Supplementary Information

An augmented transformer model trained on protein family specific variant data leads to improved prediction of variants of uncertain significance

Dinesh Joshi, Swatantra Pradhan, Rakshanda Sajeed, Rajgopal Srinivasan and Sadhna Rana*

TCS Research, Tata Consultancy Services, India

*Corresponding author: sadhna.rana@tcs.com

Methods

Confidence Interval Computation: Confidence interval of performances were computed using bootstrap estimates. We generated 1000 bootstrap samples (bootstrap samples of the same size were generated using sampling with replacement strategy) and then calculated the perfomance metric on each of the 1000 bootstrap samples. For a given performance metric, we reported the gaussian approximation based 95% confidence interval calculated as 1.96 * standard deviation obtained from the boostrap estimates (Jain et. al, 2024). The metrics for our best model, ESM models and all other state of the art methods shown in the ROC curve, Truncated ROC curve, Precision-Recall curve and bar plots are given in the tabular format along with the confidence interval values.

Statistical significance analysis using one-sided binomial test: To check whether the difference in performance was statistically significant with the state of the art models, a one-sided binomial test was used with the number of wins on 1000 bootstrap samples (bootstrap samples of same size were generated using sampling with replacement strategy) as the test statistic. The number of wins considered as the number of times the bootstrap samples performance metric is greater than the performance metrics of the model being compared. The p-value was computed as the probability that the Binomial (0.5, 1000) variable is greater than or equal to the number of wins (Jain et. al, 2024).

Packages used for developing the models, getting metrics , plotting the curves and computing the P-values for the binomial test

Deep neural network models were developed using PyTorch (torch version: 1.12.0, https://pytorch.org/) (Paszke et al. 2019). For calculating area under the curve (AUROC, AUPRC,

Truncated ROC) metrics and mathews correlation coefficient metrics (MCC) we used sklearn python package (scikit-learn version: 0.23.1, https://scikit-learn.org/stable/)(Pedregosa et al. 2011) and for calculating the correlation metrics (Pearson's correlation coefficient, Kendall's Tau) and the P-value for the binomial test used the SciPy package stats module (scipy version: 1.7.3, https://scipy.org/) (Virtanen et al. 2020). For plotting the curves and bar plots we used the MatPlotlib package pyplot module (matplotlib version: 3.4.0, https://matplotlib.org/)(Hunter 2007).

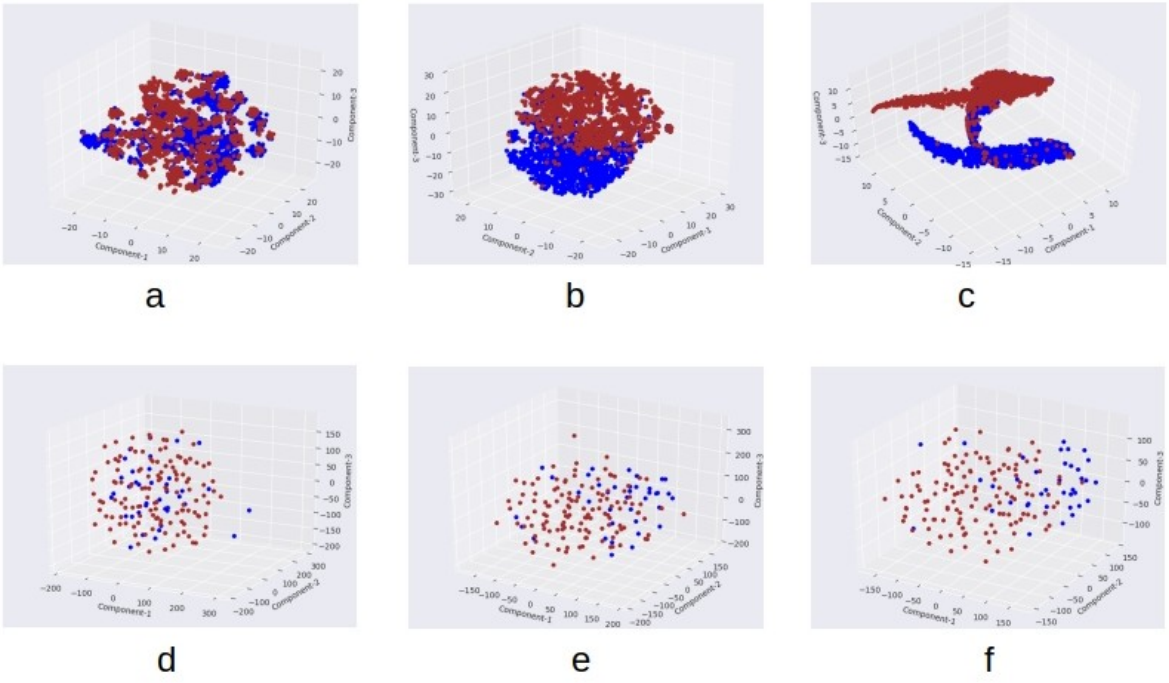

Fig. S1: 3D t-distributed stochastic neighbor embedding (tSNE) of variants from carbohydrate training set (upper row) and NAGLU VUS test set (lower row)

The plots show classifier performance in separating pathogenic (blue color) and benign variants (brown color) in both training and test set. (a,d) before classification (b,e) after zero-shot classification (c,f) family specific classifier.

Table S1: Zero-shot predictions from different ESM models on NAGLU VUS

Models	AUROC	AUPRC
ESM-1b	0.872	0.687
ESM-2	0.889	0.725
ESM-1v	0.863	0.698

Table S2: Hyperparameters used for different models using residue level embeddings (ARSA)

Training Dataset	Total data points*	Model	Model number	Number of Hidden layers (HL)	Hidden layer size	Dropout rate	Input Features
Dataset-1 (Sulfatase Family)	1116	Without PCA[a]	1	1	10	0.2	1281
			2	3	250,100,10	0.2	1281
			3	4	500,250,100,10	0.2	1281
			4	1	10	0.2	1281
			5	3	250,100,10	0.2	1281
		With PCA[b]	1	1	10	0.2	270
			2	3	120,60,10	0.2	268
			3	1	10	0.2	264
			4	2	100,10	0.2	263
			5	3	120,60,10	0.2	270
Dataset-2 (Carbohydrate Hydrolases + Sulfatse Family)	4026	Without PCA[a]	1	1	10	0.2	1281
			2	1	10	0.2	1281
			3	1	10	0.2	1281
			4	2	100,10	0.2	1281
			5	1	10	0.2	1281
		With PCA[b]	1	1	10	0.2	493
			2	4	240,120,60,10	0.2	492
			3	2	100,10	0.5	490
			4	3	250,100,10	0.2	493
			5	4	240,120,60,10	0.5	489
Dataset-3 (Control)	980	Without PCA[a]	1	1	10	-	1281
			2	2	100,10	0.2	1281
			3	1	10	0.2	1281
			4	1	10	-	1281
			5	2	100,10	0.2	1281
		With PCA[b]	1	3	120,60,10	-	247
			2	1	10	0.2	247
			3	2	100,10	0.2	244
			4	2	100,10	-	240
			5	1	10	-	245

*denotes total data points split into train and validation datasets in 80:20 ratio for each model . [a]denotes the Input (ESM-1b embeddings concatenated with log-odds feature) . [b]denotes the Input (ESM-1b embeddings obtained after doing PCA & concatenating log-odds feature). Dataset-3 contains all variants from sulfatase family genes except ARSA. Other hyperparameters used - Optimizer: AdamW, Learning rate: 1e-3, Train & Validation Batch sizes: 32, Patience value: 10

Table S3: Hyperparameters used for different models using residue level / sequence level (N/S) embedding (NAGLU)

Training Dataset	Total data points*	Model	Representation (N/S)+	Model number	Number of Hidden layers (HL)	Hidden layer size	Dropout rate	Input Features
Carbohydrate hydrolases	3276	Without PCA[a]	N	1	2	100,10	0.5	1281
				2	1	10	0.2	1281
				3	2	100,10	0.2	1281
				4	3	250,100,10	0.2	1281
				5	2	100,10	-	1281
			S	1	2	100,10	0.2	1281
				2	3	250,100,10	0.2	1281
				3	4	500,250,100,10	0.5	1281
				4	2	100,10	0.2	1281
				5	2	100,10	0.2	1281
		With PCA[b]	N	1	3	120,60,10	0.5	482
				2	4	240,120,60,10	0.2	484
				3	4	240,120,60,10	0.5	483
				4	3	120,60,10	0.2	482
				5	2	60,10	-	482
			S	1	4	120,60,30,10	0.2	191
				2	4	120,60,30,10	0.2	190
				3	3	60,30,10	-	189
				4	1	10	-	188
				5	3	60,30,10	-	190
Hydrolases	3526	Without PCA[a]	N	1	1	10	-	1281
				2	3	250,100,10	0.5	1281
				3	2	100,10	-	1281
				4	4	500,250,100,10	0.2	1281
				5	2	100,10	0.5	1281
			S	1	3	250,100,10	0.2	1281
				2	1	10	0.2	1281
				3	4	500,250,100,10	0.5	1281
				4	1	10	0.2	1281
				5	2	100,10	0.2	1281
		With PCA[b]	N	1	3	120,60,10	0.5	495
				2	3	120,60,10	0.5	496
				3	1	10	0.2	495
				4	4	240,120,60,10	0.5	495
				5	4	240,120,60,10	0.5	494
			S	1	1	60,30,10	-	198
				2	2	30,10	-	196
				3	2	30,10	-	196
				4	1	10	0.2	195
				5	4	120,60,30,10	0.2	197
Control	3340	Without PCA[a]	N	1	4	500,250,100,10	0.2	1281
				2	2	100,10	-	1281
				3	2	100,10	-	1281
				4	2	100,10	-	1281
				5	1	10	0.2	1281
			S	1	2	100,10	-	1281
				2	1	10	-	1281
				3	2	100,10	0.2	1281
				4	4	500,250,100,10	0.2	1281
				5	2	100,10	-	1281
		With PCA[b]	N	1	4	240,120,60,10	0.2	488
				2	3	120,60,10	0.5	490
				3	4	240,120,60,10	0.5	488
				4	3	120,60,10	0.2	488
				5	4	240,120,60,10	0.5	488
			S	1	2	30,10	-	195
				2	1	10	-	194
				3	4	120,60,30,10	-	194
				4	1	10	-	193
				5	1	10	-	195

[+]N/S denotes the residue level (node level) embedding and sequence level embedding respectively. [*]denotes total data points split into train and validation datasets in 80:20 ratio for each model . [a]denotes the Input (ESM-2 embeddings concatenated with log-odds feature) . [b]denotes the Input (ESM-2 embeddings obtained after doing PCA & concatenating log-odds feature). Other hyperparameters used - Optimizer: AdamW, Learning rates: 1e-3, Train & Validation Batch sizes: 32, Patience value: 10

Table S4 : Prediction performance of models on CAGI6 ARSA evaluation test set data (221 variants)

Training Dataset Identifier	Model	AUROC	AUPRC	MCC	Truncated ROC
Dataset-1	Without PCA[a]	0.807	0.718	**0.522**	0.484
	With PCA[b]	0.81 ± 0.066	0.731 ± 0.097	0.497 ± 0.119	0.499 ± 0.134
Dataset-2	Without PCA[a]	0.802	0.696	0.495	0.436
	With PCA[b]	0.805	0.714	0.472	0.471
Dataset-3	Without PCA[a]	0.80	0.718	0.467	0.482
	With PCA[b]	0.806 ± 0.068	0.734 ± 0.097	0.495 ± 0.118	0.512 ± 0.129
	Sift	0.777 ± 0.066	0.691 ± 0.087	0.438 ± 0.119	0.328 ± 0.13
	MutPred2	0.797 ± 0.062	0.687 ± 0.103	0.422 ± 0.117	0.412 ± 0.12
	Snap2	0.77 ± 0.071	0.639 ± 0.118	0.416 ± 0.125	0.419 ± 0.128
	PolyPhen2	0.807 ± 0.058	0.727 ± 0.078	0.368 ± 0.092	0.375 ± 0.125
	EVE[+]	0.804 ± 0.058	0.655 ± 0.112	0.375 ± 0.162	0.39 ± 0.139
	REVEL	0.805 ± 0.062	0.65 ± 0.116	0.42 ± 0.1	0.387 ± 0.14
	MetaRNN	0.825 ± 0.058	0.708 ± 0.11	0.458 ± 0.131	0.464 ± 0.124
	CinPred	0.838 ± 0.056	**0.738 ± 0.094**	0.46 ± 0.108	0.486 ± 0.129
	VEST4	0.793 ± 0.064	0.64 ± 0.118		0.39 ± 0.138
	AlphaMissense	**0.859 ± 0.053**	0.729 ± 0.111	0.633[*] ± 0.115	**0.53 ± 0.152**
	ESM-1b	0.807 ± 0.068	0.725 ± 0.105	0.481 ± 0.12	0.507 ± 0.129
	ESM-1v	0.702 ± 0.085	0.669 ± 0.096	0.42 ± 0.094	0.441 ± 0.116
	ESM-2	0.761 ± 0.077	0.706 ± 0.096	0.476 ± 0.123	0.483 ± 0.119

[a]denotes the input (ESM-1b embeddings concatenated with log-odds feature). [b]denotes the input (ESM-1b embeddings obtained after doing PCA & concatenating log-odds feature from zero-shot predictions). Dataset-3 denotes model trained on sulfatase family without ARSA variants. Binary predictions along with the associated scores were collected for the methods PolyPhen2 (‘Probably damaging’ and ‘Possibly damaging’ classes were merged and assigned as pathogenic) (Adzhubei et al. 2013), SIFT (‘Tolerated’ and ‘Tolerated_low_confidence’ classes were merged and assigned as benign) (Ng and Henikoff 2003), SNAP2 (Hecht et al. 2015), MetaRNN (Li et al. 2022), AlphaMissense (Cheng et al. 2023) and EVE (Frazer et al. 2021). For MutPred2 (Pejaver et al. 2020) , REVEL (Ioannidis et al. 2016) and ClinPred (Alirezaie et al. 2018) pathogenicity scores greater than 0.5 were classified as pathogenic else as benign while calculating the MCC metrics. VEST4 scores (Carter et al. 2013) were collected from the CAGI6 ARSA assessment paper supplementary data (Jain et. al, 2024). [*]AlphaMissense uses thresholds that classify variants as ‘likely_pathogenic’, ‘ambiguous’ and ‘likely_benign’, hence for AlphaMissense while calculating the MCC metrics 30 variants classified as ‘ambiguous’ were removed from the evaluation set. [+]For EVE model pathogenicity scores (EVE Score column) for 7 variants were not available, so they were removed from the evaluation set for metric calculation . EVE (EVE Class column) classifies

the variants into three classes namely ‘Benign’, ‘Pathogenic’ or ‘Uncertain’. 68 variants classified as ‘Uncertain’ were further removed to calculate the MCC metrics.

Table S5: Correlation Metrics of models on CAGI6 ARSA evaluation test set data (221 variants)

Model	Pearson’s Correlation	Kendalls’s Tau
Best Model	0.503 ± 0.099	0.369 ± 0.08
Control	0.509 ± 0.096	0.362 ± 0.083
Sift	0.435 ± 0.109	0.378 ± 0.079
MutPred2	0.512 ± 0.105	0.373 ± 0.077
Snap2	0.499 ± 0.103	0.359 ± 0.082
PolyPhen2	0.521 ± 0.106	0.44 ± 0.072
EVE[+]	0.527 ± 0.085	0.411 ± 0.072
REVEL	0.515 ± 0.102	0.38 ± 0.074
MetaRNN	0.551 ± 0.097	0.403 ± 0.071
ClinPred	0.553 ± 0.098	0.423 ± 0.068
VEST4	0.523 ± 0.098	0.373 ± 0.075
AlphaMissense	0.614 ± 0.063	0.468 ± 0.065
ESM-1b	0.476 ± 0.094	0.363 ± 0.082
ESM-1v	0.362 ± 0.098	0.226 ± 0.091
ESM-2	0.433 ± 0.097	0.316 ± 0.088

Best Model: Dataset-1(Sulfatase family (ESM-1b + Zero-shot, with PCA)). Control denotes model trained on sulfatase family without ARSA variants with (ESM-1b + Zero-shot, with PCA) [+]denotes the metrics for variants classified as ‘Benign’ or ‘Pathogenic’. [+]For EVE model pathogenicity scores (EVE Score column) for 7 variants were not available, so they were removed from the evaluation set for metric calculation.

Table S6: Statistical significance using one-sided binomial test for best model with the state of the art pathogenicity predictors on CAGI ARSA evaluation set.

Test set	Binomial Test	AUROC	Average Precision	AUPRC	Truncated ROC	PCC	Kendall's Tau	MCC*
ARSA Evaluation (with ClinPred)	n_wins	218	485	476	**591**	158	96	745
	P-value	1.0	0.837	0.939	**4.760e-09**	1.0	1.0	1.543e-56
ARSA Evaluation (with MetaRNN)	n_wins	357	**694**	**696**	**705**	165	208	**756**
	P-value	1.0	**1.33e-35**	**2.547e-36**	**1.188e-39**	1.0	1.0	**8.33e-62**
ARSA Evaluation (with MutPred2)	n_wins	**655**	**828**	**821**	**908**	445	477	**975**
	P-value	**3.388e-23**	**9.009e-104**	**4.541e-99**	**1.11e-169**	1.0	0.931	**4.563e-252**
ARSA Evaluation (with PolyPhen2)	n_wins	**531**	**985**	**560**	**965**	378	37	**980**
	P-value	**0.0268**	**6.521e-269**	**8.252e-05**	**5.133e-237**	1.0	1.0	**3.234e-260**

Bold entries indicates that the difference in performance is statistically significant for the particular metric. To check whether the difference in performance of our best models with the state of the art predictors like ClinPred, MetaRNN, MutPred2, PolyPhen2 and AlphaMissense was statistically significant or not , statistical significance analysis was done using one-sided binomial test with number of wins on 1000 bootstrap samples (boostrap samples of the same size were generated using sampling with replacement strategy) as the test statistic (Jain et. al, 2024) . *For NAGLU Best model vs AlphaMissense statistical significance test , MCC metric (Best Model: 0.474 , AlphaMissense: 0.633) for the test set excludes the 16 variants classified as 'ambiguous' by AlphaMisssense.

Table S7: Prediction performance of models on NAGLU VUS

Training Dataset	Model	Representation (N/S)[+]	AUROC	AUPRC	MCC	Truncated ROC
Carbohydrate hydrolases (family)	Without PCA[a]	N	0.871	0.738	0.472	0.591
		S	0.882	0.74	0.487	0.607
	With PCA[b]	N	**0.888 ± 0.065**	**0.77 ± 0.123**	0.472 ± 0.106	**0.627 ± 0.156**
		S	0.887	0.742	0.453	0.625
Hydrolases (super family)	Without PCA[a]	N	0.867	0.714	0.449	0.565
		S	0.885	0.732	0.475	0.616
	With PCA[b]	N	0.88	0.758	0.478	0.621
		S	0.882	0.738	0.439	0.613
Control	Without PCA[a]	N	0.883	0.764	0.465	0.618
		S	0.879	0.734	0.506	0.597
	With PCA[b]	N	0.875 ± 0.071	0.754 ± 0.124	0.476 ± 0.114	0.598 ± 0.153
		S	**0.889**	0.728	0.478	0.622
	PolyPhen2		0.778 ± 0.088	0.612 ± 0.13	0.303 ± 0.122	0.393 ± 0.149
	Snap2		0.774 ± 0.082	0.552 ± 0.155	0.295 ± 0.149	0.397 ± 0.148
	Sift		0.819 ± 0.069	0.633 ± 0.121	0.462 ± 0.114	0.371 ± 0.148
	MutPred2		0.858 ± 0.077	0.716 ± 0.135	**0.522 ± 0.135**	0.574 ± 0.158
	EVE[+]		0.627 ± 0.097	0.297 ± 0.108	0.615 ± 0.165	0.125 ± 0.109
	REVEL[#]		0.808 ± 0.12	0.475 ± 0.24	0.397 ± 0.145	0.46 ± 0.221
	MetaRNN		0.851 ± 0.078	0.647 ± 0.171	0.417 ± 0.114	0.561 ± 0.166
	ClinPred		0.832 ± 0.083	0.664 ± 0.151	0.388 ± 0.135	0.542 ± 0.165
	AlphaMissense		0.855 ± 0.078	0.719 ± 0.152	0.633[*] ± 0.157	0.613 ± 0.154
	ESM-1b		0.872 ± 0.071	0.687 ± 0.167	0.488 ± 0.164	0.623 ± 0.159
	ESM-1v		0.863 ± 0.071	0.698 ± 0.145	0.505 ± 0.158	0.583 ± 0.152
	ESM-2		0.889 ± 0.063	0.725 ± 0.151	0.60 ± 0.149	0.631 ± 0.15

[a]denotes the input (ESM-2 embeddings concatenated with log-odds feature). [b]denotes the input (ESM-2 embeddings obtained after doing PCA & concatenating log-odds feature from zero-shot predictions). [+]N and S denote residue (node) level and sequence level embeddings respectively. Control denotes model trained on hydrolases without NAGLU variants. Binary predictions along with the associated scores were collected for the methods PolyPhen2 ('Probably damaging' and 'Possibly damaging' classes were merged and assigned as pathogenic) (Adzhubei et al. 2013), SIFT ('Tolerated' and 'Tolerated_low_confidence' classes were merged and assigned as benign) (Ng and Henikoff 2003), SNAP2 (Hecht et al. 2015), MetaRNN (Li et al. 2022), AlphaMissense (Cheng et al. 2023) and EVE (Frazer et al. 2021). For MutPred2 (Pejaver et al. 2020), REVEL (Ioannidis et al. 2016) and ClinPred (Alirezaie et al. 2018) pathogenicity scores greater than 0.5 were classified as pathogenic else as benign while calculating the MCC metrics. [*]AlphaMissense uses thresholds that classify variants as 'likely_pathogenic', 'ambiguous' and 'likely_benign', hence for AlphaMissense while calculating the MCC metrics 16 variants classified as 'ambiguous' were removed from the test set. [+]For EVE

model pathogenicity scores (EVE Score column) for 15 variants were not available, so they were removed from the test set for metric calculation . EVE (EVE Class column) classifies the variants into three classes namely ‘Benign’, ‘Pathogenic’ or ‘Uncertain’. 23 variants classified as ‘Uncertain’ were further removed to calculate the MCC metrics. [#]For REVEL pathogenicity scores for 52 variants were not available, so they were removed from the test set for metric calculation.

Table S8: Correlation Metrics of models for NAGLU VUS (164 variants)

Model	Pearson’s Correlation	Kendalls’s Tau
Best Model	0.584 ± 0.118	0.442 ± 0.09
Control	0.568 ± 0.123	0.437 ± 0.092
Sift	0.322 ± 0.142	0.346 ± 0.091
PolyPhen2	0.421 ± 0.144	0.357 ± 0.1
MutPred2	**0.606 ± 0.109**	0.428 ± 0.086
Snap2	0.459 ± 0.127	0.334 ± 0.094
EVE[+]	0.272 ± 0.151	0.183 ± 0.099
REVEL[#]	0.482 ± 0.163	0.343 ± 0.12
MetaRNN	0.503 ± 0.132	0.377 ± 0.096
ClinPred	0.471 ± 0.137	0.36 ± 0.101
AlphaMissense	0.56 ± 0.09	0.441 ± 0.087
ESM-1b	0.570 ± 0.12	0.432 ± 0.092
ESM-1v	0.549 ± 0.128	0.411 ± 0.095
ESM-2	0.594 ± 0.115	**0.448 ± 0.087**

Best Model: Carbohydrate Hydrolases (ESM-2 + Zero-shot, with PCA)). Control denotes model trained on hydrolases without NAGLU variants (ESM-2 + Zero-shot, with PCA, Representation: N). [+]For EVE model pathogenicity scores (EVE Score column) for 15 variants were not available, so they were removed from the test set for metric calculation. [#]For REVEL pathogenicity scores for 52 variants were not available, so they were removed from the test set for metric calculation.

Table S9: Statistical significance using one-sided binomial test comparing different models on NAGLU test set

	Binomial Test	AUROC	AUPRC	Truncated ROC	PCC	Kendall's Tau	MCC*
Best Model vs ESM-1b	n_wins	**712**	**877**	**533**	**614**	**583**	384
	P-value	**2.33e-42**	**3.553e-141**	**0.02**	**2.815e-13**	**8.509e-08**	1.0
Best Model vs ESM-2	n_wins	498	**732**	491	465	461	5
	P-value	0.563	**1.146e-50**	0.726	0.988	0.99	1.0
Best model vs AlphaMissense	n_wins	**806**	**778**	**572**	**666**	525	4
	P-value	**1.79e-89**	**3.349e-73**	**2.957e-06**	**2.147e-26**	0.061	1.0
ESM-2 vs AlphaMissense	n_wins	**854**	528	**567**	**714**	**577**	**838**
	P-value	**1.391e-122**	0.041	**1.267e-05**	**3.762e-43**	**6.255e-07**	**9.492e-111**

Bold entries indicates that the difference in performance is statistically significant for the particular metric . Best Model: Carbohydrate Hydrolases (ESM-2 + Zero-shot, with PCA)). To check whether the difference in performance of our best models with the ESM zero-shot models was statistically significant or not , statistical significance analysis was done using one-sided binomial test with number of wins on 1000 bootstrap samples (boostrap samples of the same size were generated using sampling with replacement strategy) as the test statistic (Jain et. al, 2024). *denotes MCC metric (Best Model: 0.472, ESM-1b: 0.488, ESM2: 0.60) calculated using ESM-1b/ESM-2 threshold score of -7.5. For ESM-2 vs AlphaMissense statistical significance test , MCC metric (ESM-2: 0.703, AlphaMissense: 0.633) for the test set excludes the 16 variants classified as 'ambiguous' by AlphaMisssense.

Table S10: Zero-shot predictions from ESM-2-Evotuned models on NAGLU VUS data

Models	AUROC	AUPRC	MCC	Truncated ROC
ESM-2-Evotuned[a]	0.874	0.695	0.547	0.605
ESM-2-Evotuned[b]	0.483	0.217	0	0.065
ESM-2-Evotuned[c]	0.454	0.208	0	0.082

ESM-2-Evotuned model is ESM-2 model fine-tuned using homologous sequences of NAGLU. [a]ESM-2-Evotuned model trained with hyperparameters as AdamW optimizer, learning rate as 1e-5, batch size as 2 and number of epochs as 2. [b]ESM-2-Evotuned model trained with linear rate scheduler having hyperparameters as Adam optimizer with beta1 = 0.9 and beta2 = 0.98, L2 weight decay of 0.01 and the learning rate is warmed up over the first 2000 steps to a peak value of 4e-4 and then linearly decayed to one tenth of its peak value (4e-5) for rest of the training duration, and trained for 3 number of epochs with batch size 2. [c]ESM-2-Evotuned model trained with linear rate scheduler having hyperparameters as Adam optimizer with beta1= 0.9 and beta2 = 0.98, L2 weight decay of 0.01 and the learning rate is linearly decayed from a starting learning rate 4e-4 to ending learning rate 4e-5 and trained for 3 number of epochs with batch size 2.

Table S11 : Top 10 hard-to-predict pathogenic/benign variants NAGLU VUS Predictions

Variant	EA	True label	Best Model score[+]	Best Model label[+]	Alpha Missense score	Alpha Missense label	PolyPhen2 score	PolyPhen2 label	Snap2 score	Snap2 label	Sift score	Sift label	Mutpred2 score	MutPred2 label	DDMut
P283L	1.19	B	0.081	P	0.19	B	0.139	P	59	P	0.03	P	0.714	P	-0.3
K729R	0.08	P	0.860	B	0.078	B	0.999	B	-95	B	0.55	B	0.057	B	0.01
V369L	0.14	P	0.990	B	0.19	B	0.999	B	-85	B	1	B	0.318	B	-0.31
E251K	0.18	B	0.772	B	0.05	B	0.997	B	-96	B	1	B	0.046	B	-0.12
S449N	0.21	B	0.965	B	0.1	B	1	B	-93	B	1	B	0.261	B	-0.49
I290V	0.21	B	0.922	B	0.07	B	1	B	-97	B	0.34	B	0.031	B	-0.15
A627V	0.08	P	0.766	B	0.09	B	0.979	B	-80	B	0.2	B	0.127	B	0.07
G596C	1.02	B	0.581	B	0.37	ambiguous	0	P	-35	B	0	P	0.601	P	-0.72
T339P	0.05	P	0.040	P	0.49	ambiguous	0.963	B	-11	B	0.03	P	0.333	B	-2.36
R377H	0.06	P	0.226	P	0.10	B	0.223	P	-58	B	0.05	P	0.249	B	-2.04

Best Model: Carbohydrate Hydrolases (ESM-2 + Zero-shot, with PCA)) P – Pathogenic, B – Benign, EA: Experimental enzyme activity

References

Adzhubei I, Jordan DM, Sunyaev SR (2013) Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. Curr Protoc Hum Genet 76:7.20.1-7.20.41. https://doi.org/10.1002/0471142905.hg0720s76

Alirezaie N, Kernohan KD, Hartley T, et al (2018) ClinPred: Prediction Tool to Identify Disease-Relevant Nonsynonymous Single-Nucleotide Variants. Am J Hum Genet 103:474–483. https://doi.org/10.1016/j.ajhg.2018.08.005

Carter H, Douville C, Stenson PD, et al (2013) Identifying Mendelian disease genes with the Variant Effect Scoring Tool. BMC Genomics 14:S3. https://doi.org/10.1186/1471-2164-14-S3-S3

Cheng J, Novati G, Pan J, et al (2023) Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science 381:eadg7492. https://doi.org/10.1126/science.adg7492

Frazer J, Notin P, Dias M, et al (2021) Disease variant prediction with deep generative models of evolutionary data. Nature 599:91–95. https://doi.org/10.1038/s41586-021-04043-8

Hecht M, Bromberg Y, Rost B (2015) Better prediction of functional effects for sequence variants. BMC Genomics 16:S1. https://doi.org/10.1186/1471-2164-16-S8-S1

Hunter JD (2007) Matplotlib: A 2D Graphics Environment. Comput Sci Eng 9:90–95. https://doi.org/10.1109/MCSE.2007.55

Ioannidis NM, Rothstein JH, Pejaver V, et al (2016) REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am J Hum Genet 99:877–885. https://doi.org/10.1016/j.ajhg.2016.08.016

Jain S, Trinidad M, Nguyen TB, et al (2024) Evaluation of enzyme activity predictions for variants of unknown significance in Arylsulfatase A. 2024.05.16.594558

Li C, Zhi D, Wang K, Liu X (2022) MetaRNN: differentiating rare pathogenic and rare benign missense SNVs and InDels using deep learning. Genome Med 14:115. https://doi.org/10.1186/s13073-022-01120-z

Ng PC, Henikoff S (2003) SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res 31:3812–3814. https://doi.org/10.1093/nar/gkg509

Paszke A, Gross S, Massa F, et al (2019) PyTorch: an imperative style, high-performance deep learning library. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA

Pedregosa F, Varoquaux G, Gramfort A, et al (2011) Scikit-learn: Machine Learning in Python. J Mach Learn Res 12:2825–2830

Pejaver V, Urresti J, Lugo-Martinez J, et al (2020) Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. Nat Commun 11:5918. https://doi.org/10.1038/s41467-020-19669-x

Virtanen P, Gommers R, Oliphant TE, et al (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 17:261–272. https://doi.org/10.1038/s41592-019-0686-2