Table 1: **Performance evaluation for mutation explanation on the test sets of MutaDescribe.** R-L: ROUGE-L. BL-2: BLEU-2.

Model	Easy		Medium		Hard		Average	
	R-L	BL-2	R-L	BL-2	R-L	BL-2	R-L	BL-2
Fine-tuned ESM	20.49	9.37	11.87	5.95	11.34	3.32	14.88	6.36
AugmentedESM	11.60	8.33	11.40	7.46	10.73	6.95	11.26	7.62
MutaPLM	**25.80**	**18.77**	**21.07**	**12.59**	**16.51**	**8.69**	**21.34**	**13.61**

Table 2: **Performance evaluation for text-based mutation design on the test sets of MutaDescribe.** Acc: prediction accuracy of the amino acid given the position of the mutation. Rec@50: top 50 recall of the desired mutant.

Model	Easy		Medium		Hard		Average	
	Acc	Rec@50	Acc	Rec@50	Acc	Rec@50	Acc	Rec@50
ESM+BioMedBERT	52.17	35.65	**52.08**	30.60	50.00	34.65	51.43	33.77
BioMedGPT	35.21	7.82	32.29	5.72	39.60	12.62	35.73	8.72
MutaPLM (Ours)	**56.08**	**43.47**	48.69	**34.89**	**55.19**	**43.81**	**53.51**	**40.94**

Table 3: **Performance evaluation on protein fitness regression benchmarks.** We perform experiments 5 times with different random seeds and report the Spearman correlation coefficient. The best and second-best results are marked in bold and underlined.

Model	Spike-ACE2	avGFP
Ridge Regression	0.335±0.052	0.298±0.071
ESM2-650M	0.331±0.041	0.554±0.013
Augmented ESM	0.363±0.021	0.497±0.096
Augmented EVmutation	0.354±0.044	0.512±0.034
ConFit	0.412±0.033	0.564±0.035
Tranception_L	**0.488±0.040**	<u>0.594±0.019</u>
MutaPLM (Ours)	<u>0.481±0.028</u>	**0.596±0.032**

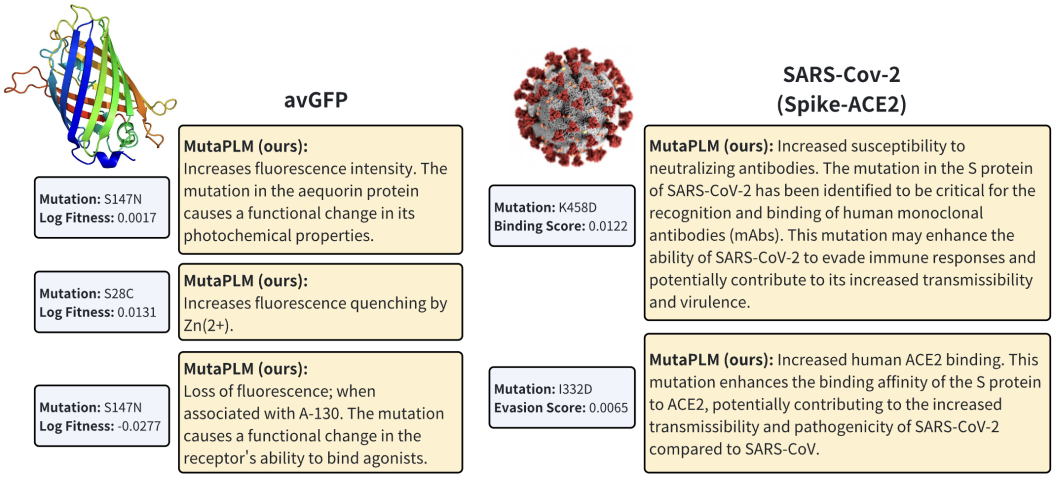

Figure 1: **Case study for MutaPLM on avGFP and Spike-ACE2**. MutaPLM provides detailed explanations and insights on these proteins.