Supplementary Materials

MVP: predicting pathogenicity of missense variants by deep learning
Qi et. al.

Table of Contents

Supplementary Figures ***2***

Supplementary Figure S1 ..... **2**

Supplementary Figure S2. ..... **3**

Supplementary Figure S3 ..... **4**

Supplementary Figure S4 ..... **5**

Supplementary Figure S5 ..... **6**

Supplementary Figure S6 ..... **7**

Supplementary Figure S7 ..... **8**

Supplementary Figure S8 ..... **9**

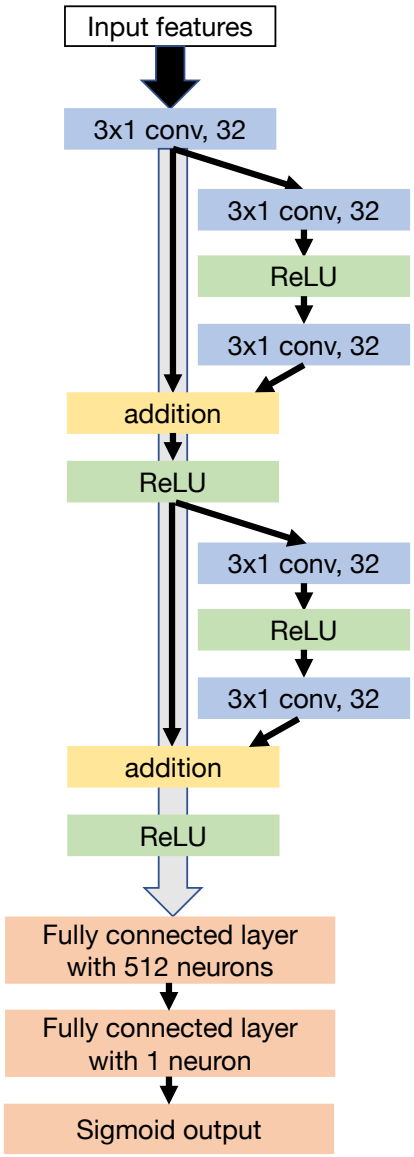
Supplementary Tables ***11***

Supplementary Table S1. ..... **11**

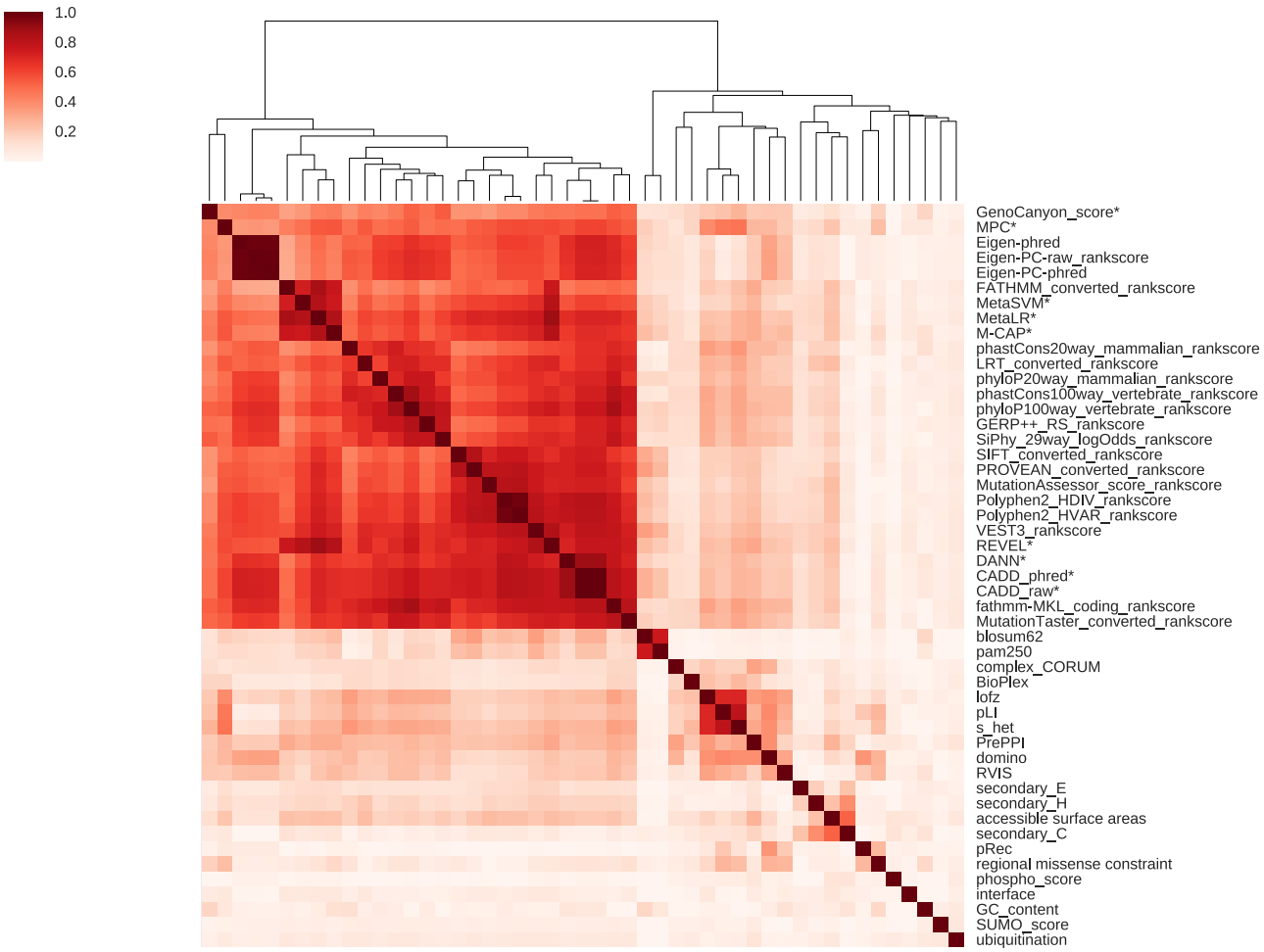
Supplementary Table S10. Comparison of cases and controls in rate of synonymous *de novo* variants **13**

Supplementary Notes ***15***

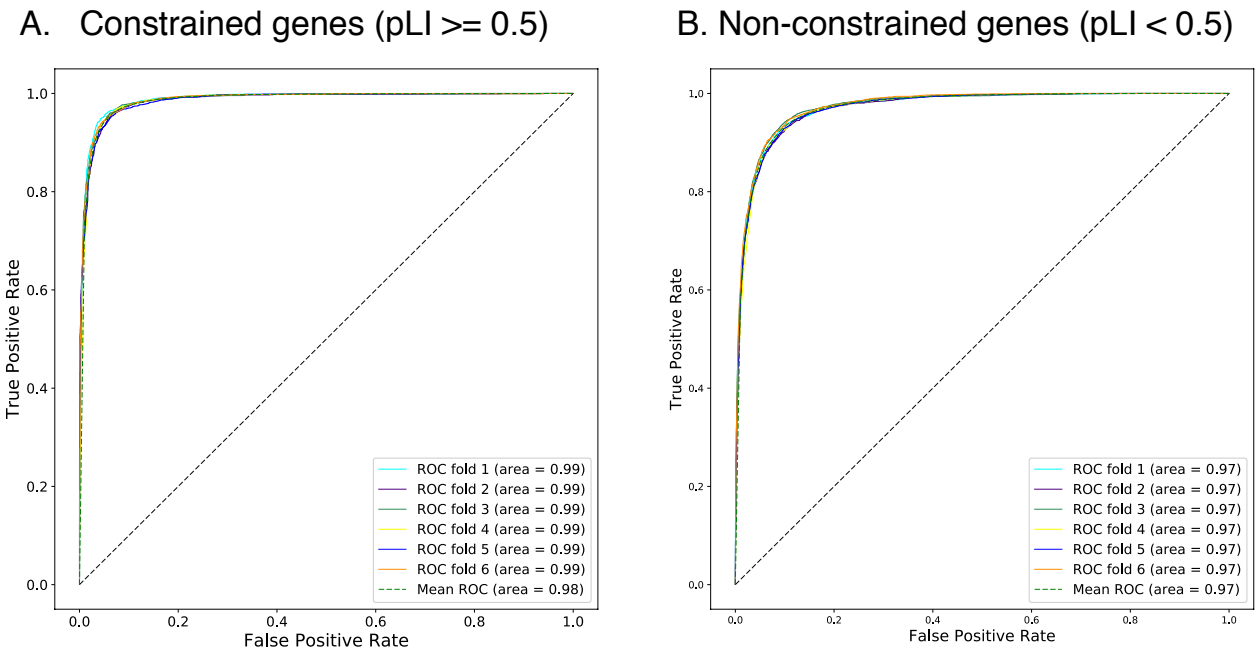
Supplementary Figures



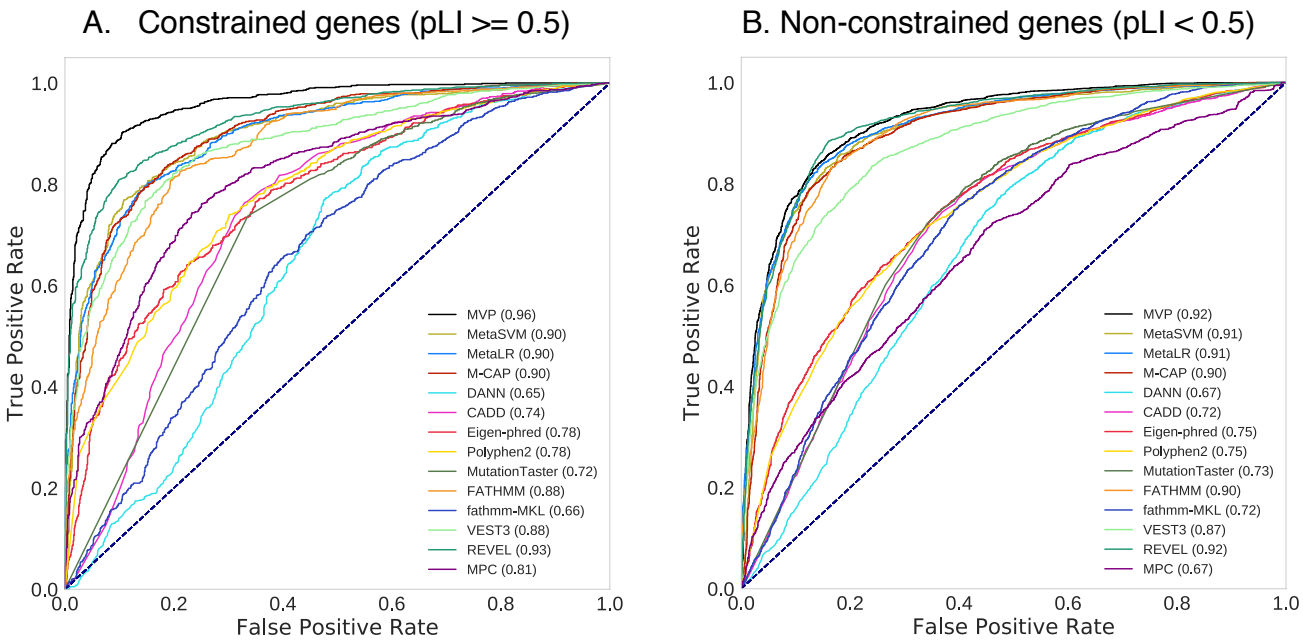
Supplementary Figure S1. The ResNet neural network architecture of MVP. Building blocks are arranged as shown in the figure. Parameters and dimensions of input and output are indicated in the boxes. Blue boxes are convolutional filters, green boxes are ReLU activation, yellow boxes are addition of output from 2 layers, orange boxes are fully connected layers.



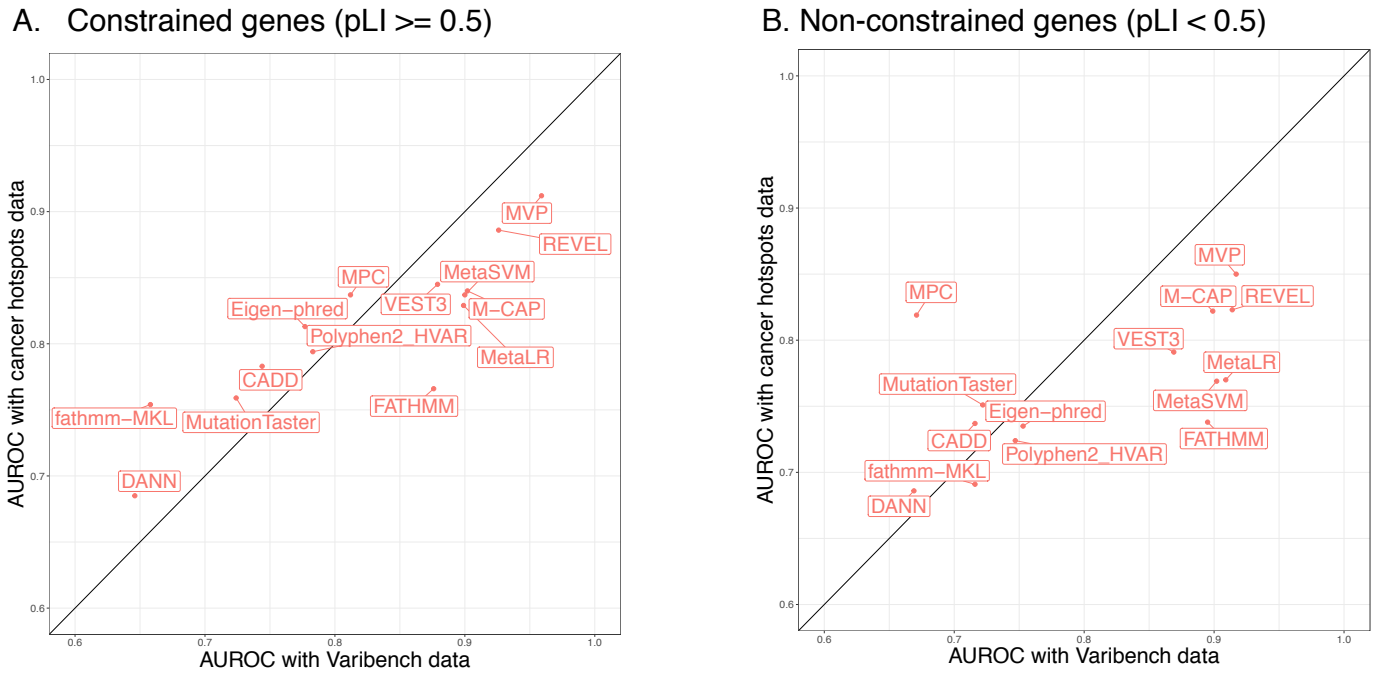
Supplementary Figure S2. Corre and hierarchical clustering of features and additional published methods. We calculated pairwise Spearman correlation of all features and additional published methods across data points used in the training**.** Color key indicates absolute value of Spearman correlation coefficient among features and predictors. Columns are ordered by hierarchical clustering. Published methods marked with * are not used in training.



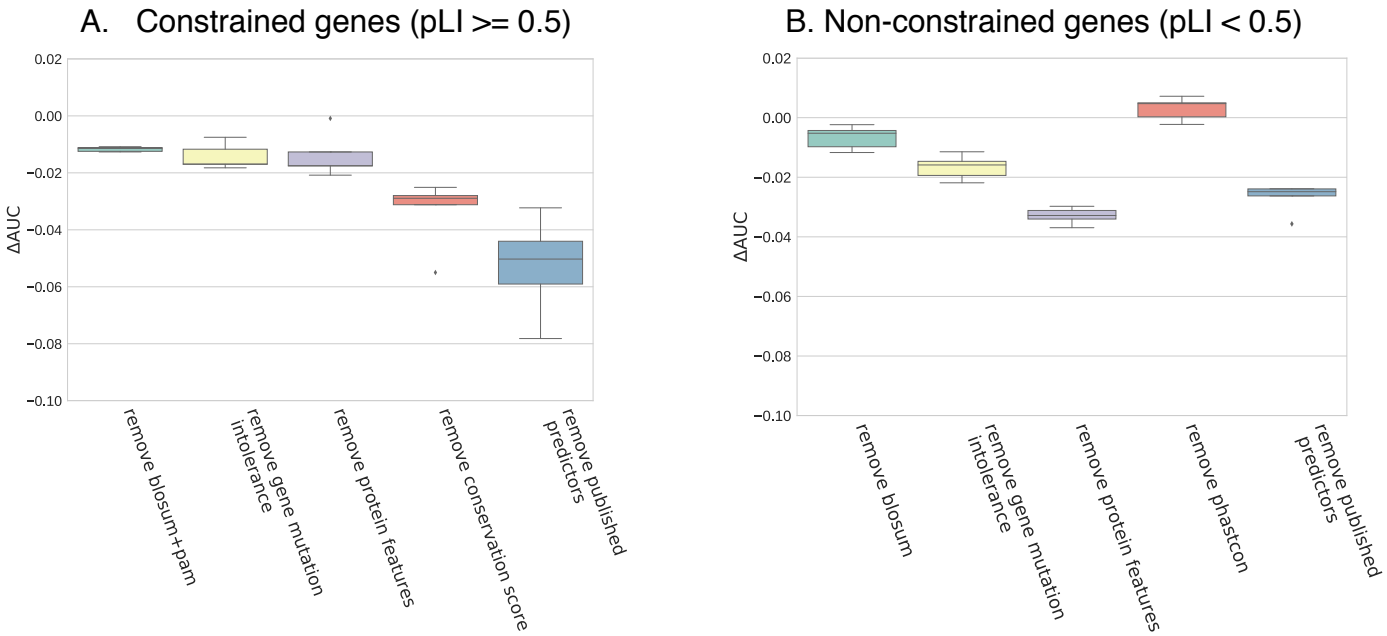
Supplementary Figure S3. Receiver operating characteristic (ROC) curves of MVP with 6-fold cross validation in the training dataset. (A) Performance evaluation in constrained genes (ExAC pLI $\geq$ 0.5). (B) Performance evaluation in non-constrained genes (ExAC pLI $<$ 0.5). The performance of MVP in each fold is evaluated by the ROC curve and Area Under Curve (AUC) score indicated in parenthesis. Higher AUC score indicates better performance.



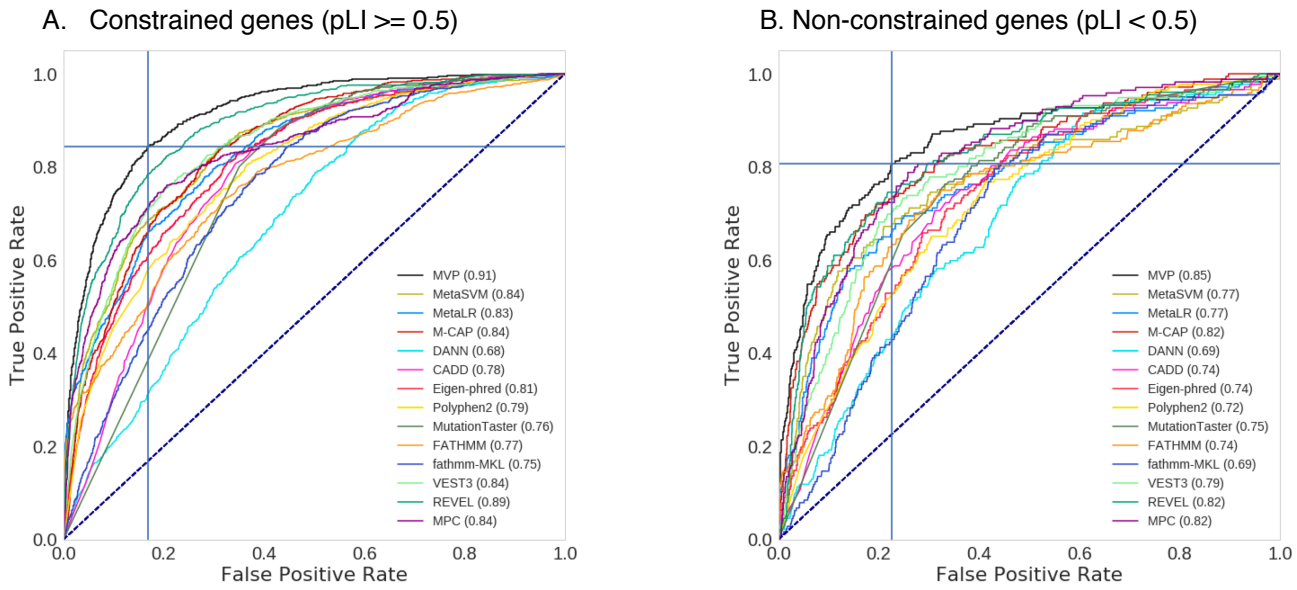
Supplementary Figure S4. Comparing MVP with previous methods by ROC curves using VariBench testing data. (A) Performance evaluation in constrained genes. (B) Performance evaluation in non-constrained genes. The performance of each method is evaluated by the ROC curve and AUC score indicated in parenthesis. Higher AUC score indicates better performance.



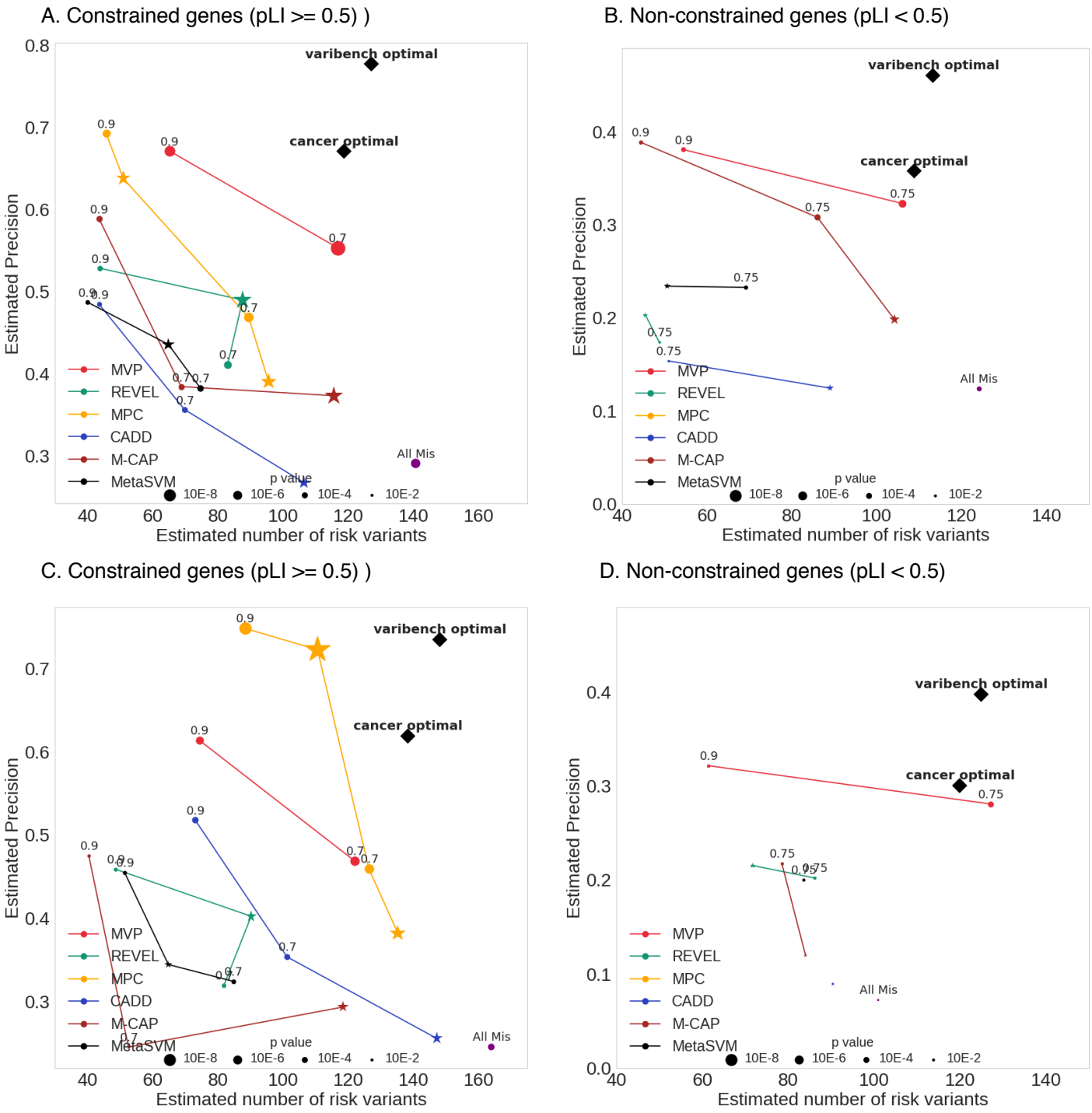
Supplementary Figure S5. Comparison of AUC using VariBench data versus cancer mutation hotspots data for MVP and previous methods. X-axis indicates the AUC with VariBench data; y-axis indicates the AUC with cancer hotspots data. (A) comparison in constrained genes. (B) comparison in non-constrained genes.



Supplementary Figure S6. Measuring the contribution of features to MVP prediction performance in cancer mutation hotspots data. Performance contribution is measured by AUC reduction (Δ AUC) from excluding a group of features. Since features within a group is often highly correlated, we did measure the contribution of an entire group instead of individual features in the group. (A) Constrained genes; (B) Non-constrained genes. Error bar is estimated by subsampling of large number of negatives.



Supplementary Figure S7. Optimal threshold of MVP score based on ROC curve using cancer somatic mutation hotspots data. Horizontal line and vertical line indicated the optimal threshold in which the ROC curve has the maximum distance to the diagonal line; (A) Constrained genes: MVP score 0.7 is best threshold; (B) Non-constrained genes: MVP score 0.75 is best threshold.



Supplementary Figure S8. Comparison of MVP and previous methods using *de novo* missense mutations from CHD and ASD studies by precision-recall-like curves. Numbers on each point indicate rank percentile thresholds; star points indicate thresholds recommended by publications. The position of "All Mis" points are estimated from all missense variants in the gene set without using any pathogenicity prediction method, ***black diamonds*** indicate estimated precision and

number of variants from cancer hotspot ROC curve and VariBench ROC curve. The size of each point is proportional to –log(p-value). P-value is calculated by Binomial test, and only points with p-value < 0.05 are shown. (A, B) Performance in CHD *de novo* data in constrained genes and non-constrained genes, respectively. (D, E) Performance in ASD *de novo* data in constrained genes and non-constrained genes, respectively.

Supplementary Tables

Supplementary Table S1.

Estimated number of pathogenic missense *de novo* mutations using published methods by recommended thresholds. The table indicates their thresholds, estimated number of risk variants and positive predictive values in Congenital heart disease and Autism spectrum disorder data.

a) Evaluation among all genes

		Congenital heart disease (CHD)		Autism spectrum disorder (ASD)	
	Threshold	Estimated number of risk variants	Estimated Precision	Estimated number of risk variants	Estimated precision
All missense	N/A	264	0.17	264	0.13
M-CAP	> 0.025	219	0.26	202	0.18
Meta-SVM	> 0	115	0.31	105	0.22
MutationTaster	> 0.5	187	0.18	201	0.14
Polyphen	> 0.5	170	0.22	183	0.17
SIFT	< 0.05	151	0.17	183	0.15
VEST3	> 0.8	115	0.28	134	0.24
CADD	> 15	195	0.17	237	0.15
REVEL	> 0.5	133	0.33	162	0.3

b) Evaluation among constrained genes (ExAC pLI≥0.5)

		Congenital heart disease (CHD)		Autism spectrum disorder (ASD)	
	Threshold	Estimated number of risk variants	Estimated precision	Estimated number of risk variants	Estimated precision
All missense	N/A	140	0.29	163	0.25
M-CAP	> 0.025	115	0.37	118	0.29
Meta-SVM	> 0	64	0.44	64	0.34
MutationTaster	> 0.5	102	0.25	134	0.23
Polyphen	> 0.5	85	0.32	113	0.30
SIFT	< 0.05	90	0.29	107	0.24
VEST3	> 0.8	83	0.44	96	0.39
CADD	> 15	106	0.27	147	0.26
REVEL	> 0.5	87	0.49	90	0.40

c) Evaluation among non-constrained genes (ExAC pLI<0.5)

		Congenital heart disease (CHD)		Autism spectrum disorder (ASD)	
	Threshold	Estimated number of risk variants	Estimated precision	Estimated number of risk variants	Estimated precision
All missense	N/A	124	0.12	101	0.07
M-CAP	> 0.025	104	0.20	84	0.11
Meta-SVM	> 0	50	0.23	40	0.14
MutationTaster	> 0.5	85	0.13	66	0.07
Polyphen	> 0.5	84	0.17	70	0.10
SIFT	< 0.05	61	0.11	76	0.09
VEST3	> 0.8	31	0.14	38	0.12
CADD	> 15	89	0.12	90	0.09
REVEL	> 0.5	45	0.2	71	0.21

Supplementary Table S2. (In separate data files) Features in the MVP model. The table lists details of features used in constrained genes model and non-constrained genes model group by different categories.

Supplementary Table S3. (In separate data files) Summary statistics of training and testing data sets. The table indicates number of genes and variants from different data sets used in training and testing. Genes are grouped as constrained genes (ExAC pLI≥0.5) and non-constrained genes (ExAC pLI<0.5)

Supplementary Table S4. (In separate data files) Performance comparison of different methods in VariBench dataset and Cancer hotspot dataset. The table indicated the AUC performance of different predictors in VariBench data and cancer hotspot data, genes are grouped as constrained gene (ExAC pLI>=0.5) and non-constrained gene (ExAC pLI<0.5).

Supplementary Table S5. (In separate data files) Number and percentage of genes and variants in testing datasets that are overlapped with genes used in training.

Supplementary Table S6. (In separate data files) CHD de novo variants D-mis enrichment using different methods by various rank percentile thresholds. The table indicates rank percentile threshold for each method, number of variants in cases and controls passing the criteria, enrichment, binomial test pvalue, estimated number of risk variants and positive predictive values and estimated recall.

Supplementary Table S7. (In separate data files) ASD de novo variants D-mis enrichment using different methods by various rank percentile thresholds. The table indicates rank percentile threshold for each method, number of variants in cases and controls passing the criteria, enrichment, binomial test pvalue, estimated number of risk variants and positive predictive values and estimated recall.

Supplementary Table S8. (In separate data files) Percentage of CHD isolated cases by damaging variants. We define damaging missense variants using various rank percentile thresholds, for metaSVM prediction we used recommended score of 0. The table indicates rank percentile threshold for each method, number of variants in cases and controls passing the criteria, enrichment, binomial test pvalue, estimated number of risk variants, positive predictive values, estimated recall, percentage of the cases explained by de novo loss of function variants and damaging missense variants with 95% confident interval.

Supplementary Table S9. (In separate data files) Predicted pathogenic missense variants in isolated CHD cases. There are 175 predicted pathogenic variants in isolated CHD cases. We selected variants located in constrained genes with MVP score larger than 0.7 and variants located in non-constrained genes with MVP score larger than 0.75 as pathogenic variants. The genomic position (hg19) of each variant and the alternative alleles were indicated. The function predicted values of each variant was given by CADD, metaSVM, M-CAP, MPC and REVEL. The rank percentile of function predicted values of each variant was given by CADD_rank, metaSVM_rank, M-CAP_rank, MPC_rank, REVEL_rank and MVP_rank. The higher rank value, the more likely to be pathogenic. The ExAC pLI value indicates gene intolerance, we only consider constrained genes with value of PLI≥0.5 and minor allele frequency (MAF) smaller than 1e-6 and non-constrained genes with value of pLI<0.5 and MAF smaller than 1e-4.

Supplementary Table S10. Comparison of cases and controls in rate of synonymous *de novo* variants

	Number of synonymous variants	Rate per cases compared to controls
Autism spectrum disorder (ASD)	1026	1.027
Congenital heart disease (CHD)	701	1.049
Simons Simplex Collection unaffected siblings (controls)	483	N/A

Supplementary Table S11. CHD *de novo* missense variants with annotation. The genomic position (hg19) of each variant and the alternative alleles were indicated. The

function predicted values of each variant was given by CADD, metaSVM, M-CAP, MPC and REVEL. The rank percentile of function predicted values of each variant was given by CADD_rank, metaSVM_rank, M-CAP_rank, MPC_rank, REVEL_rank and MVP_rank. The higher rank value, the more likely to be pathogenic. The ExAC pLI value indicates gene intolerance, we only consider constrained genes with value of PLI≥0.5 and minor allele frequency (MAF) less than 1e-6 and non-constrained genes with value of pLI<0.5 and MAF less than 1e-4.

Supplementary Table S12 ASD *de novo* missense variants with annotation The genomic position (hg19) of each variant and the alternative alleles were indicated. The function predicted values of each variant was given by CADD, metaSVM, M-CAP, MPC and REVEL. The rank percentile of function predicted values of each variant was given by CADD_rank, metaSVM_rank, M-CAP_rank, MPC_rank, REVEL_rank and MVP_rank. The higher rank value, the more likely to be pathogenic. The ExAC pLI value indicates gene intolerance, we only consider constrained genes with value of PLI≥0.5 and minor allele frequency (MAF) less than 1e-6 and non-constrained genes with value of pLI<0.5 and MAF less than 1e-4.

Supplementary Table S13 SSC control *de novo* missense variants with annotations. The genomic position (hg19) of each variant and the alternative alleles were indicated. The function predicted values of each variant was given by CADD, metaSVM, M-CAP, MPC and REVEL. The rank percentile of function predicted values of each variant was given by CADD_rank, metaSVM_rank, M-CAP_rank, MPC_rank, REVEL_rank and MVP_rank. The higher rank value, the more likely to be pathogenic. The ExAC pLI value indicates gene intolerance, we only consider constrained genes with value of PLI≥0.5 and minor allele frequency (MAF) less than 1e-6 and non-constrained genes with value of pLI<0.5 and MAF less than 1e-4.

Supplementary Notes

Performance inflation in different datasets

Databases of pathogenic variants curated from the literature are known to have a substantial frequency of false positives. There are likely similar factors causing false positives across different databases. Therefore, dividing the datasets into training and testing data does not create truly independent data for performance assessment, and as a result, the AURC calculated from VariBench data is likely inflated for methods trained on these dataset, including MVP and other methods with best AUROC values. This is supported by results in Supplementary Figure S5: using cancer somatic mutation hotspots as positives, and randomly selected rare variants from DiscovEHR as negatives, the area under receiver operating characteristic curve (AUROC) of all methods trained by HGMD or UniProt is substantially decreased (Supplementary Figure S5). Notably, MPC, which was trained on a small set of high-confidence ClinVar data, saw increased performance in cancer data, especially in non-constrained genes.

The results from *de novo* mutations provide further support. In Supplementary Figure S8, we estimated the precision of the optimal MVP score based on ROC curves with cancer and VariBench data, and used baseline precision (i.e. precision of "all missense") to bridge ROC and Precision-Recall calculation (see details below). The figure shows that the Precision-Recall point of optimal MVP score in *de novo* mutations is much closer to the estimated point based on cancer ROC curves than VariBench ROC curve in both constrained and non-constrained genes (supplementary Figure S8).

The procedure to estimate precision for a method at a certain threshold based on ROC curves

Denote the number of all true positives (pathogenic variants in cases) in a *de novo* mutation data set as $\boldsymbol{P}$, the estimated number of true positive detected by all methods at any threshold (including estimation from "all missense" without prediction methods) as a set $\boldsymbol{\mathcal{P}}$, the number of all negatives (non-pathogenic variants in cases) in the de novo mutation data as $\boldsymbol{N}$, the number of true positives by a method at a threshold as TP, the

number of false positives by a method at a threshold as *FP*, and the baseline precision as $\boldsymbol{B}$, defined as:

$$B \equiv \frac{P}{P+N}$$

$P+N$ is just the total number of *de novo* mutations in cases. We can estimate B by:

$$\hat{B} = \frac{\max(\boldsymbol{\mathcal{P}})}{P+N}$$

Therefore, N/P can be estimated as:

$$\frac{N}{P} = \frac{1}{1/\hat{B} - 1}$$

From the ROC curve, denote true positive rate (which is also called *recall* or *sensitivity*) as *TPR*, and false positive rate as *FPR*. We obtain *FPR* and *TPR* for a method at a certain threshold from cancer or VariBench ROC curves, and then use them to estimate number of true and false positives:

$$\widehat{TP} = P \cdot TPR$$
$$\widehat{FP} = N \cdot FPR$$

Therefore, the estimated precision of a method at a threshold based on ROC curve is:

$$\widehat{Precision} = \frac{\widehat{TP}}{\widehat{TP}+\widehat{FP}} = \frac{1}{1+\frac{\widehat{FP}}{\widehat{TP}}} = \frac{1}{1+\frac{FPR}{TPR}\cdot\frac{N}{P}} = \frac{1}{1+\frac{FPR}{TPR}*(\frac{1}{\hat{B}}-1)}$$