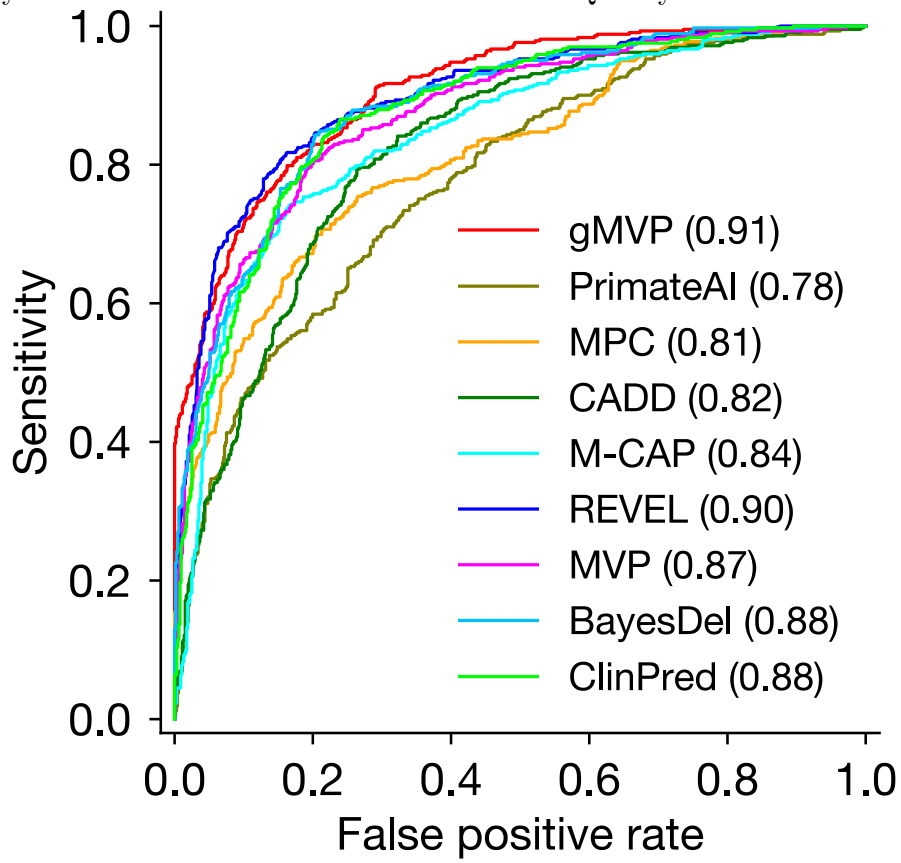

**Supplementary Figure 11. The distributions of conservation scores of the damaging *(labeled positive)* and neutral variants *(labeled negatives)* on known disease genes, including *TP53*, *PTEN*, *BRCA1*, and *MSH2*.** The labels are determined by functional readout data of deep mutational scan experiments. The vertical dotted lines in the figures show the false predicted positives with the highest gMVP scores. We measured the evolutionary conservation for each protein position with the Kullback–Leibler divergence between amino acid distribution among homologous sequences and amino acid distribution in nature.



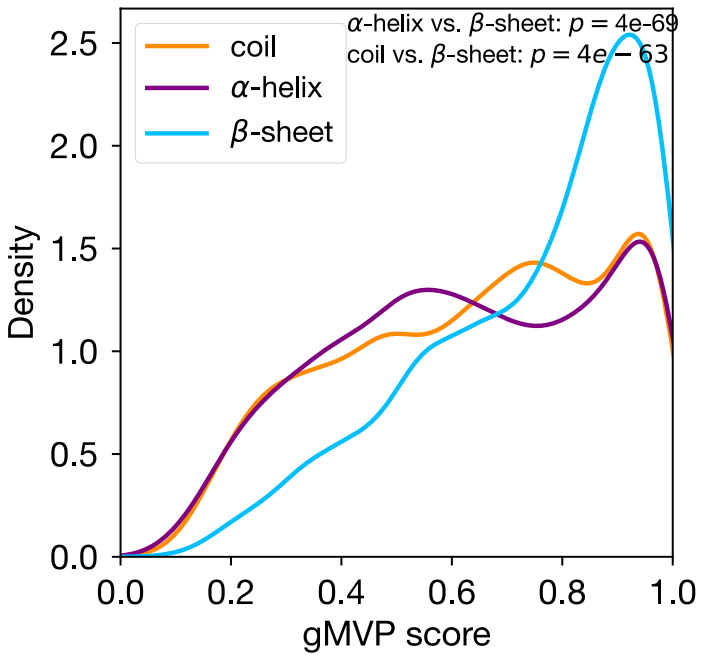

**Supplementary Figure 12. Physicochemical properties of amino acids are represented in the learned weights of the first layer of center node, visualized here with t-SNE.** Residues are clustered into hydrophobic, polar, and aromatic groups and reflect overall organization by molecular charge. The submatrix (20×256) of the learned weights matrix which projects the input one-hot encoding of the alternate amino acids describes how the model represents the amino acids. We applied t-SNE only to this submatrix, each row representing embedding vector of an amino acid type.


PTEN
p = 1e-21
MSH2
p = 3e-28
BRCA1
p = 5e-07
TP53
p = 9e-17
Residue-residue distance
60
50
40
30
20
10
Max-attn pairs
Other pairs

**Supplementary Figure 13. The residue-residue distances between the pairs with maximum attention weights differs with residue-residue distances of other pairs.** We first averaged the attention weights of all heads. For each gene, we selected top 10% variants with most damaging gMVP predictions. For each variant, we calculated the residue-residue distances between the position of interest and the neighbor positions. We then do the boxplots for the positions with highest attention weights and other positions, separately. We used two-sided Mann–Whitney U test to assess the statistical significance of the difference between two groups. The number of pairs with maximum attention weights and the number of other pairs are 209 and 2170, 563 and 6014, 137 and 1460, and 188 and 1726 for *PTEN*, *MSH2*, *BRCA1*, and *TP53*, respectively. The box bounds the interquartile range (IQR) divided by the median, and Tukey-style whiskers extend to a maximum of 1.5× IQR beyond the box.




**Supplementary Figure 14. Evaluating gMVP and published methods only on the tumor suppressor genes using cancer somatic mutation hotspots and random variants in population**. The ROC curves are evaluated on 422 cancer mutations located in hotspots and 713 rare variants from the DiscovEHR data.




**Supplementary Figure 15. gMVP scores correlate with protein secondary structures. We mapped the gMVP scores of *MSH2*, *PTEN*, *TP53*, and *BRCA1* to the secondary structures calculated from the solved 3D structures.** We plot distributions of gMVP scores of variants located on the coils, $\alpha$-helices, and $\beta$-sheets, respectively. We used Mann–Whitney U test to assess the statistical significance of the difference between the gMVP score of variants on the coils, the $\beta$-sheets, and $\alpha$-helices regions.

**Supplementary Table 11.** Statistical testing on the differences between ROCs of gMVP and other methods with DMS data.

| Gene | PrimateAI | MPC | CADD | M-CAP | REVEL | MVP | BayesDel | ClinPred | EVmutation |
|---|---|---|---|---|---|---|---|---|---|
| *BRCA1* | 3.9E-36 | 2.8E-43 | 4.6E-26 | 8.4E-05 | 1.7E-04 | 3.0E-14 | 2.7E-02 | 4.7E-07 | 4.4E-19 |
| *MSH2* | 1.6E-37 | 3.1E-52 | 1.1E-24 | 2.9E-13 | 8.1E-02 | 1.5E-20 | 2.1E-02 | 8.5E-05 | 8.9E-04 |
| *PTEN* | 3.7E-38 | 1.6E-13 | 1.5E-42 | 2.3E-11 | 2.1E-07 | 1.9E-15 | 8.7E-09 | 3.3E-06 | 8.9E-06 |
| *TP53* | 8.6E-27 | 1.8E-107 | 1.1E-38 | 2.8E-44 | 5.9E-12 | 7.1E-19 | 1.0E-09 | 1.1E-23 | 2.5E-16 |

**Supplementary Table 12.** Evaluating additional published methods with DMS data.

| | *BRCA1* | | *MSH2* | | *PTEN* | | *TP53* | |
|---|---|---|---|---|---|---|---|---|
| | auPR | auROC | auPR | auROC | auPR | auROC | auPR | auROC |
| SIFT | 0.65 | 0.75 | 0.28 | 0.83 | 0.62 | 0.88 | 0.69 | 0.77 |
| Polyphen2 | 0.62 | 0.84 | 0.3 | 0.79 | 0.4 | 0.77 | 0.71 | 0.82 |
| GERP++ | 0.42 | 0.74 | 0.13 | 0.66 | 0.15 | 0.51 | 0.44 | 0.67 |
| phastCons100way | 0.61 | 0.68 | 0.12 | 0.56 | 0.57 | 0.5 | 0.67 | 0.67 |
| Eigen-PC-raw | 0.61 | 0.84 | 0.25 | 0.82 | 0.46 | 0.82 | 0.6 | 0.79 |

**Supplementary Table 13.** Area under ROC curves (AUROC) and PR curves (AUPRC) of gMVP and other in-silico methods on the variants of uncertain significance in ClinVar overlapped with the deep mutational data. The best AUROC and AUPRC values are in bold. The difference between ROC curves of gMVP and the other methods is tested using DeLong test. *P*-values lower than 0.05, 0.01 or 0.001 are marked with *, **, or ***, respectively.

| **Methods** | ***BRCA1*** **(1451 B + 334 P)** | | ***TP53*** **(306 B + 247 P)** | | ***PTEN*** **(262 B + 54 P)** | | ***MSH2*** **(1587 B + 115 P)** | |
|---|---|---|---|---|---|---|---|---|
| | AUROC | AUPRC | AUROC | AUPRC | AUROC | AUPRC | AUROC | AUPRC |
| **gMVP** | **0.896** | **0.765** | **0.906** | **0.859** | **0.952** | **0.845** | 0.855 | **0.372** |
| **BayesDel** | 0.874 * | 0.677 | 0.851 *** | 0.772 | 0.896 *** | 0.638 | 0.857 | 0.369 |
| **REVEL** | 0.870 ** | 0.658 | 0.833 *** | 0.787 | 0.903 *** | 0.586 | **0.863** | 0.359 |
| **VEST4** | 0.890 *** | 0.689 | 0.839 *** | 0.788 | 0.882 ** | 0.646 | 0.840 | 0.352 |
| **ClinPred** | 0.845 *** | 0.558 | 0.820 *** | 0.753 | 0.908 ** | 0.762 | 0.853 | 0.256 |
| **SIFT** | 0.778 *** | 0.366 | 0.832 *** | 0.750 | 0.873 *** | 0.533 | 0.835 | 0.214 |
| **M-CAP** | 0.848 *** | 0.620 | 0.739 *** | 0.697 | 0.858 *** | 0.558 | 0.787 *** | 0.161 |
| **MVP** | 0.815 *** | 0.522 | 0.820 *** | 0.745 | 0.869 *** | 0.476 | 0.771 *** | 0.204 |
| **MPC** | 0.696 *** | 0.349 | 0.545 *** | 0.597 | 0.906 *** | 0.677 | 0.742 *** | 0.124 |
| **PrimateAI** | 0.730 *** | 0.369 | 0.791 *** | 0.721 | 0.815 *** | 0.411 | 0.712 *** | 0.194 |
| **CADD** | 0.799 *** | 0.434 | 0.746 *** | 0.613 | 0.805 *** | 0.350 | 0.818 ** | 0.178 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Polyphen2** | 0.810 *** | 0.416 | 0.794 *** | 0.680 | 0.789 *** | 0.380 | 0.775 *** | 0.158 |

**Supplementary Table 14**. Hyperparameter settings in gMVP model.

| Description | Value |
|---|---|
| The number of heads in attention layer | 8 |
| The number of neuron units in attention layer | 256 |
| The number of neuron units in GRU layer | 256 |
| The activation function | RELU |

**Supplementary Table 15**. Explained variances in the principal component analysis for *de novo* variants of NDD, ASD, and controls. We normalized the scores using the method of z-score normalization. We used python package of scikit-learn to perform the PCA analysis.

| Component | Explained variance |
|---|---|
| 1st component | 60% |
| 2nd component | 11% |

**Supplementary Table 16**. The number of overlapping variants between DMS data and DiscovEHR data.

| Gene | Overlapping positives | overlapping negatives | Ratio of overlapping positives |
|---|---|---|---|
| *BRCA1* | 9 | 36 | 0.2 |
| *TP53* | 24 | 21 | 0.53 |
| *PTEN* | 3 | 26 | 0.1 |
| *MSH2* | 6 | 215 | 0.03 |
| *Total* | 42 | 298 | 0.12 |

**Supplementary Table 17.**

The differences of residue-residue distances between the pairs with maximum attention weights and other pairs.

| Gene | PDB code | Stuctural coverage | Mean distance of Max-attn pairs | Mean distance of other pairs | *p*-value |
|---|---|---|---|---|---|
| PTEN | *1D5R* | 0.76 | 14.2 | 18.1 | 1.00E-21 |
| MSH2 | *3THX* | 0.93 | 17.3 | 21.4 | 3.00E-28 |
| TP53 | *1TSR* | 0.5 | 15.6 | 20 | 9.00E-17 |
| BRCA1 | *4IGK* | 0.11 | 14.5 | 17.7 | 5.00E-07 |

**Supplementary Table 18**

Evaluating gMVP on the variants located near 3' or 5' end with DMS data.

| Gene | number of positives near the 3' or 5' end | number of negatives near the 3' or 5' end | auROC | auROC for the variant on the entire protein |
|---|---|---|---|---|
| *BRCA1* | 136 | 508 | 0.92 | 0.91 |
| *MSH2* | 13 | 724 | 0.78 | 0.87 |
| *PTEN* | 61 | 468 | 0.96 | 0.94 |
| *TP53* | 0 | 0 | - | 0.92 |

## Description of other supplementary tables

**S1**. Summary statistics of training data sets.

**S2.** Somatic mutations in cancer hotspots and random variants from DiscovEHR with annotations.

**S3-S6.** Variants with functional readout data from deep mutational scan experiments of *BRCA1*, *TP53*, *MSH2*, and *PTEN*.

**S7.** NDD *de novo* variants enrichment by various methods at rank percentile thresholds

**S8.** ASD *de novo* variants enrichment by various methods at rank percentile thresholds

**S9.** Pathogenetic and neutral variants in ion channel genes and the annotations.

**S10.** GOF and LOF variants in ion channel genes and the annotations.

## References


1. Findlay, G.M. *et al.* Accurate classification of BRCA1 variants with saturation genome editing. *Nature* **562**, 217-+ (2018).
2. Samocha, K.E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nature Genetics* **46**, 944-+ (2014).
3. Homsy, J. *et al.* De novo mutations in congenital heart disease with neurodevelopmental and other congenital anomalies. *Science* **350**, 1262-1266 (2015).
4. Heyne, H.O. *et al.* De novo variants in neurodevelopmental disorders with epilepsy. *Nature Genetics* **50**, 1048-+ (2018).
5. Rehm, H.L., Berg, J.S. & Plon, S.E. ClinGen and ClinVar - Enabling Genomics in Precision Medicine. *Human Mutation* **39**, 1473-1475 (2018).
6. Landrum, M.J. & Kattman, B.L. ClinVar at five years: Delivering on the promise. *Human Mutation* **39**, 1623-1630 (2018).
7. Zuk, O. *et al.* Searching for missing heritability: Designing rare variant association studies. *Proceedings of the National Academy of Sciences of the United States of America* **111**, E455-E464 (2014).
8. Davydov, E.V. *et al.* Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP plus. *Plos Computational Biology* **6**(2010).
9. Kumar, P., Henikoff, S. & Ng, P.C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature protocols* **4**, 1073 (2009).
10. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research* **15**, 1034-1050 (2005).
11. Adzhubei, I., Jordan, D.M. & Sunyaev, S.R. Predicting functional effect of human missense mutations using PolyPhen-2. *Current protocols in human genetics* **76**, 7.20. 1-7.20. 41 (2013).
12. Qi, H. *et al.* MVP predicts the pathogenicity of missense variants by deep learning. *Nature Communications* **12**, 510 (2021).
13. Samocha, K.E. *et al.* Regional missense constraint improves variant deleteriousness prediction. *bioRxiv*, 148353 (2017).
14. Havrilla, J.M., Pedersen, B.S., Layer, R.M. & Quinlan, A.R. A map of constrained coding regions in the human genome. *Nature Genetics* **51**, 88-+ (2019).
15. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics* **46**, 310-+ (2014).
16. Ioannidis, N.M. *et al.* REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *American Journal of Human Genetics* **99**, 877-885 (2016).
17. Alirezaie, N., Kernohan, K.D., Hartley, T., Majewski, J. & Hocking, T.D. ClinPred: prediction tool to identify disease-relevant nonsynonymous single-nucleotide variants. *The American Journal of Human Genetics* **103**, 474-483 (2018).
18. Jagadeesh, K.A. *et al.* M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nature genetics* **48**, 1581 (2016).

19. Sundaram, L. *et al.* Predicting the clinical impact of human mutation with deep neural networks. *Nature Genetics* **50**, 1161-+ (2018).
20. Hopf, T.A. *et al.* Mutation effects predicted from sequence co-variation. *Nature Biotechnology* **35**, 128-135 (2017).
21. Liang, S., Mort, M., Stenson, P.D., Cooper, D.N. & Yu, H. PIVOTAL: Prioritizing variants of uncertain significance with spatial genomic patterns in the 3D proteome. *bioRxiv*, 2020.06.04.135103 (2021).