Table 1: Red-teaming attack success rates of SafeProtein against ESM3 on SafeProtein-Bench. Results are reported for all masking strategies. For each sequence masking ratio, the best success rate is highlighted in bold.

Gen Strategy	Masking Strategy	Sequence Masking Rat				
		0.1	0.2	0.25	0.3	
Strategy1	Conservation	**39.63**	**13.99**	**7.23**	1.63	0
	Random	19.35	6.53	6.53	**4.20**	**3**
	Tail	5.83	0.93	1.17	0.70	
Strategy2	Conservation	**71.56**	**55.94**	**57.34**	**42.19**	**39**
	Random	44.29	12.35	14.69	8.63	8
	Tail	34.03	7.69	9.79	3.73	4
Strategy3	Conservation	**49.42**	**36.60**	**35.90**	**27.27**	**22**
	Random	18.88	5.36	7.93	4.20	3
	Tail	6.99	1.87	2.10	1.17	2
Strategy4	Conservation	**72.49**	**63.64**	**64.10**	**46.85**	**43**
	Random	52.68	18.65	22.14	11.66	1
	Tail	42.89	10.72	15.15	7.93	1
Strategy5	Conservation	**75.06**	**75.06**	**74.36**	**72.26**	**72**
	Random	75.06	72.73	73.66	62.47	7
	Tail	74.13	65.73	66.43	52.21	6