Deep structured learning realizes variant prioritization for Mendelian diseases

Matt C. Danzi, Maike F. Dohrn, Sarah Fazal, Danique Beijer, Adriana Rebelo, Vivian Cintra, and Stephan Züchner

Supplementary Figure Legends

Supplementary Figure 1: MAVERICK sub-model architecture 1. Inputs are shown as grey boxes. Transformer-based layers are shown in blue. Densely-connected linear layers are shown in orange. The number of parameters is shown to the left of each layer. The size of the output of each layer is shown on its right side. The size of each densely-connected linear layer is given in parentheses within the layer. For the multi-head attention layers, sixteen attention heads were used. The weights are shared between the two stacks of transformer layers.

Supplementary Figure 2: MAVERICK sub-model architecture 2. Inputs are shown as grey boxes. Transformer-based layers are shown in blue. Densely-connected linear layers are shown in orange. The number of parameters is shown to the left of each layer. The size of the output of each layer is shown on its right side. The size of each densely-connected linear layer is given in parentheses within the layer. For the multi-head attention layers, sixteen attention heads were used. ProtT5-XL BFD was used as an additional feature extractor in this architecture.

Supplementary Figure 3: MAVERICK exhibits excellent classification performance for benign, dominant, and recessive variants. A-C) Violin plot of distributions of MAVERICK predictions for each of the three classes, separated by true class label. A) Validation set. B) Known genes test set. C) Novel genes test set. D) Relative performance of MAVERICK with different input components ablated by dropout. Performance is measured by the area under the precision-recall curve, averaged among the benign, dominant and recessive scores.

Supplementary Figure 4: MAVERICK's predictions are well-calibrated. A-B) Scatter plots of binned prediction values on the x-axis plotted against the proportion of those predictions for which this was the correct class on the y-axis. A perfectly calibrated model would have all its points fall on the x=y line. Calibration curves above the x=y line indicate under-confidence, while those under the x=y line indicate over-confidence. A) Calibration curve for the known genes test set. B) Calibration curve for the novel genes test set.

Supplementary Figure 5: MAVERICK reliably prioritizes causal variants. Cumulative proportion of cases solved by MAVERICK's rank ordering of variants when 98 control samples had pathogenic variants from the known and novel genes test sets spiked in. A) Performance on

the known genes test set separated according to the inheritance pattern of the spiked-in variant. B) Performance on the novel genes test set separated according to the inheritance pattern of the spiked-in variant.

Supplementary Tables

Supplementary Table 1: Classification performance of MAVERICK on the validation set, the known genes test set and the novel genes test set. For each class in each test set, the number of variants, precision, recall, F1-score, area under the receiver operating characteristic curve, and area under the precision-recall curve are given.

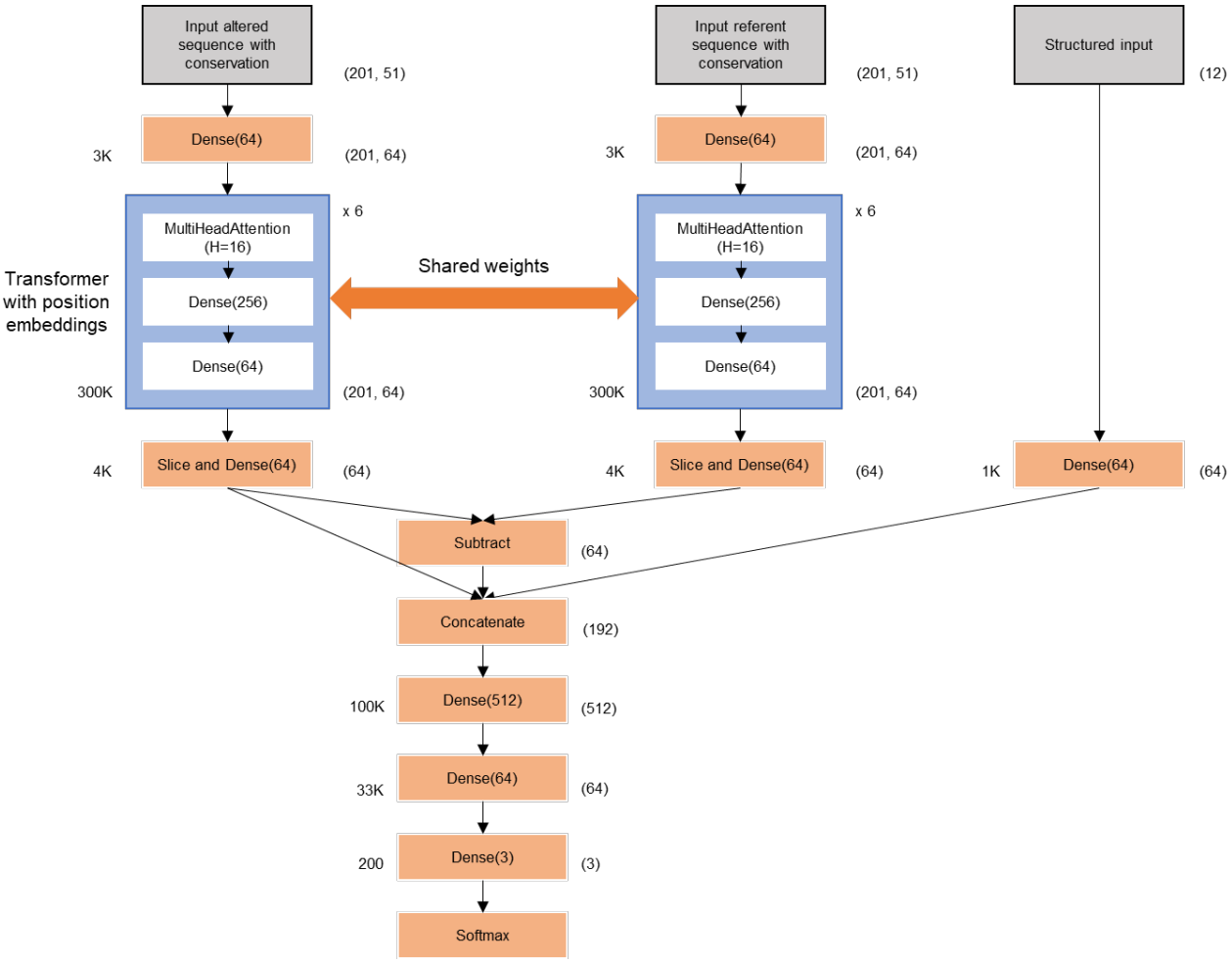
Supplementary Table 2: Comparison of MAVERICK classification performance to MAPPIN. For each class in the known and novel genes test sets, the number of variants evaluated is given, along with the precision and recall of MAVERICK and MAPPIN.

Supplementary Table 3: Comparison of MAVERICK classification performance to ALoFT. For each class in the known and novel genes test sets, the number of variants evaluated is given, along with the precision and recall of MAVERICK and ALoFT.

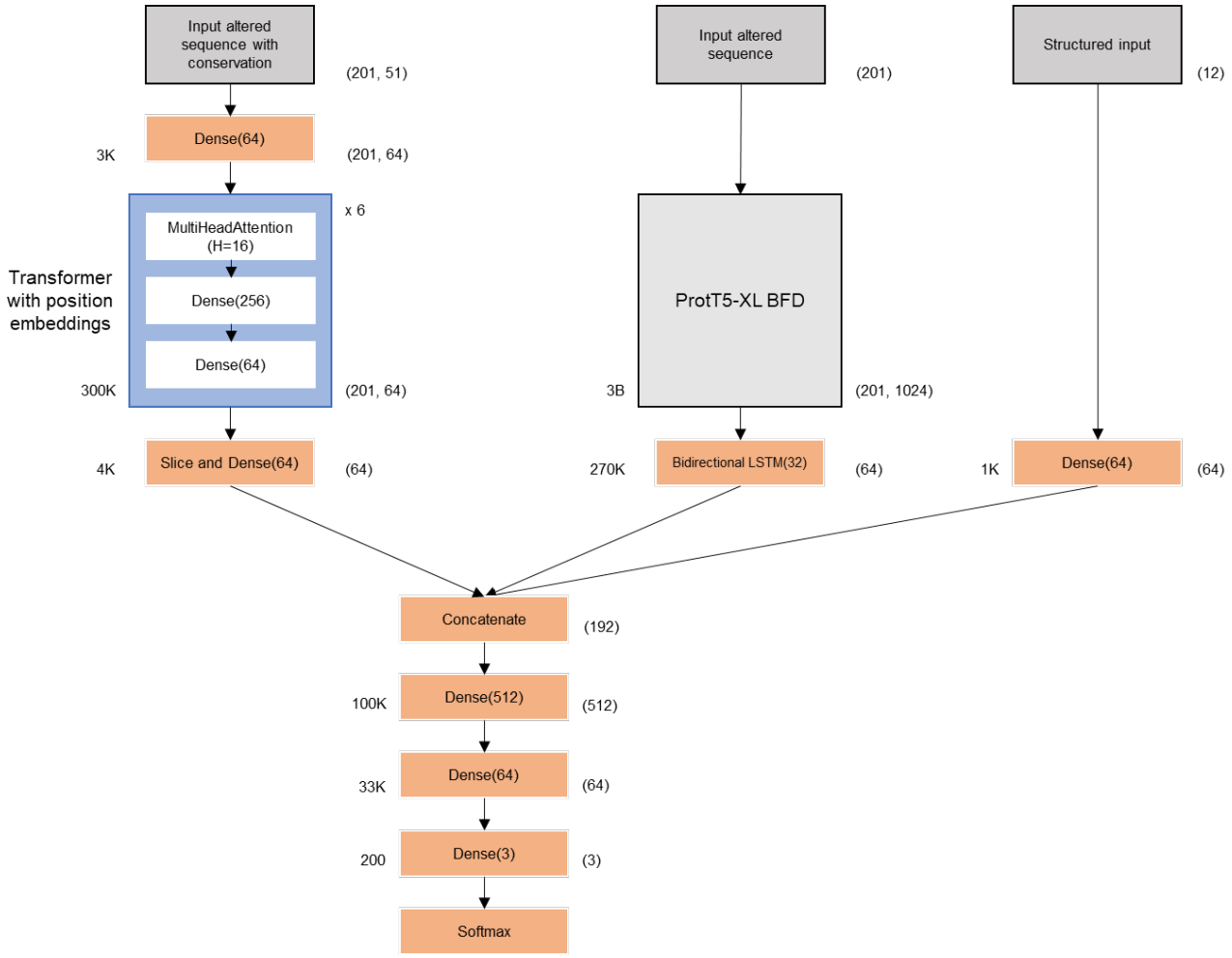
Supplementary Table 4: Comparison of MAVERICK classification performance to PrimateAI. For each class in the known and novel genes test sets, the number of variants evaluated is given, along with the precision and recall of MAVERICK and PrimateAI. Dominant and recessive variants were both included to create the 'pathogenic' class for this comparison.

Supplementary Table 5: Comparison of MAVERICK to Exomiser in solving real novel cases with and without phenotype information. For each case, the causal gene is shown, along with its mode of inheritance, the patient's HPO terms, and the rank of the causal variant or variant pair when evaluated by MAVERICK using only genotype information, MAVERICK score averaged with HiPhive phenotype score, Exomiser using only genotype information, and Exomiser combined gene-level genotype and phenotype score.

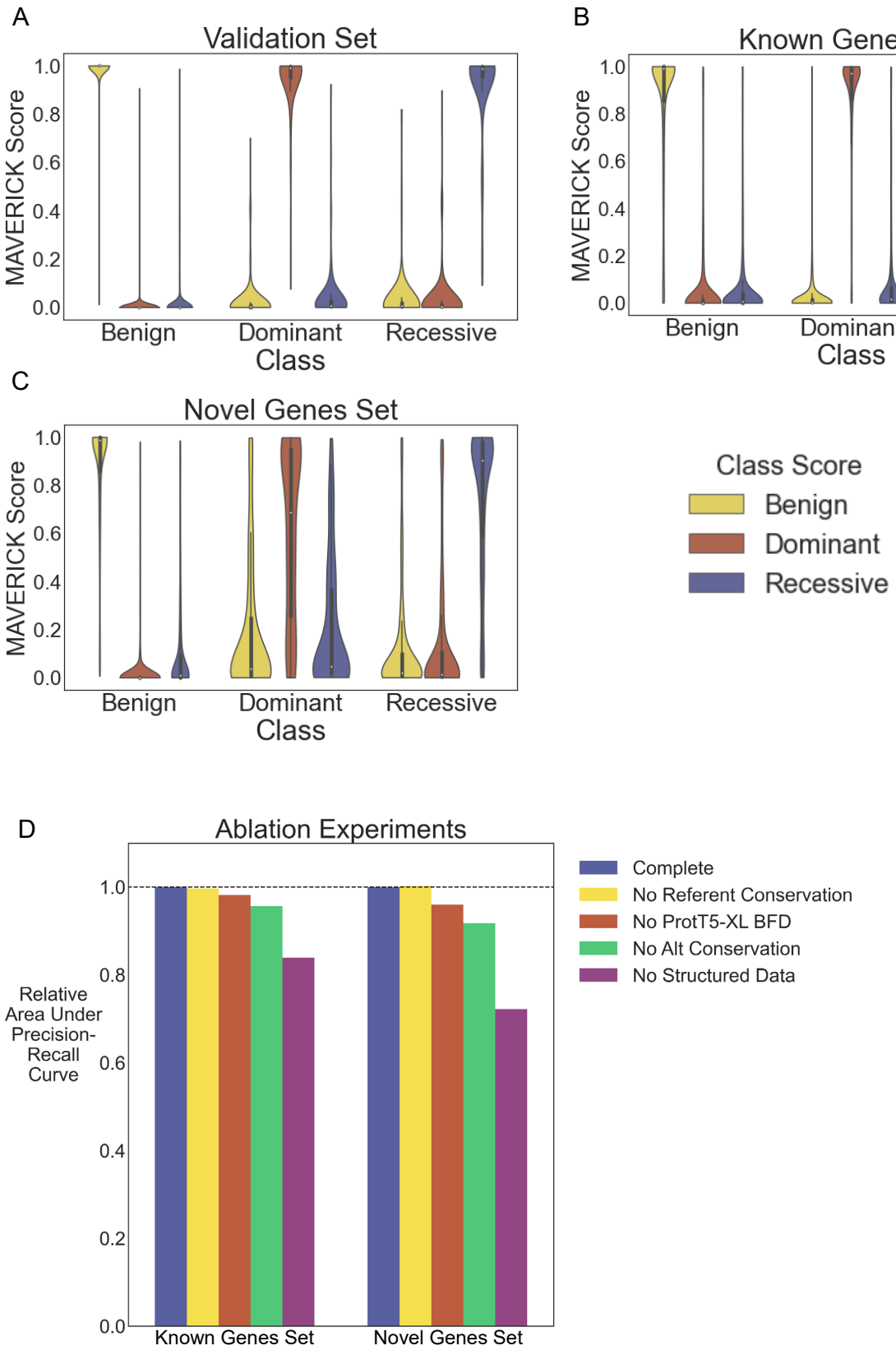
Supplementary Table 6: Classification performance of MAVERICK on the X chromosome. For each class, the number of variants, precision, recall, F1-score, area under the receiver operating characteristic curve, and area under the precision-recall curve are given. These scores are again reported for the "binary" case where dominant and recessive variants are grouped together as "pathogenic".



Supplementary Figure 1

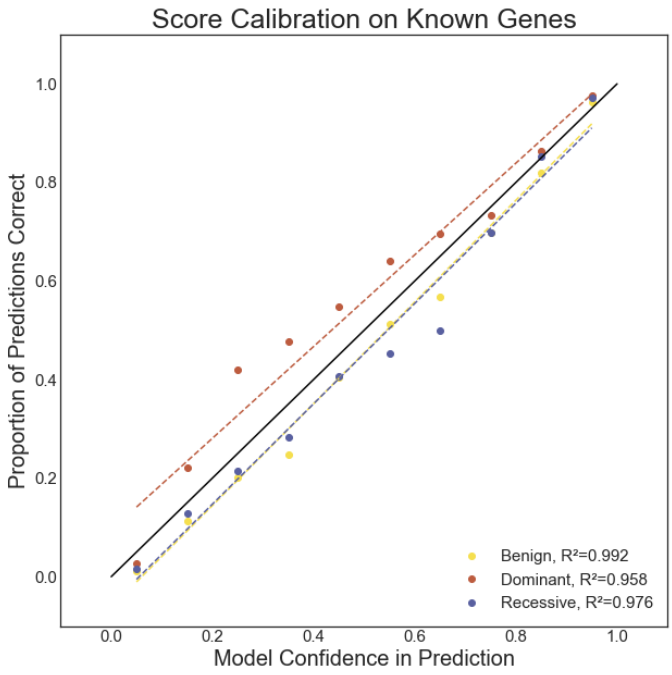


Supplementary Figure 2

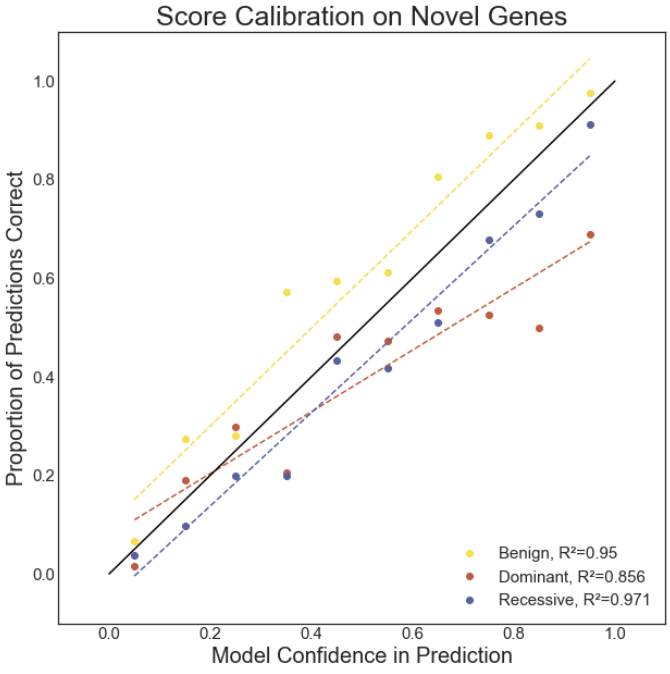


Supplementary Figure 3

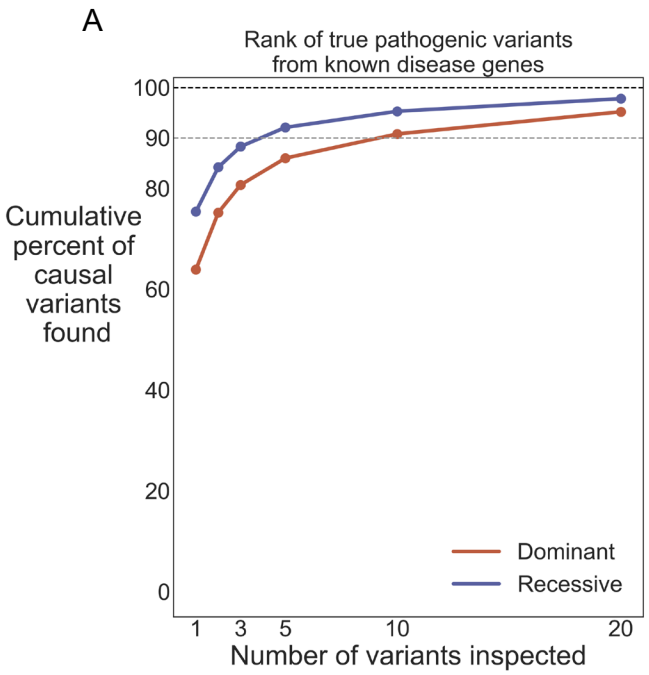
A

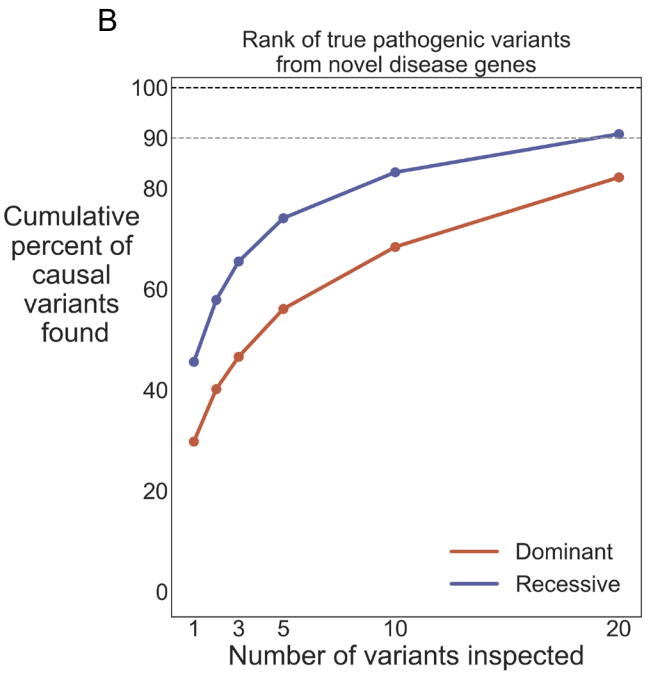


B



Supplementary Figure 4





Supplementary Figure 5

Supplementary Table 1

Validation Set

Class Label	Precision	Recall	F1	N	auROC	auPRC
Benign	0.9923	0.9884	0.9903	778	0.9987	0.9996
Dominant	0.9541	0.963	0.9585	108	0.9991	0.9929
Recessive	0.9310	0.9474	0.9391	114	0.9975	0.9789

Known Genes Set

Class Label	Precision	Recall	F1	N	auROC	auPRC
Benign	0.8824	0.8797	0.881	2917	0.9799	0.9411
Dominant	0.9239	0.8902	0.9068	6085	0.9752	0.9632
Recessive	0.9130	0.9431	0.9278	7010	0.9805	0.9751

Novel Genes Set

Class Label	Precision	Recall	F1	N	auROC	auPRC
Benign	0.9343	0.9109	0.9224	1234	0.9623	0.9771
Dominant	0.5930	0.6448	0.6178	183	0.9359	0.6253
Recessive	0.7765	0.7992	0.7877	513	0.9337	0.8598

Supplementary Table 2

Known Genes Set	MAVERICK		MAPPIN		
Variant Class	Precision (%)	Recall (%)	Precision (%)	Recall (%)	N
Benign	91.1	89.6	94.1	7.5	2541
Dominant	87.8	86.3	60.5	81.5	2576
Recessive	89.3	91.9	60.1	89.6	2945

Novel Genes Set	MAVERICK		MAPPIN		
Variant Class	Precision (%)	Recall (%)	Precision (%)	Recall (%)	N
Benign	94.8	91.0	97.7	8.0	1072
Dominant	62.6	62.1	21.9	59.9	132
Recessive	69.1	79.7	24.1	87.8	286

Supplementary Table 3

Known Genes Set	MAVERICK		ALoFT		
Variant Class	Precision (%)	Recall (%)	Precision (%)	Recall (%)	N
Benign	50.0	9.1	0.0	0.0	22
Dominant	92.4	86.0	69.4	71.1	1001
Recessive	90.7	95.8	80.1	79.8	1538

Novel Genes Set	MAVERICK		ALoFT		
Variant Class	Precision (%)	Recall (%)	Precision (%)	Recall (%)	N
Benign	100.0	22.2	100	11.1	9
Dominant	54.6	64.3	40.0	42.9	28
Recessive	89.3	90.5	75.3	88.5	157

Supplementary Table 4

Known Genes Set	MAVERICK		PrimateAI		
Variant Class	Precision (%)	Recall (%)	Precision (%)	Recall (%)	N
Benign	91.5	89.7	83.0	52.2	2473
Pathogenic	91.8	93.3	70.3	91.4	3062

Novel Genes Set	MAVERICK		PrimateAI		
Variant Class	Precision (%)	Recall (%)	Precision (%)	Recall (%)	N
Benign	95.2	91.6	94.7	56.3	1041
Pathogenic	68.00	79.6	30.8	86.00	235

Supplementary Table 6

X-chromosome

Class Label	Precision	Recall	F1	N	auROC	auPRC
Benign	0.9718	0.9727	0.9723	3333	0.9924	0.9955
Dominant	0.3553	0.8801	0.5063	667	0.8849	0.4623
Recessive	0.7383	0.1519	0.2520	1244	0.8749	0.5914

X-chromosome binary

Class Label	Precision	Recall	F1	N	auROC	auPRC
Benign	0.9738	0.9697	0.9717	3333	0.9924	0.9955
Pathogenic	0.9475	0.9545	0.951	1911	0.9924	0.9879