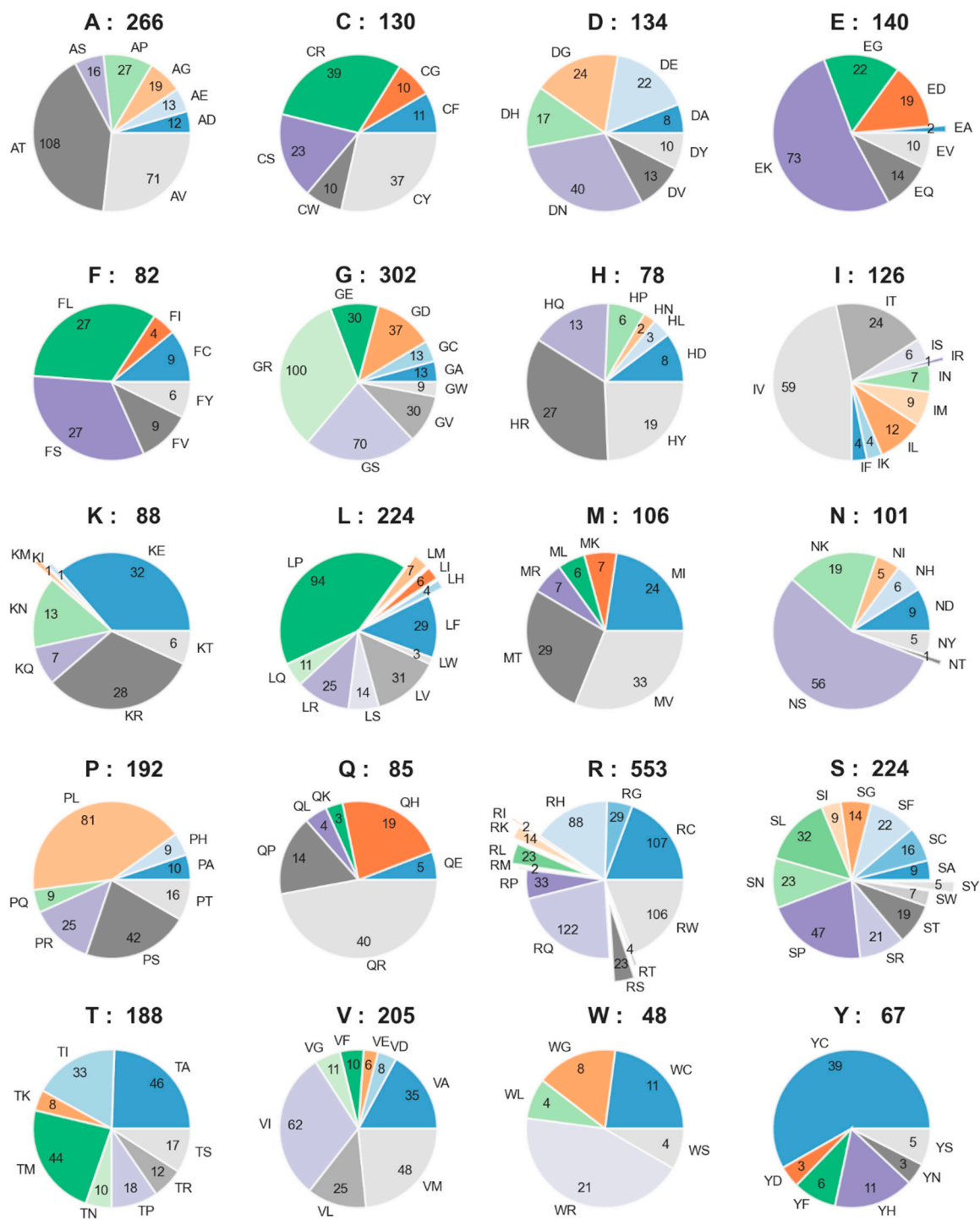


Most monogenic disorders are caused by mutations altering protein folding free energy

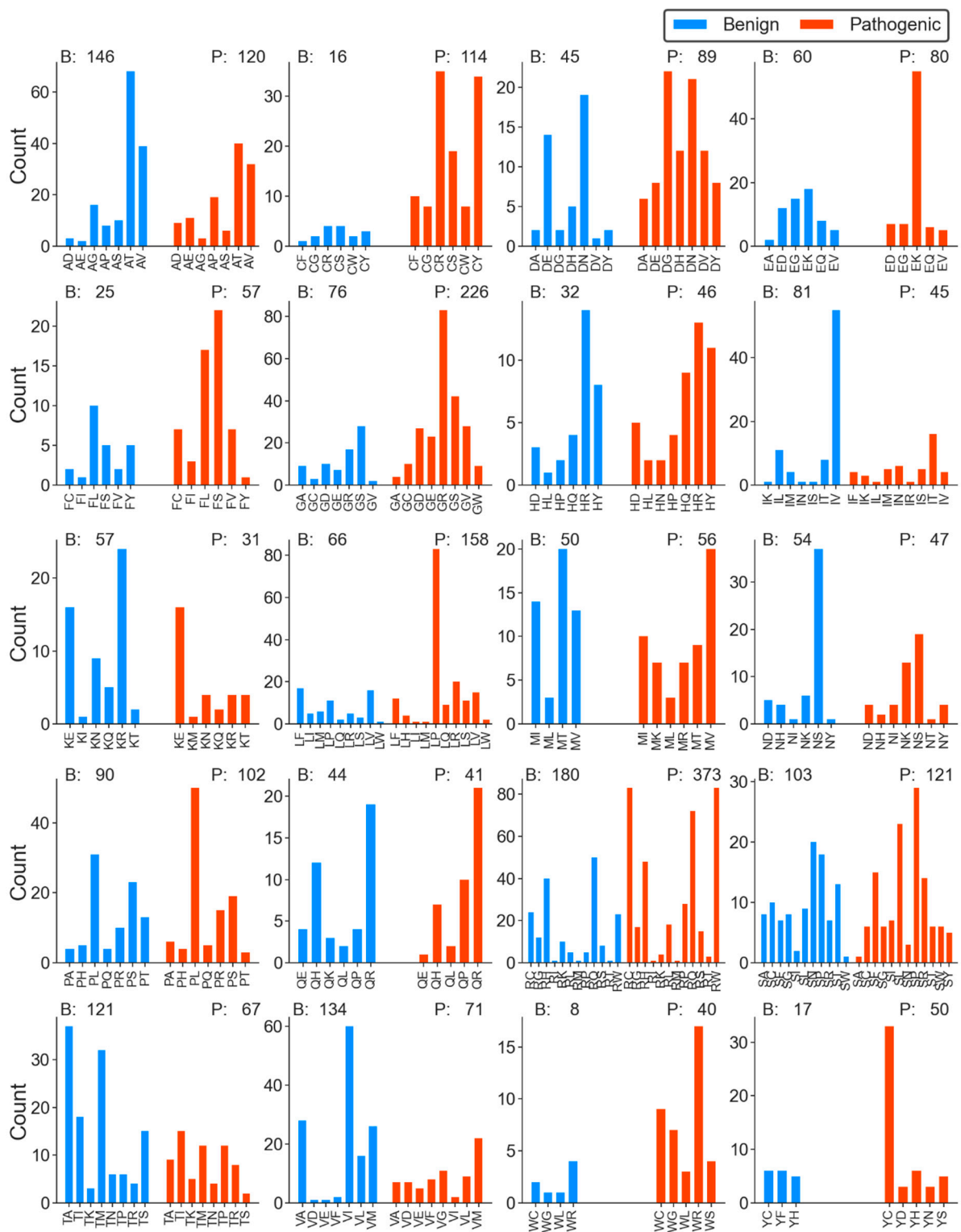
Preeti Pandey¹ and Emil Alexov^{1,*}

Department of Physics and Astronomy, Clemson University, Clemson, SC 29634, USA

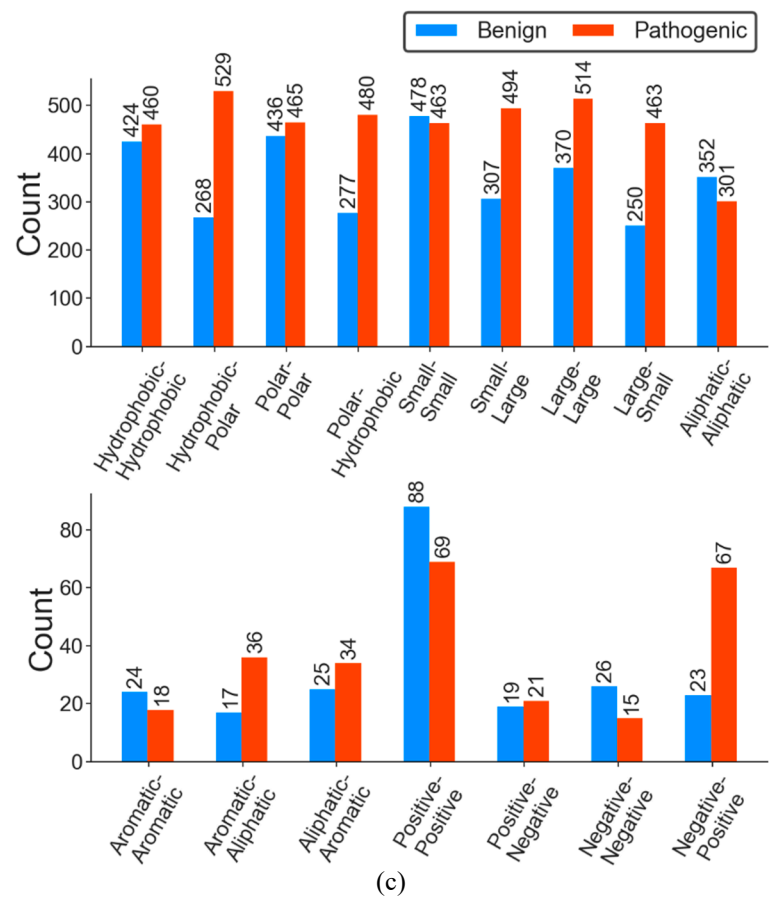
* Correspondence: Corresponding author email: calexov@clemson.edu



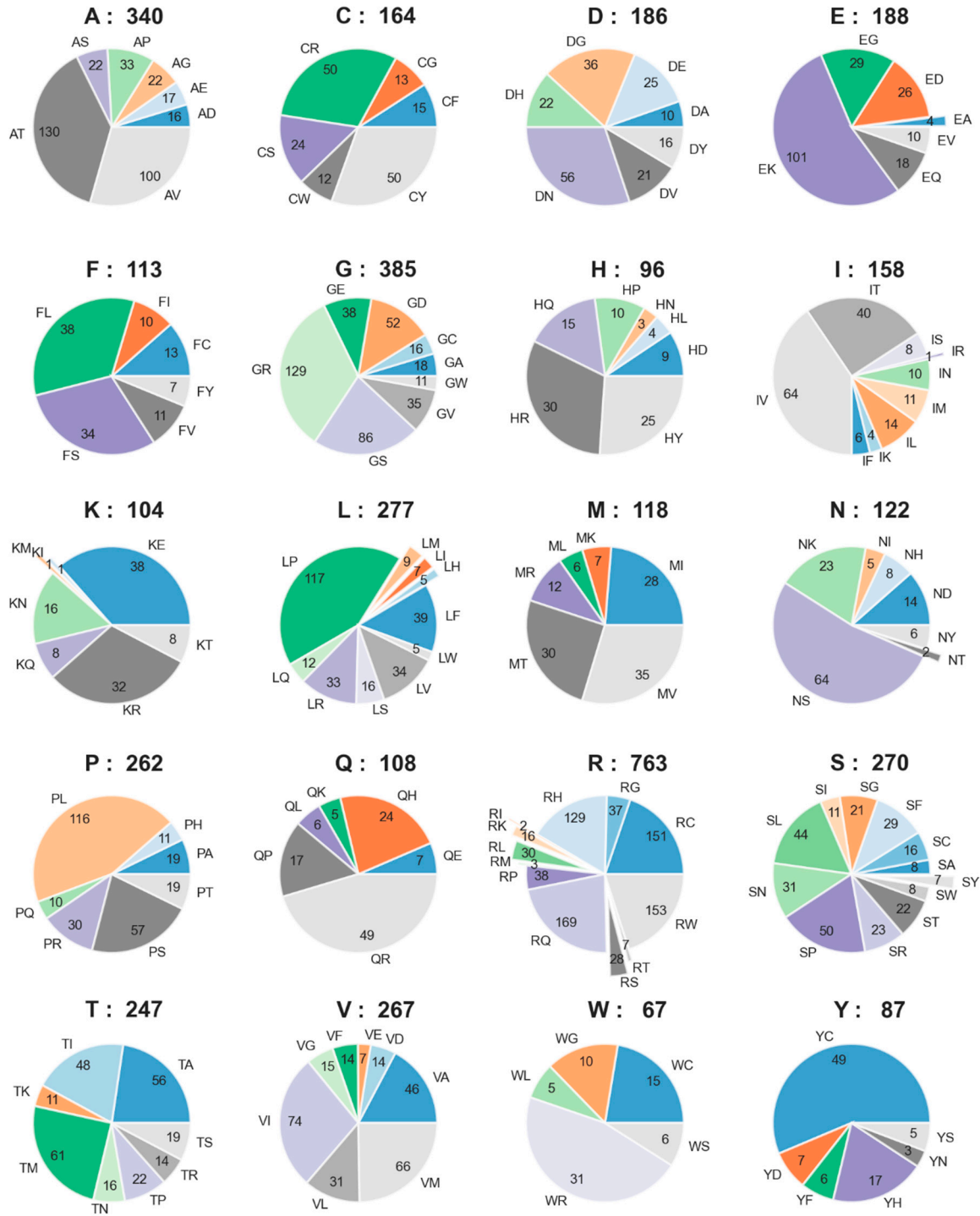
(a)



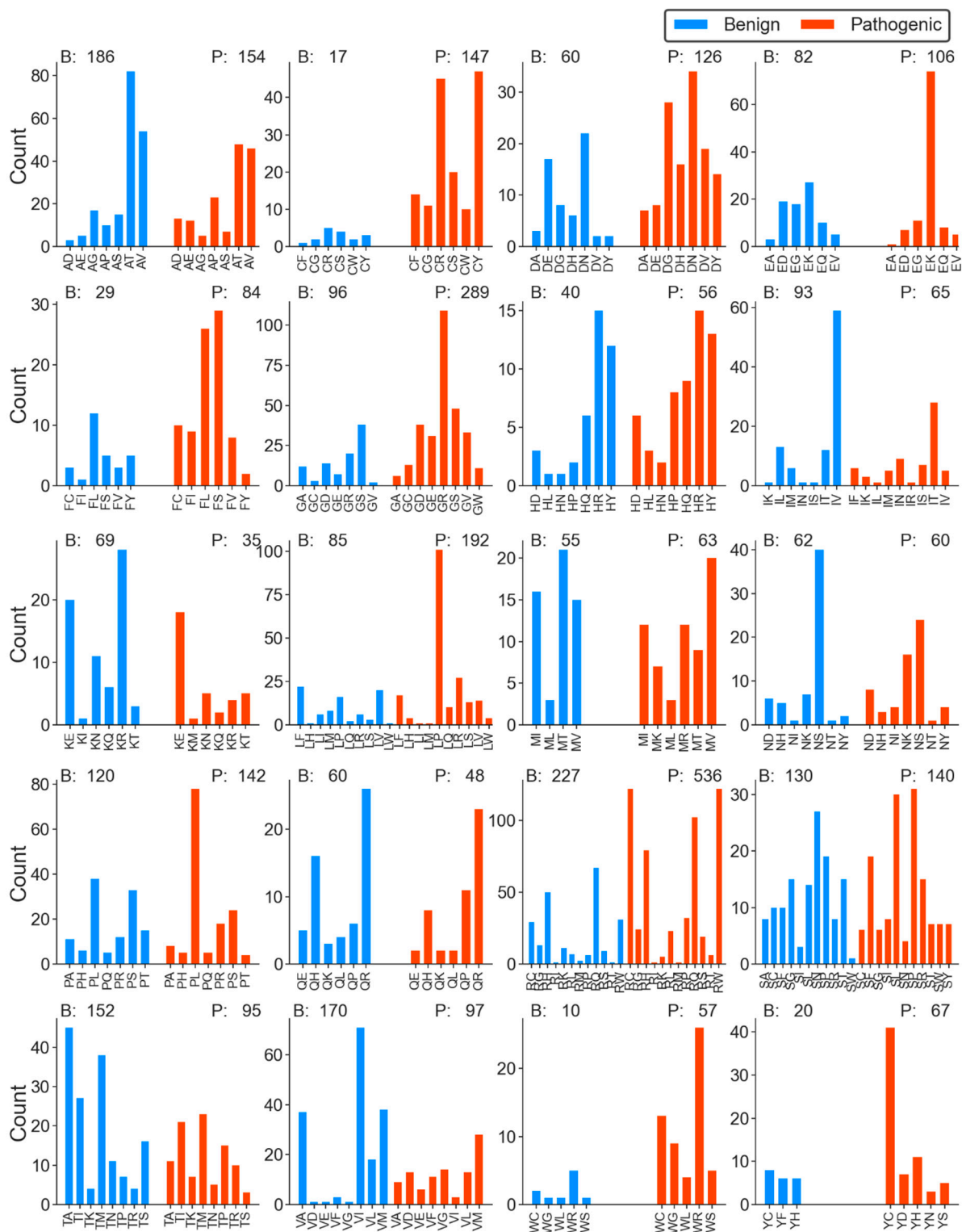
(b)



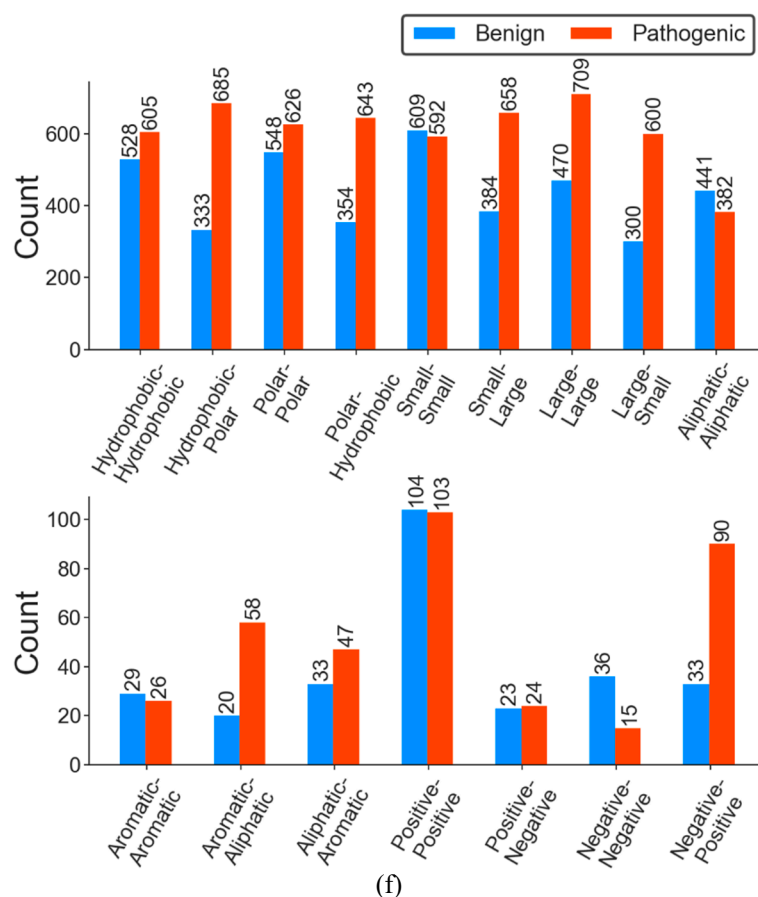
(c)



(d)



(e)



(f)

Figure S1. (a). Pie-chart summarising count of different amino acid mutations in the Monogenic Disorder Dataset 1 (no likely cases). In each subplot, the title shows the count of wild-type amino acids in the whole dataset, and the two-letter label of every pie-chart denotes the one-letter codes of the wild-type and mutant amino acid pair, respectively. (b). Bar plots summarising count of benign and pathogenic mutations in the Monogenic Disorder Dataset 1 (no likely cases). The two-letter label of every bar denotes the one-letter codes of the wild-type and mutant amino acid pairs, respectively. B denotes benign and P denotes pathogenic mutations. (c). Benign vs. pathogenic mutations assigned according to the property of amino acid mutation in the Monogenic Disorder Dataset 1 (no likely cases). The two-word label in the bar plot represents the chemical nature of the wild-type and mutant residues pairs, respectively. The amino acids according to their chemical properties are categorised as follows: Ala, Cys, Gly, Ile, Leu, Met, Phe, Pro, Trp, and Val are categorised as hydrophobic. Asp, Glu, Lys, Arg, His, Asn, Gln, Ser, Thr, and Tyr are categorised as polar. His, Phe, Trp, and Tyr as aromatic. Ala, Ile, Lys, Leu, Met, Pro, and Val as aliphatic. His, Lys, Arg as positive. Asp, and Glu as negative. Ala, Cys, Gly, Ser, Asn, Asp, Pro, Thr, and Val as small. And Arg, Gln, Glu, His, Ile, Leu, Lys, Met, Phe, Trp, and Tyr as large amino acids. (d). Pie-chart summarising count of different amino acid mutations in the Monogenic Disorder Dataset 2 (including likely cases). In each subplot, the title shows the count of wild-type amino acids in the whole dataset, and the two-letter label of every pie-chart denotes the one-letter codes of the wild-type and mutant amino acid pair, respectively. (e). Bar plots summarising count of benign and pathogenic mutations in the Monogenic Disorder Dataset 2 (including likely cases). The two-letter label of every bar denotes the one-letter codes of the wild-type and mutant amino acid pairs, respectively. B denotes benign and P denotes pathogenic mutations. (f). Benign vs. pathogenic mutations assigned according to the property of amino acid mutation in the Monogenic Disorder Dataset 2 (including likely cases). The two-word label in the bar plot represents the chemical nature of the wild-type and mutant residues pairs, respectively. The amino acids according to their chemical properties are categorised as follows: Ala, Cys, Gly, Ile, Leu, Met, Phe, Pro, Trp, and Val are categorised as hydrophobic. Asp, Glu, Lys, Arg, His, Asn, Gln, Ser, Thr, and Tyr are categorised as polar. His, Phe, Trp, and Tyr as aromatic. Ala, Ile, Lys, Leu, Met, Pro, and Val as aliphatic. His, Lys, Arg as positive. Asp, and Glu as negative. Ala, Cys, Gly, Ser,

Asn, Asp, Pro, Thr, and Val as small. And Arg, Gln, Glu, His, Ile, Leu, Lys, Met, Phe, Trp, and Tyr as large amino acids.

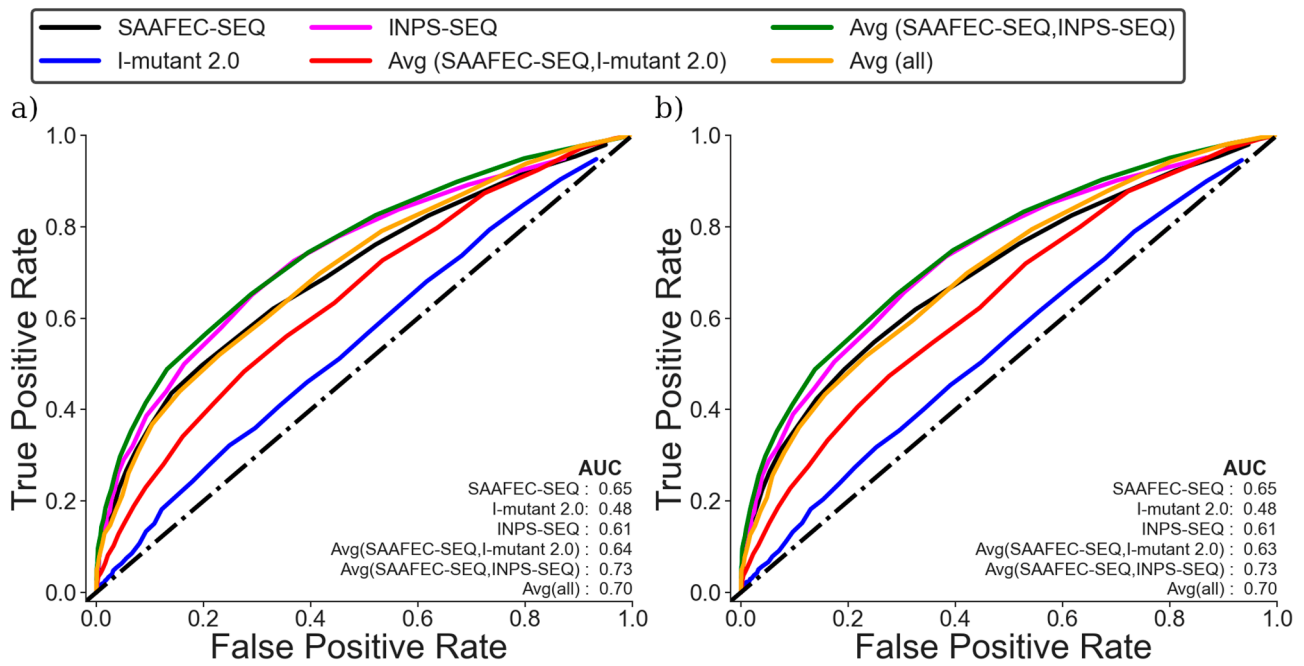


Figure S2. ROC curve for monogenic disorder dataset using absolute $\Delta\Delta G$ values. (a) Monogenic Disorder Dataset 1 (no likely cases) (b) Monogenic Disorder Dataset 2 (including likely cases).

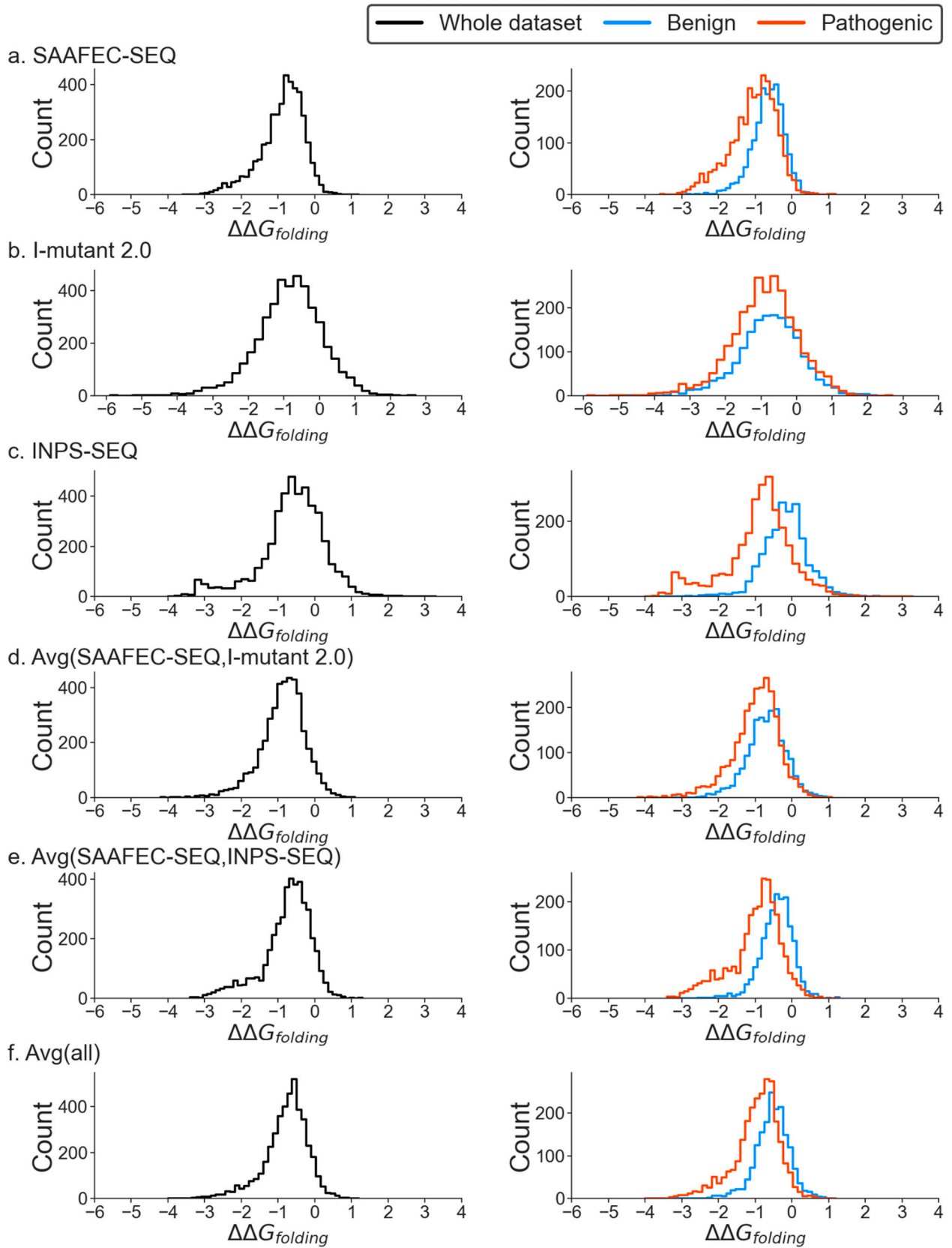
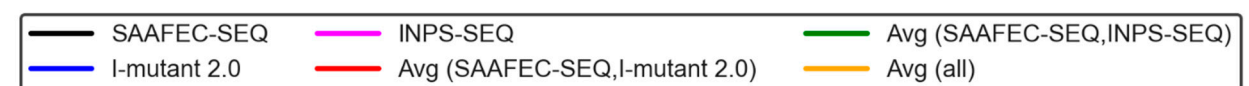
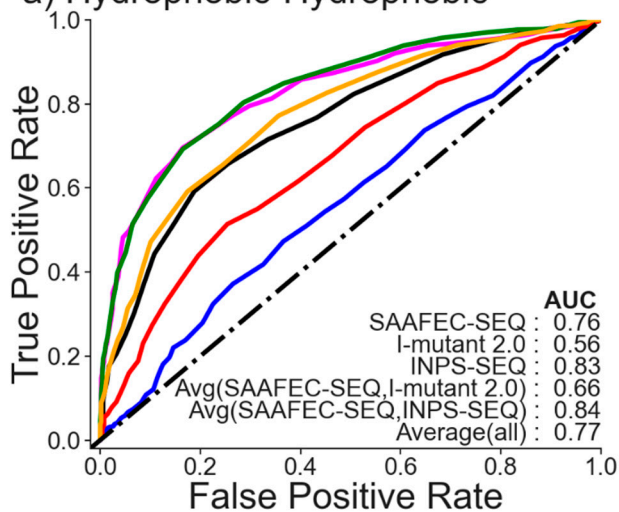


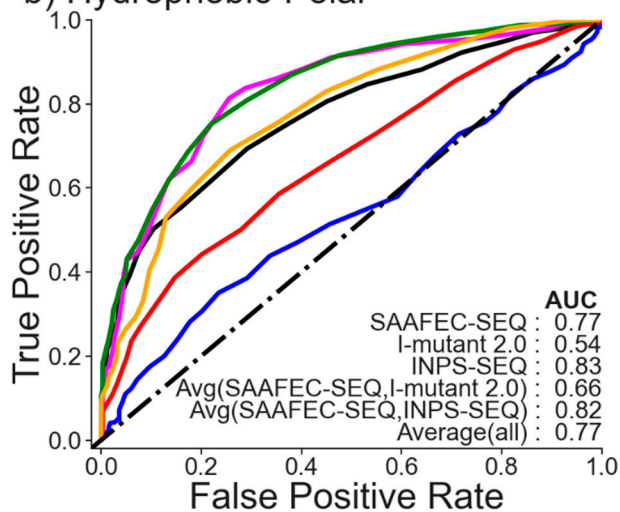
Figure S3. Distribution of change in folding free energy using different predictors using Monogenic Disorder Dataset 2 (including likely cases).



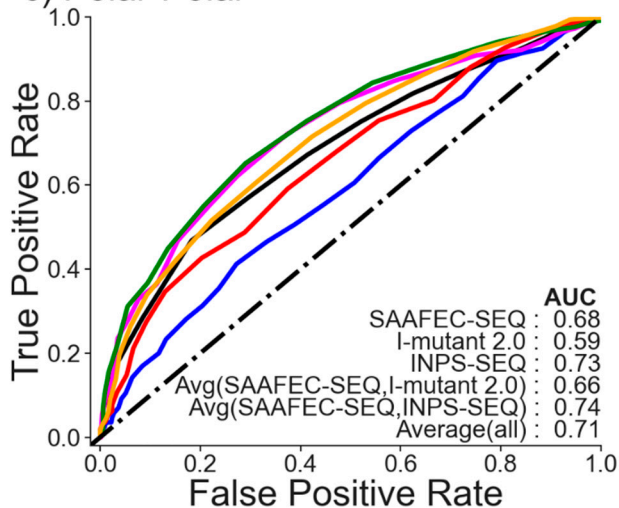
a) Hydrophobic-Hydrophobic



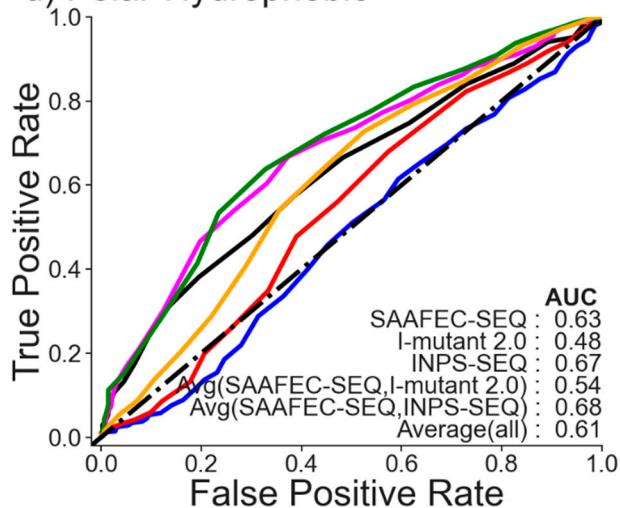
b) Hydrophobic-Polar



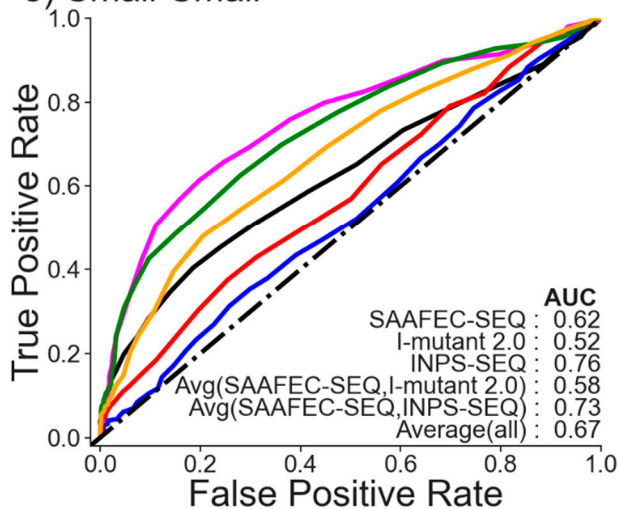
c) Polar-Polar



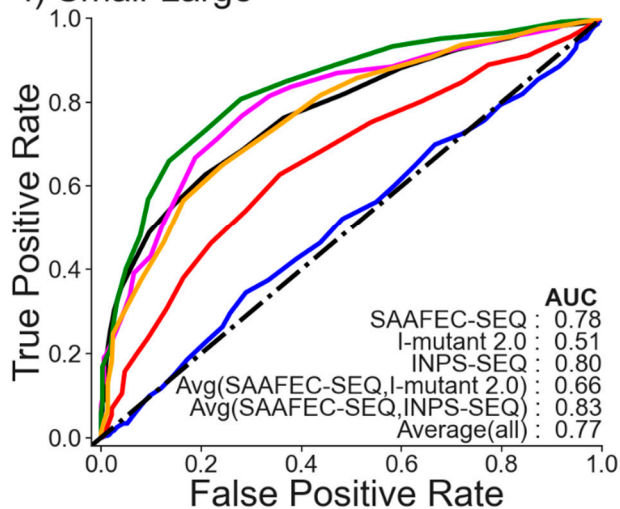
d) Polar-Hydrophobic



e) Small-Small



f) Small-Large



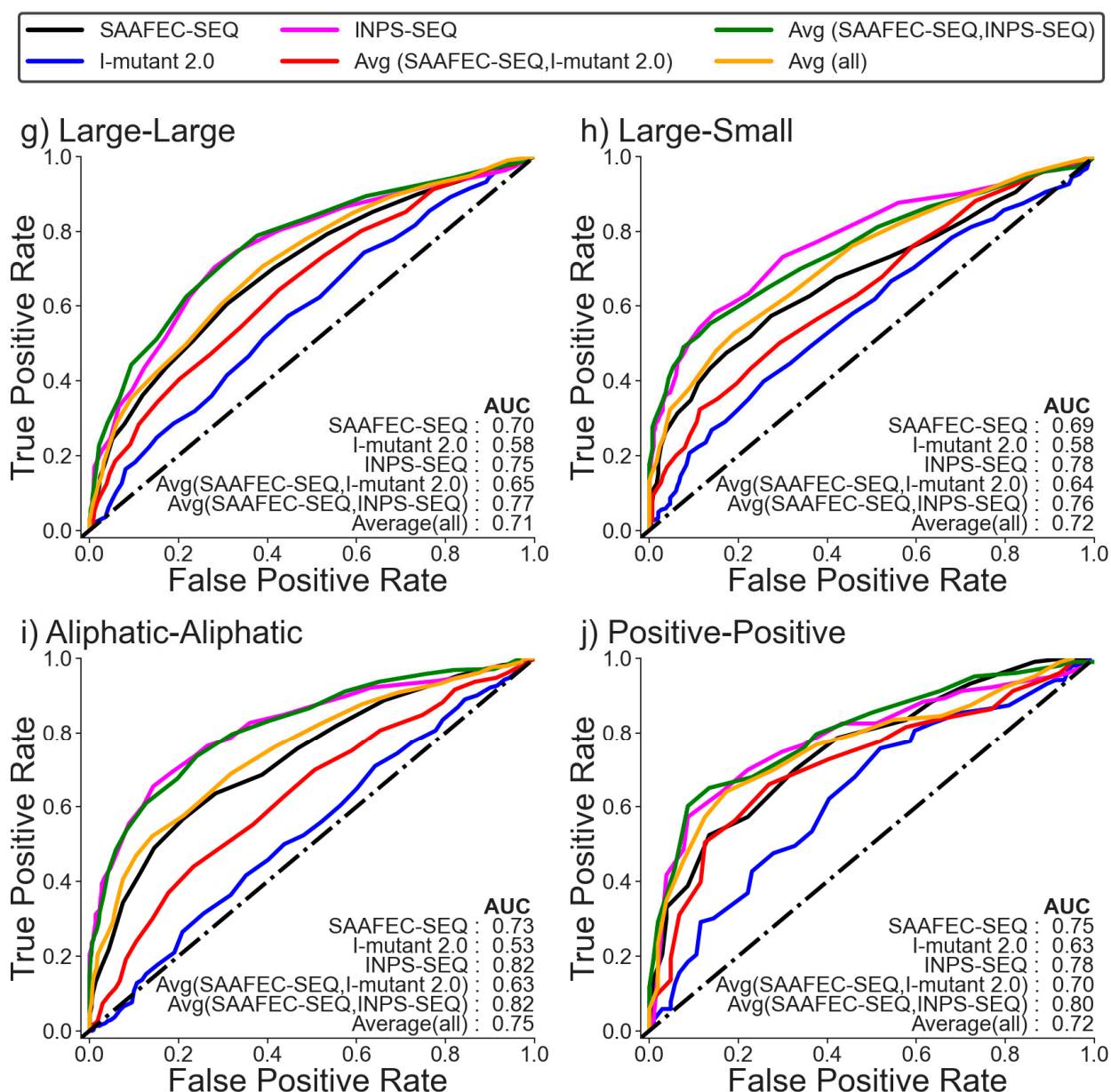
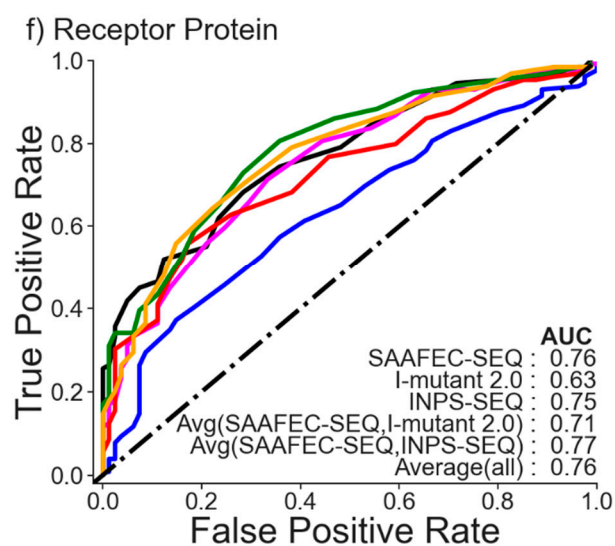
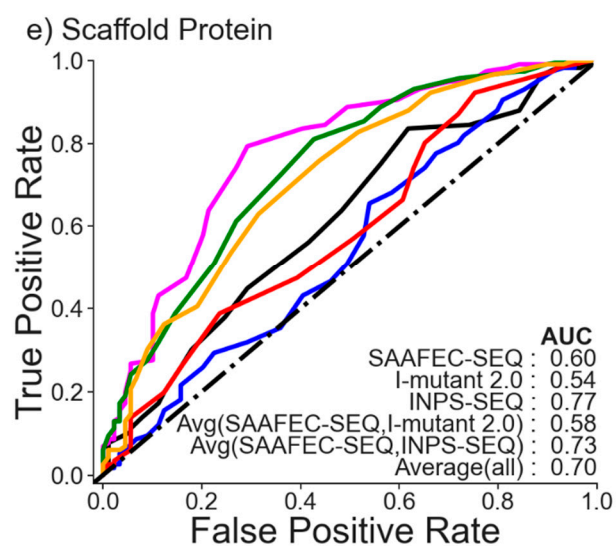
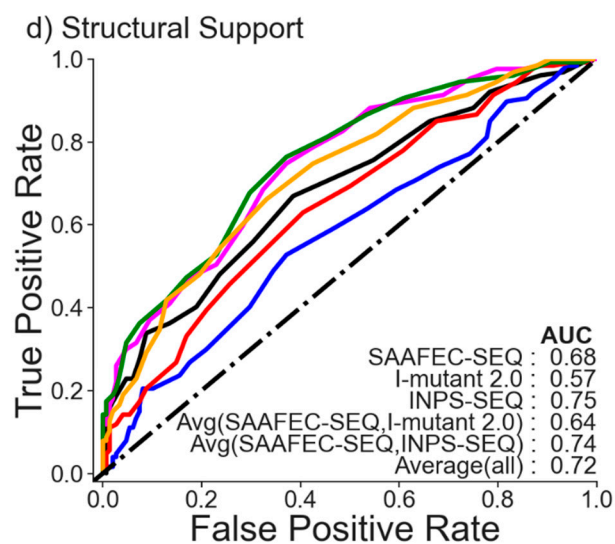
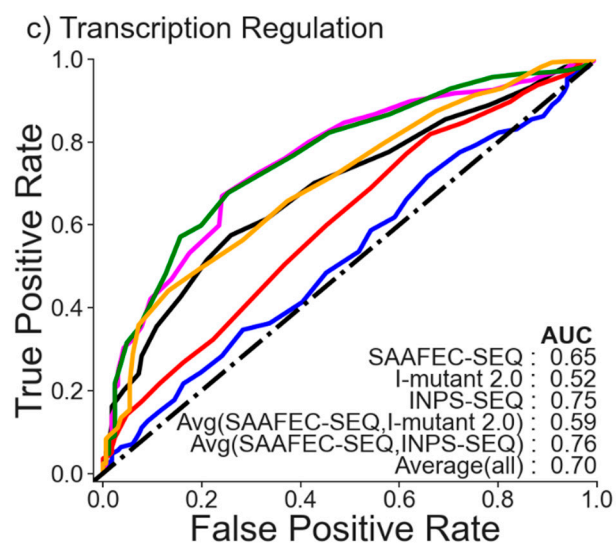
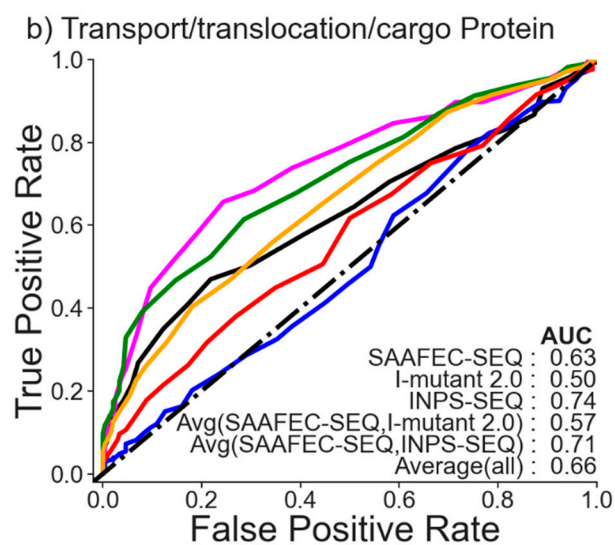
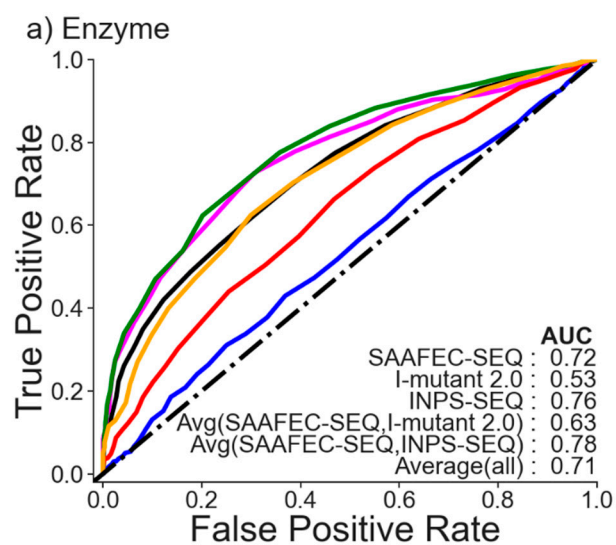
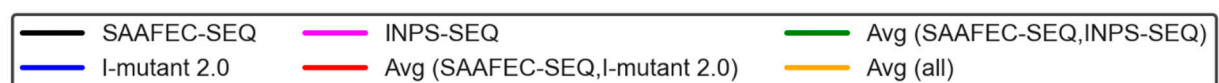


Figure S4. ROC curve for Monogenic Disorder Dataset 2 (including likely cases) for different categories of amino acid mutations. The amino acids according to their chemical properties are categorised as follows: Ala, Cys, Gly, Ile, Leu, Met, Phe, Pro, Trp, and Val are categorised as hydrophobic. Asp, Glu, Lys, Arg, His, Asn, Gln, Ser, Thr, and Tyr are categorised as polar. His, Phe, Trp, and Tyr as aromatic. Ala, Ile, Lys, Leu, Met, Pro, and Val as aliphatic. His, Lys, Arg as positive. Asp, and Glu as negative. Ala, Cys, Gly, Ser, Asn, Asp, Pro, Thr, and Val as small. And Arg, Gln, Glu, His, Ile, Leu, Lys, Met, Phe, Trp, and Tyr as large amino acids.



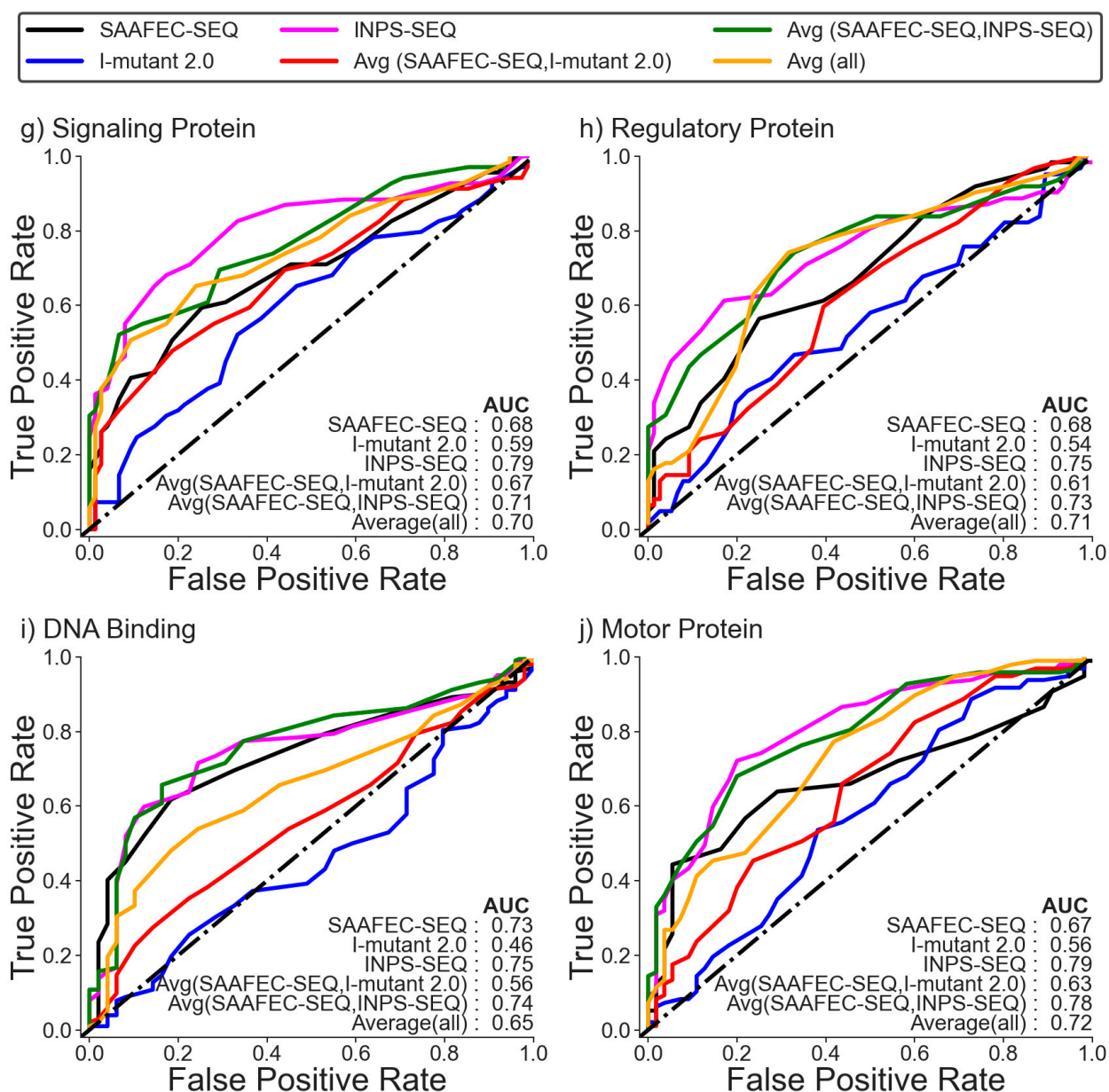


Figure S5. ROC curve for Monogenic Disorder Dataset 2 (including likely cases) for different functional categories.

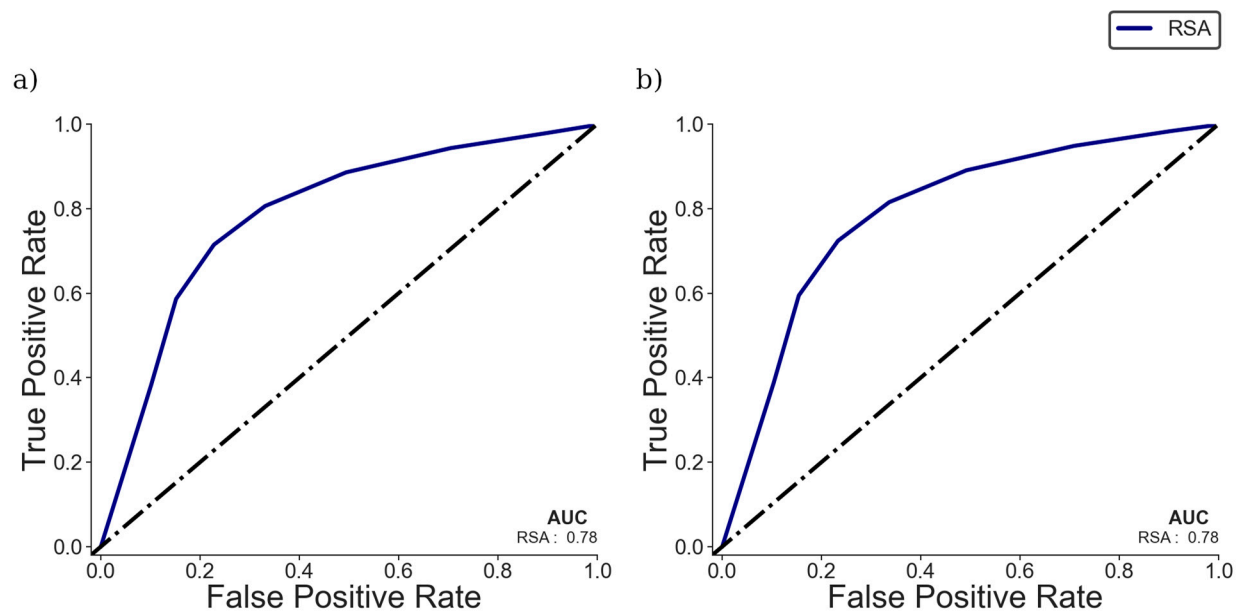
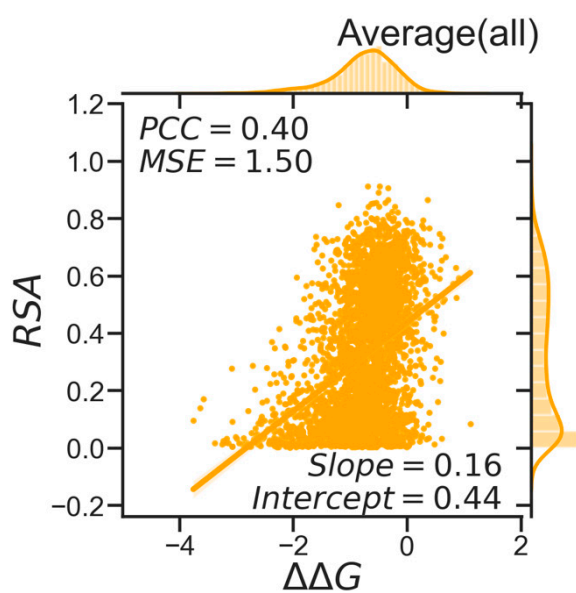
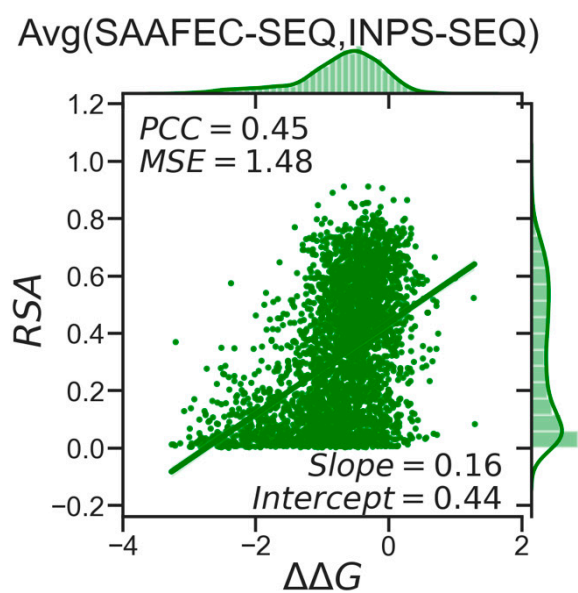
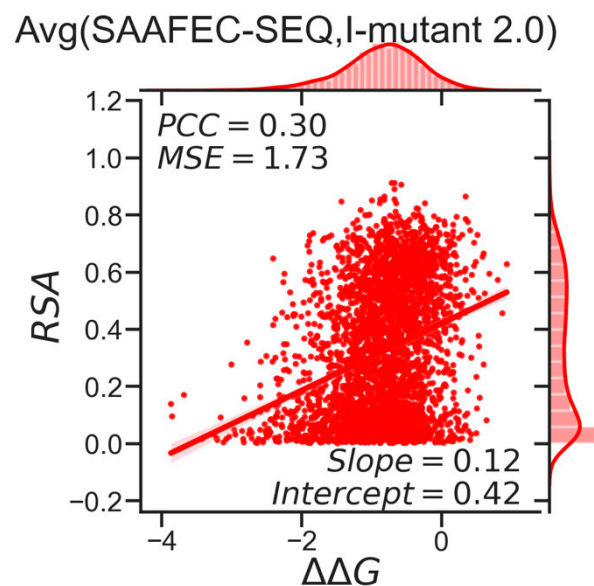
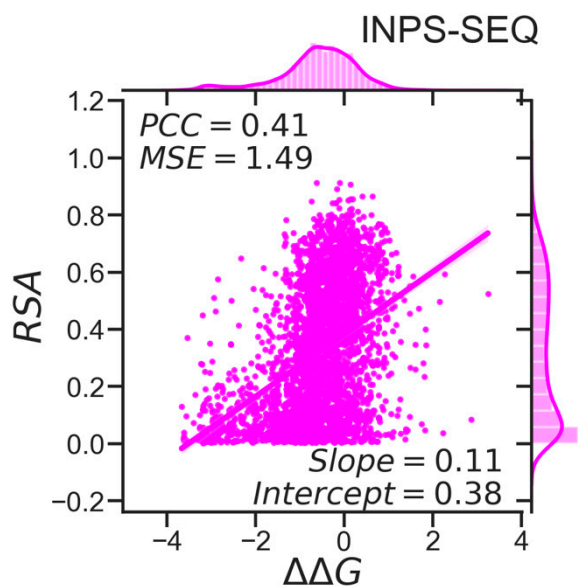
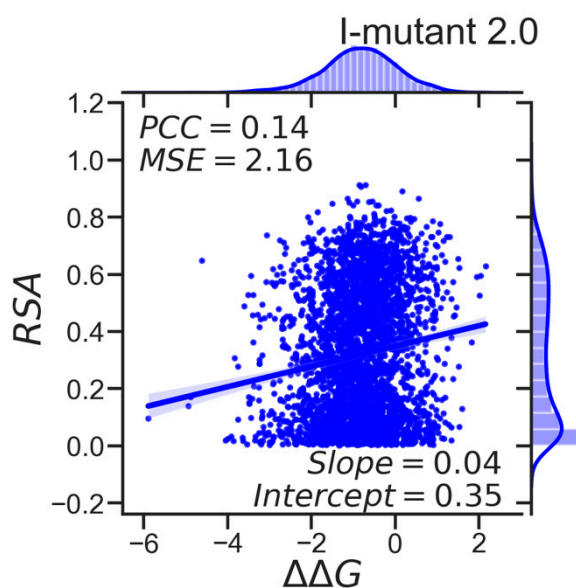
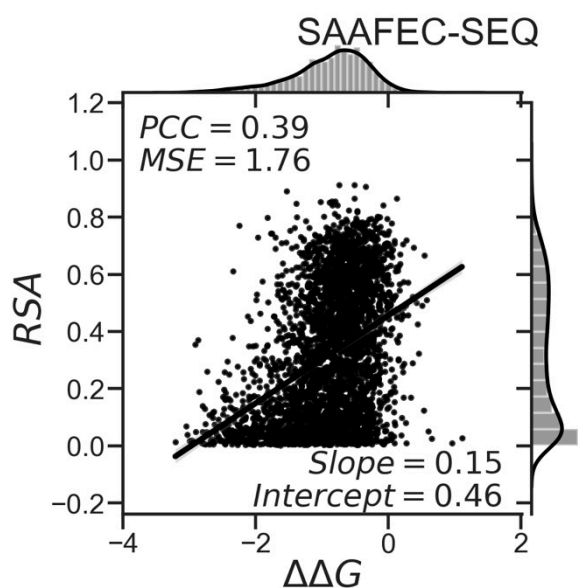
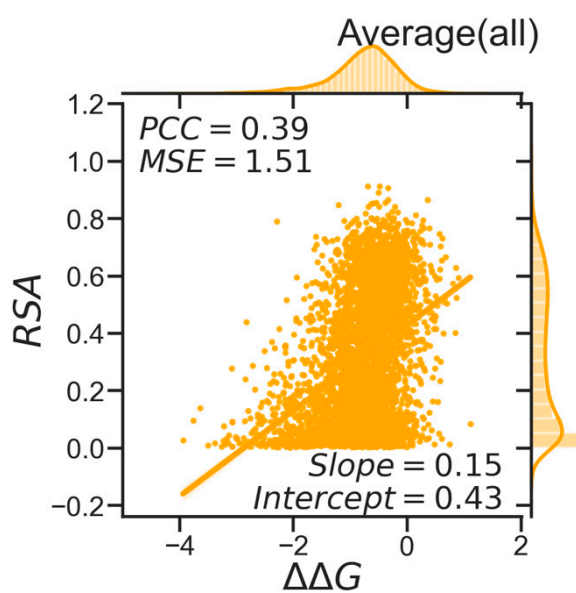
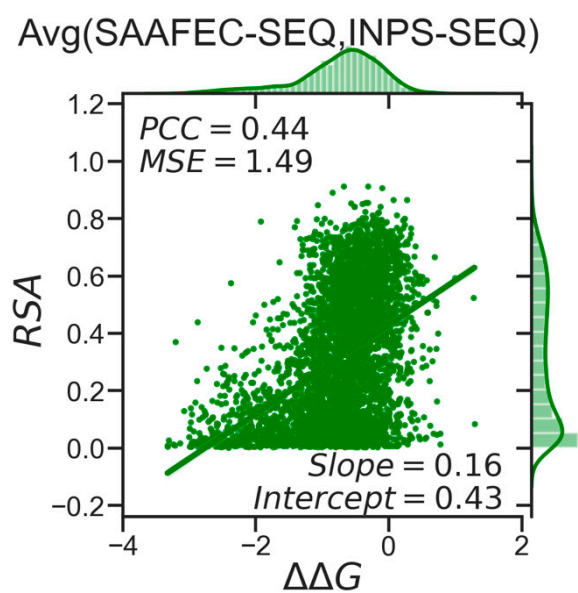
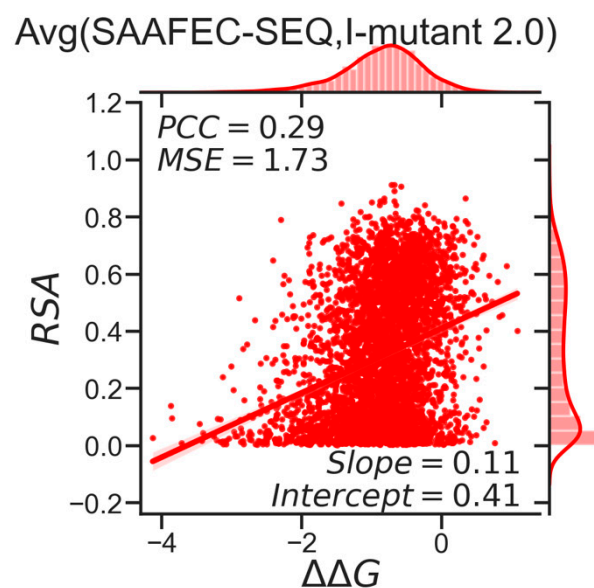
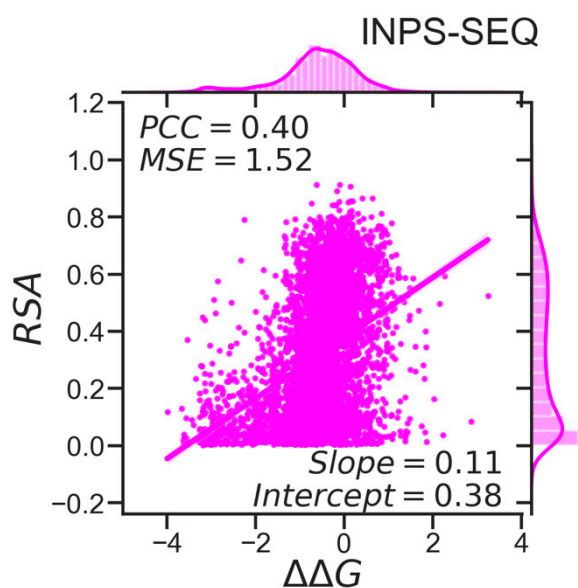
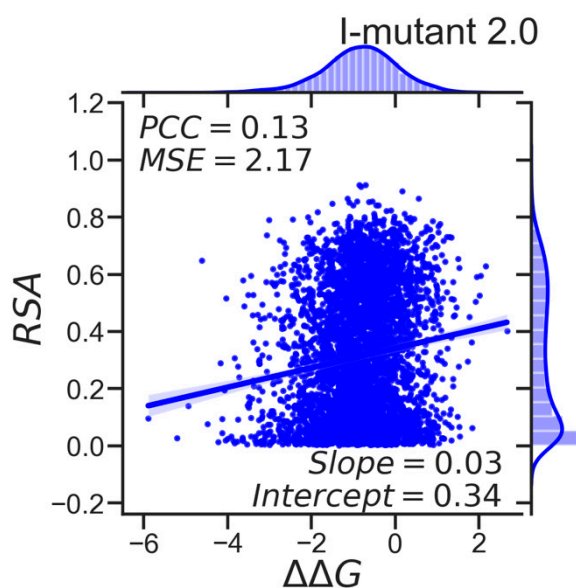
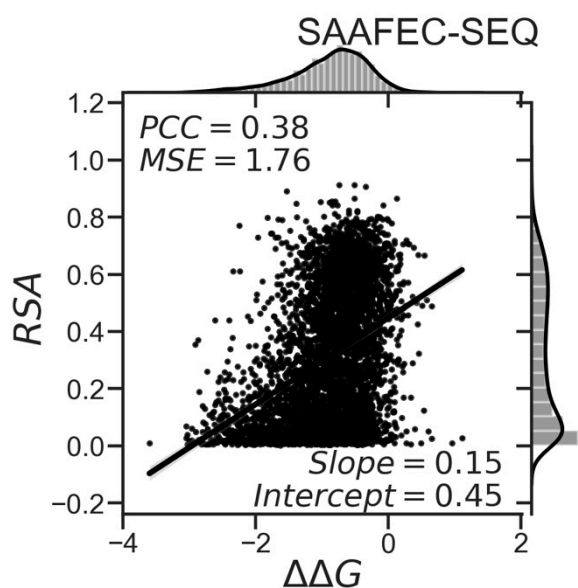


Figure S6. ROC curve for monogenic disorder dataset using Relative Surface Area (RSA) of amino acid. (a) Monogenic Disorder Dataset 1 (no likely cases) (b) Monogenic Disorder Dataset 2 (including likely cases).



(a)



(b)

Figure S7. (a). Plots of $\Delta\Delta G$ against RSA for Monogenic Disorder Dataset 1 (no likely cases). (b). Plots of $\Delta\Delta G$ against RSA for Monogenic Disorder Dataset 2 (including likely cases).

Supplementary Tables

Table S1. Total number of amino acid mutations based on chemical nature of amino acid mutations in monogenic disorder database.

Se. No.	Type of amino acid mutation	Dataset 1			Dataset 2		
		Benign	Pathogenic	Total	Benign	Pathogenic	Total
1.	Hydrophobic-Hydrophobic	424	460	884	528	605	1133
2.	Hydrophobic-Polar	268	529	797	333	685	1018
3.	Polar-Polar	436	465	901	548	626	1174
4.	Polar-Hydrophobic	277	480	757	354	643	997
5.	Small-Small	478	463	941	609	592	1201
6.	Small-Large	307	494	801	384	658	1042
7.	Large-Large	370	514	884	470	709	1179
8.	Large-Small	250	463	713	300	600	900
9.	Aliphatic-Aliphatic	352	301	653	441	382	823
10.	Aliphatic-Aromatic	25	34	59	33	47	80
11.	Aromatic-Aromatic	24	18	42	29	26	55
12.	Aromatic-Aliphatic	17	36	53	20	58	78
13.	Positive-Positive	88	69	157	104	103	207
14.	Positive-Negative	19	21	40	23	24	47
15.	Negative-Negative	26	15	41	36	15	51
16.	Negative-Positive	23	67	90	33	90	123

Table S2. Total number of amino acid mutations based on functional annotation of proteins in monogenic disorder database.

Se. No.	Functional Class	No. of entries in Dataset 1	No. of entries in Dataset 2
1.	Enzyme	1295	1646
2.	Transport/translocation/cargo Protein	445	570
3.	Transcription Regulation	343	420
4.	Structural Support	224	275
5.	Scaffold Protein	152	205
6.	Receptor Protein	141	210
7.	Signalling Protein	128	144
8.	Regulatory Protein	112	138
9.	DNA Binding	90	151
10.	Motor Protein	87	152
11.	Secretory Proteins	74	80
12.	Adhesion Protein	66	87
13.	Chaperons	66	74
14.	Membrane Protein	32	40
15.	RNA Binding	31	50
16.	Antigen-Antibody	14	18
17.	Miscellaneous	39	62