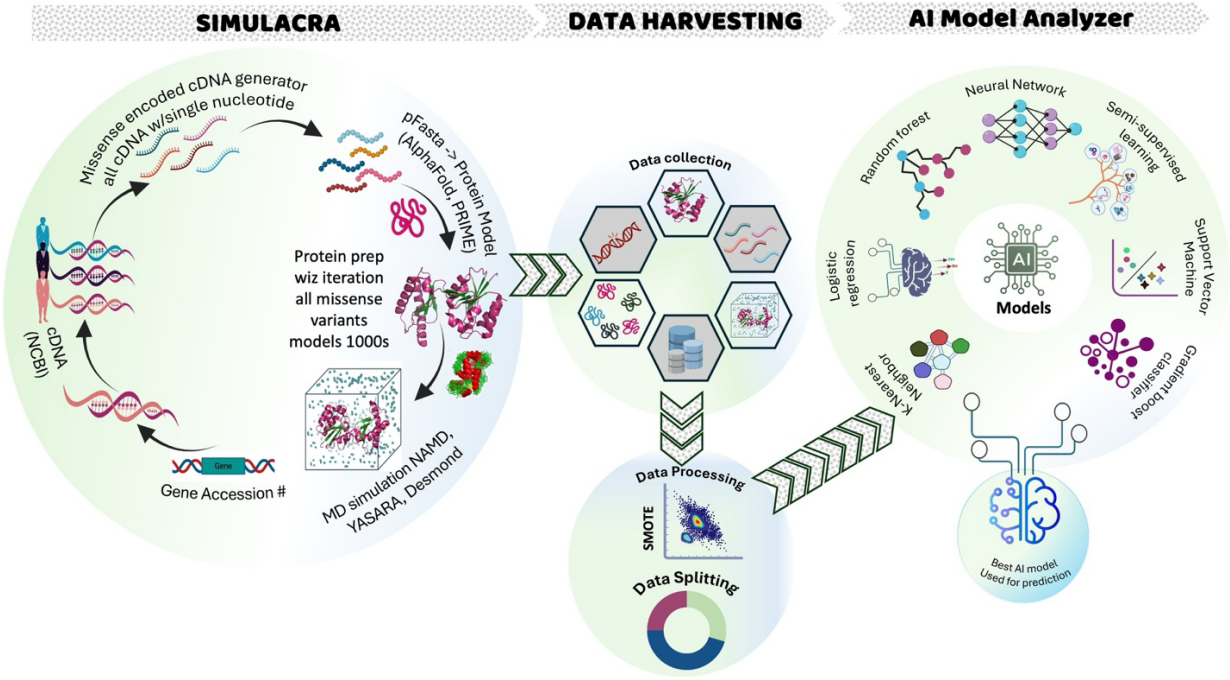
Supplemental Figures and Data.

Schematic S1. Overview of the Gene-to-Protein through Data Harvesting and AI Differentiation.



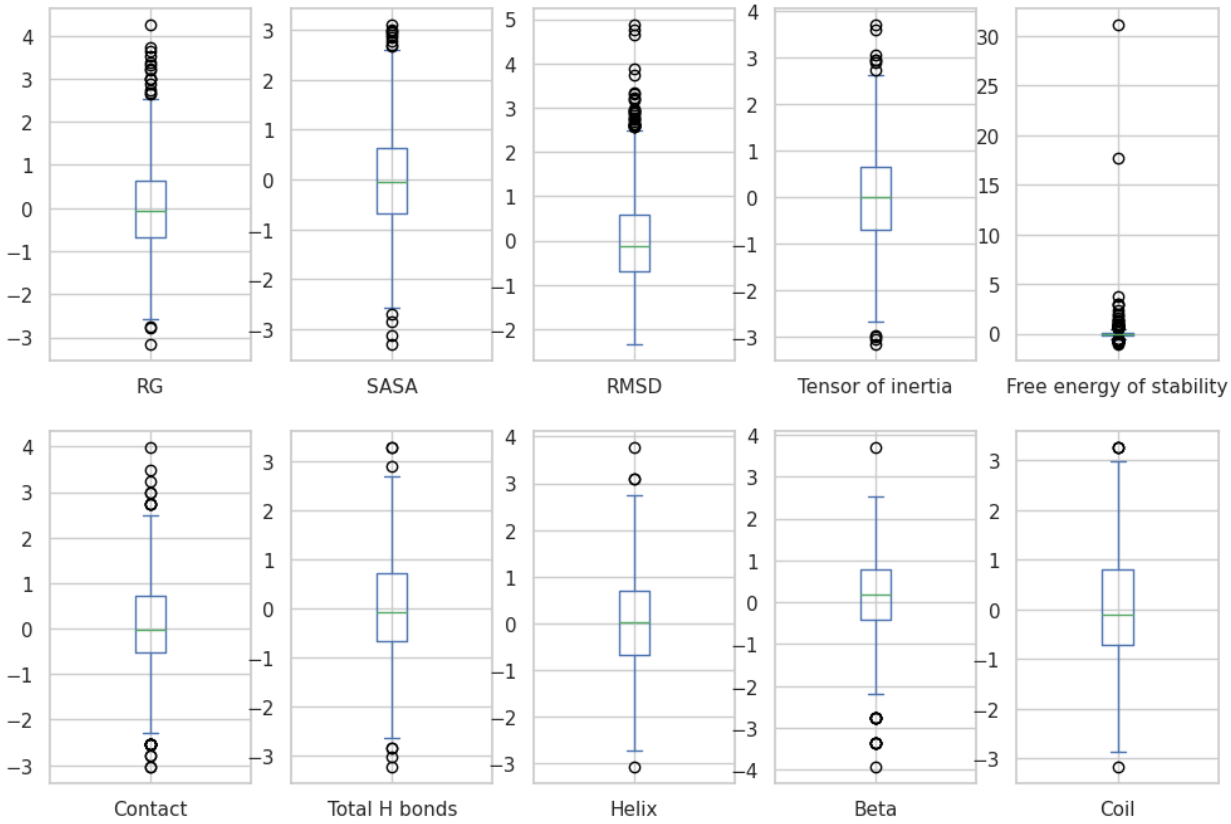


Fig. S1. MDS-extracted features show varied distributions across PMM2 mutations. Box plots of the distribution of indicated MDS-extracted features across all PMM2 missense mutations. Each plot depicts the range as a line, the interquartile range as a box, the median as a horizontal line inside the box, and outliers as circles. “Contact” refers to all amino acid contacts; “H bonds” refers to hydrogen bonds; “Helix”, “Beta”, and “Coil” refer to secondary structures.

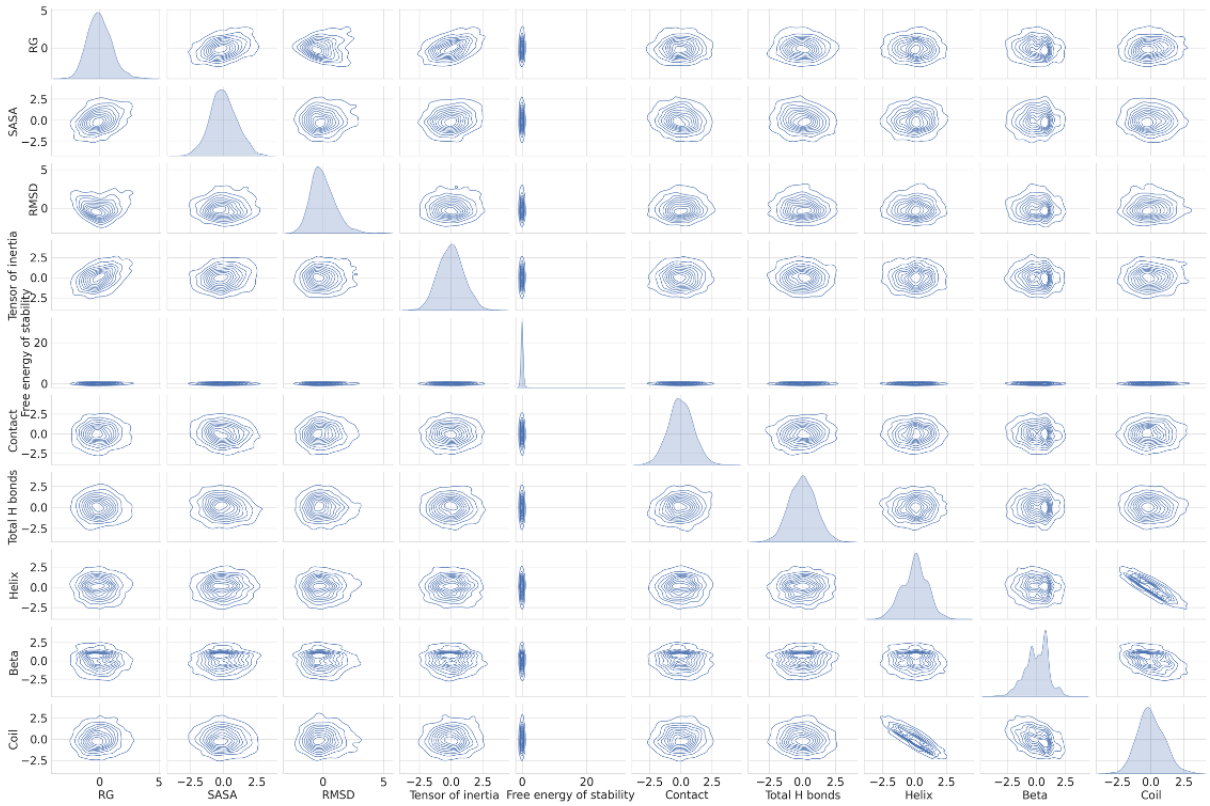


Fig. S2. Pair plots reveal correlations between MDS-extracted features of PMM2 variants. Scaled readouts of each feature extracted from MDS of PMM2 variants are plotted against each other to reveal correlations.

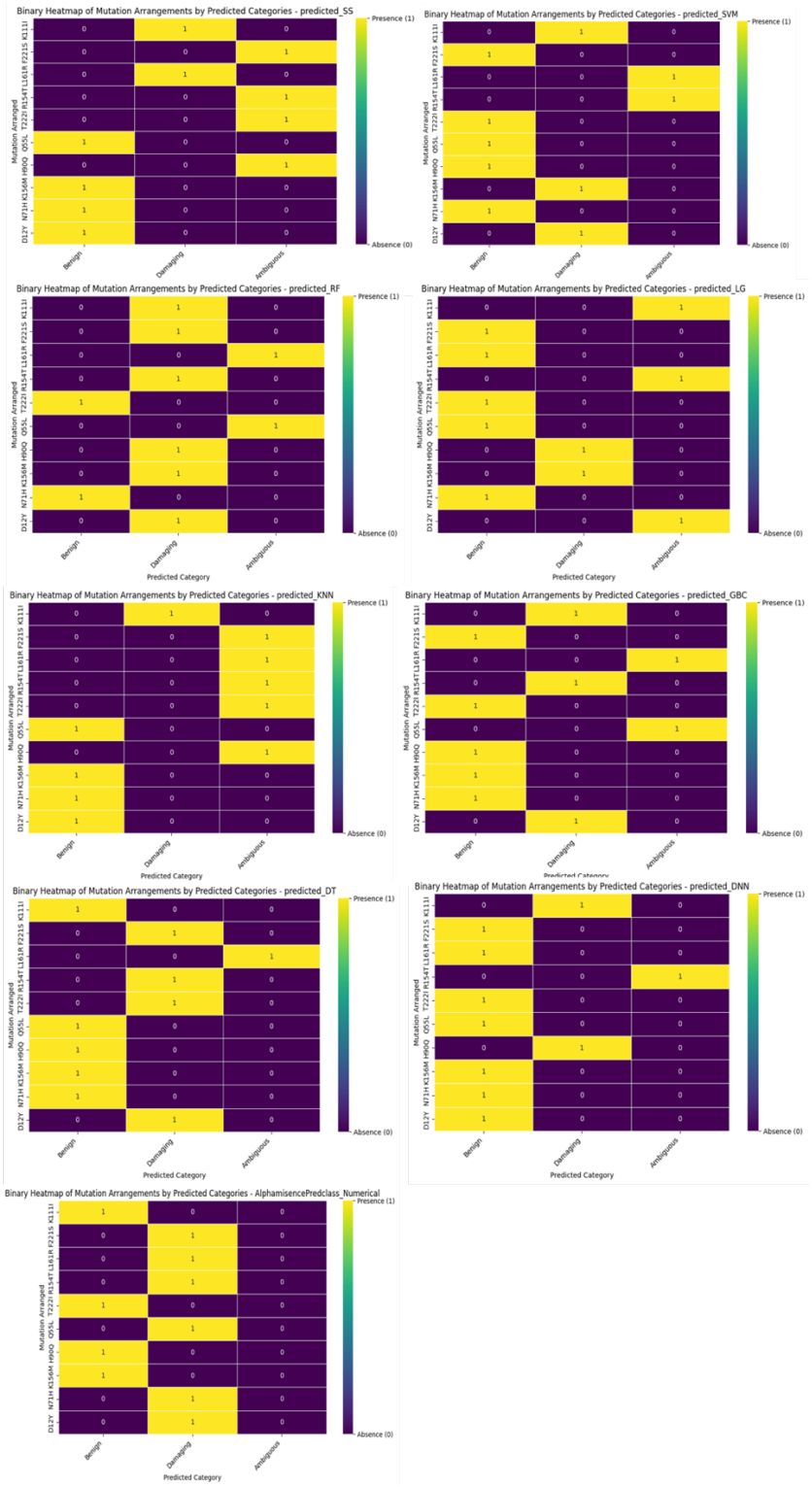


Fig. S3. Advanced AI models show variations in their predictions for individual PMM2 mutations. Binary heat maps display how indicated advanced AI models called a subset of PMM2 mutations of unknown significance. Yellow denotes the category called by the model (benign, damaging, or ambiguous), while purple denotes the other two categories not called.

Table S1. Clinical significance of all labeled PMM2 missense variants in ClinVar.

aa in WT PMM2	Residue Number in PMM2	aa post-mutation	ClinVar Classification	Second ClinVar Classification
C	9	F	LP	
C	9	Y	P	LP
T	18	S	P	LP
R	21	G	P	
L	32	V	LP	
L	32	R	P	
Q	37	L	B	LB
V	43	M	LP	
V	44	A	P	LP
D	65	Y	P	
P	69	T	LP	
N	101	K	P	
L	104	V	P	
A	108	V	P	LP
P	113	L	P	
P	113	T	LP	
G	117	R	LP	
F	119	L	P	
I	120	M	LP	
I	120	T	P	LP
R	123	Q	P	LP
V	129	L	P	
V	129	M	P	LP
P	131	A	P	LP
I	132	T	P	
I	132	F	P	
E	139	K	P	LP
R	141	H	P	LP
F	144	C	LP	
F	144	V	LP	
F	144	L	P	LP
D	148	N	P	LP
I	153	T	P	LP
F	157	C	LP	

F	157	S	P	LP
R	162	P	LP	
R	162	W	P	LP
G	176	V	LP	
G	176	S	LP	
F	183	S	P	
G	186	R	P	
D	188	E	LP	
D	188	G	P	LP
D	188	Y	LP	
H	195	R	LP	
E	197	A	B	LB
F	207	S	P	LP
G	208	A	P	LP
N	216	I	P	LP
H	218	D	LP	
D	223	E	P	
D	223	N	P	
T	226	S	P	LP
G	228	C	P	LP
V	231	M	P	
T	237	M	P	LP
T	237	R	P	LP
R	238	P	P	LP
R	239	S	LP	
R	239	W	P	
C	241	W	LP	
C	241	S	P	

aa: amino acid; P: pathogenic; LP: likely pathogenic; LB: likely benign; B: benign. Note: some mutations were submitted to ClinVar multiple times, leading to a second classification of pathogenicity in some instances. If these classifications conflicted, the mutation was considered of uncertain significance. Mutations labeled as uncertain were omitted from this table.

Table S2. **Allele frequency, heterozygosity, and ClinVar clinical significance of PMM2 variants in gnomAD deemed benign by our evaluation.**

aa in WT PMM2	Residue Number in PMM2	aa Post-mutation	ClinVar Classification	Second ClinVar Classification	Allele Frequency	Number of Homozygotes in gnomAD
E	197	A	B	LB	0.023	535
Q	37	L	B	LB	0.000549	16
A	228	V	LB		0.000902	8
D	30	E	VUS		0.00038	8
R	238	C	VUS		0.000556	8
M	212	V	VUS		0.000473	4
M	227	T	VUS		0.0000116	2
V	196	M	VUS		0.0000279	1
E	219	D	VUS		0.0000756	1
V	182	I	-		0.0000116	1
H	221	Q	-		0.0000716	1

aa: amino acid; LB: likely benign; B: benign; VUS: variant of unknown significance.

Table S3. Performance comparison of various machine learning models in multi calss mutation prediction, evaluated using F1 score, precision, and recall. RF and SSL exhibit strong, balanced performance, while LR and KNN show weaker results. Benchmark models REVEL, PROVEN, and Alphamissence are also included for comparison.

Models	F1 Score	Precision	Recall
DT	0.655	0.655	0.661
GBC	0.778	0.778	0.779
KNN	0.629	0.713	0.661
LR	0.413	0.396	0.423
RF	0.804	0.841	0.813
SSL	0.803	0.813	0.836
SVM	0.700	0.712	0.711
DNN	0.774	0.801	0.779
AlphaMissense	0.682	0.660	0.793

PROVEN	0.769	0.773	0.754
REVEL	0.853	0.869	0.861