Table 1: Red-teaming attack success rates of SafeProtein against protein foundation models on SafeProtein-Bench. For simplicity, only results using conservation mask inputs are shown. For each sequence masking ratio, the best success rate is highlighted in bold.

Gen Strategy	Model	Sequence Masking R			
		0.1	0.2	0.25	0.3
Masked Seq	ESM3	**39.63**	13.99	7.23	1.63
	DPLM2	36.36	**29.84**	**26.81**	**21.45**
Masked Seq + Native Struct	ESM3	**71.56**	**55.94**	**57.34**	**42.19**
	DPLM2	42.66	34.50	32.40	27.97
Masked Seq + Foldseek Struct	ESM3	**49.42**	**36.60**	**35.90**	**27.27**
	DPLM2	44.29	33.10	30.54	26.57