SUPPLEMENTARY MATERIAL

The rvET optimized alignments when input into ivET or Shannon Entropy ranking method increased the average z-score $\langle z_o \rangle$ of functional overlap from 2.98 to 3.45 (16%) and from 3.61 to 3.82 (6%), respectively (see Figure 1). The ivET and Shannon Entropy optimized sequence selections also triggered improvements in rvET, but these were quite slight (1-4%). This may be explained by the intrinsic robustness of rvET compared to the coarseness of ivET and Shannon Entropy methods.

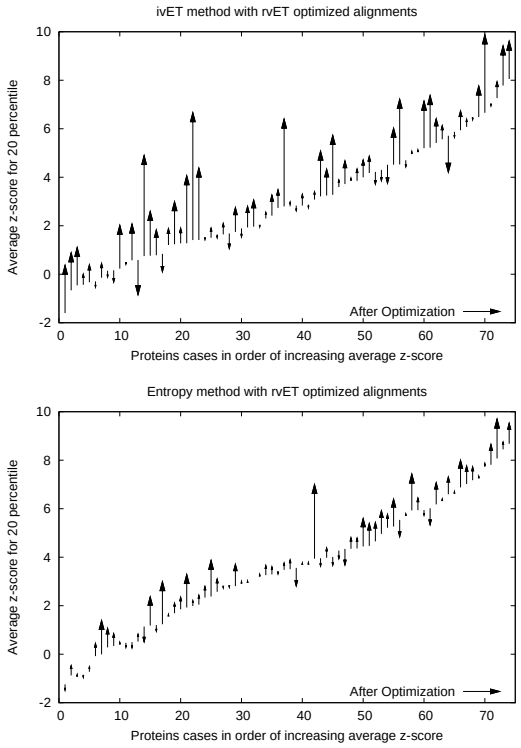

Figure 1: The ivET and Shannon Entropy ranking methods were tested on the rvET optimized alignments. The sequence selection optimization using the rvET also helps the control methods improve site prediction.

Testset 1		Number of sequences			
PDBID	unoptimized	optimized	PDBID	unoptimized	optimized
16pk	492	132	1dam	462	119
1a09	295	179	1dig	467	207
1a0oE	481	136	1dqr	467	169
1a22A	319	137	1dqx	152	81
1a22B	141	49	1e96A	439	144
1a2kA	124	77	1e96B	62	41
1a2kD	424	218	1ee9	164	109
1a3k	349	127	1efaB	470	73
1a48	425	121	1eg2	264	85
1a4mA	295	89	1eje	126	81
1a53	468	145	1elrA	385	182
1a59	478	216	1elwA	406	137
1a6m	344	212	1f6mA	488	180
1a6q	238	162	1f88A	260	118
1a80	482	199	1finA	444	163
1aca	352	219	1finB	417	204
1ad3A	459	224	1fjmA	422	167
1ai2	350	192	1fqjB	281	133
1aj2	471	143	1gnjA	92	41
1aj8A	475	181	1jfiB	133	74
1aky	454	226	1k7vA	206	87
1am1	354	156	1ng1	464	204
1amk	416	148	1nzcA	478	120
1aonF	475	261	1pvdA	239	127
1ars	388	162	1qumA	310	133
1aru	93	49	1qupA	47	34
1ast	366	125	1rrpA	411	196
1axn	441	157	1rrpB	207	99
1b54	483	94	1vh4A	243	116
1bag	41	24	1w1uA	319	170
1bqk	63	42	1ycsA	112	60
1bto	490	184	1ycsB	57	49
1c1bA	474	301	2bif	253	96
1cg0	479	244	2mjpA	488	325
1cio	396	220	2msbA	298	142
1cvjA	310	139	3hhrA	339	173
1cxzA	446	176	6gst	361	113

Table 1: The change in sequence count for training set due to the optimization is shown.

Testset 2		Number of sequences			
PDBID	unoptimized	optimized	PDBID	unoptimized	optimized
1aa6	438	193	1fca	428	138
1aac	126	89	1ffh	465	188
1ah7	29	24	1fit	170	98
1ako	480	166	1fnc	200	124
1amj	255	135	1fsu	182	69
1apq	254	127	1fxd	24	19
1arv	89	56	1gai	79	61
1at0	82	64	1gcb	151	68
1ayl	417	179	1gpl	192	100
1bdb	482	196	1han	153	91
1bia	225	114	1htn	219	111
1bif	240	118	1hyt	323	96
1bip	64	41	1iba	130	67
1bor	26	19	1ido	280	110
1btl	449	201	1ig5	51	28
1cfb	183	96	1iyu	335	127
1chc	93	49	1krn	190	92
1chd	457	164	1lam	406	240
1csh	200	119	1lay	31	23
1ctn	145	58	1lcf	191	93
1ctt	79	51	1led	349	123
1cvl	116	67	1lgr	455	180
1def	474	180	1lml	178	74
1drw	469	225	1mla	487	177
1dxy	449	176	1mup	86	41
1e70	449	186	1nif	37	31
1ecl	473	171	1nir	34	27
1emn	1277	36	1nox	49	27
1esl	73	39	1onc	43	25
1far	81	32	1opc	422	125

Table 2: The change in sequence count for testset due to the optimization is shown.

Testset 2		Number of sequences			
PDBID	unoptimized	optimized	PDBID	unoptimized	optimized
1osa	397	204	1vii	47	40
1pbn	331	144	1vsd	190	56
1pda	458	211	1whi	493	188
1pdc	218	147	1xnb	243	115
1pii	156	84	2abk	488	203
1pkp	465	233	2ace	409	139
1poa	379	149	2af8	46	35
1poc	38	30	2asi	306	96
1pth	83	53	2cba	376	151
1put	443	124	2cmd	281	147
1qli	326	115	2dkb	391	209
1rfs	72	56	2dln	447	146
1rie	395	179	2fha	399	165
1rnl	487	262	2hft	32	25
1se4	87	44	2rn2	479	125
1snc	164	61	2sil	48	23
1sp2	164	128	2vil	286	215
1sra	63	44	3dni	110	58
1thg	285	106	3ebx	176	54
1thm	441	358	3ssi	31	26
1thx	483	271	4enl	446	179
1tmy	460	305	4rhn	474	205
1uch	165	62	5eat	409	147
1uxc	130	82	5ptp	417	139
1vhh	66	33	7rsa	341	114

Table 3: The change in sequence count for testset due to the optimization is shown.