Supplementary information

ZEAL: Protein structure alignment based on shape similarity

Filip Ljung and Ingemar André

Division of Biochemistry and Structural Biology, Department of Chemistry, Lund University,

POB 124, SE-22100 Lund, Sweden

March 23, 2021

S1. Molecular surfaces from EDT

Our algorithm to generate molecular surfaces using the EDT deviates slightly from that by Xu and Zhang[6], and we therefore outline the main steps below with the corresponding results shown in Figure S1 after each step.

1. **Step 1.** Scale and translate all atoms to fit inside a bounding box with side length L and with integer coordinates (voxels). The resolution of this grid is thus L^3 voxels. After this scaling, the vdW radii r_i and solvent probe radius r_p becomes sr_i and sr_p voxels.

2. **Step 2.** Create a solvent-accessible (SA) solid by assigning voxels a value of 1 if they are within $sr_i + sr_p$ for each atom, or 0 otherwise. While not performed in the work presented here, any cavities in the SA solid can be removed using a flood-fill operation on background voxels; the cavities are those voxels that cannot be reached by filling in the background from the sides of the box.

3. **Step 3.** Create the molecular surface (MS) by keeping voxels at the boundary (perimeter) of the SA solid. A voxel is part of the perimeter if it is non-zero and it is connected to at least one non-zero valued voxel within a 6-connected neighbourhood (i.e. voxel faces touch). The perimeter is found by comparing the SA solid and its image-eroded version using the connectivity neighbourhood. The vdW surface is obtained by setting $r_p = 0$ in step 2.

4. **Step 4.** Map voxels inside the MS by using a flood-fill operation on background pixels.

5. **Step 5.** Perform a Euclidean distance transform (EDT) to generate the Euclidean distance map (EDM) for the MS.

6. **Step 6.** Change the sign for EDM-values belonging to interior voxels mapped in step (4). This is the signed EDM (sEDM).

7. **Step 7.** Generate the (voxelized) molecular surface by extracting the isosurface from the sEDM with isovalues equal to $-sr_p$.

The vdW radii are taken from Bondi[1] and, by default, the probe radius r_p is set to 1.4 Å. The EDT method integrates naturally with the Novotni and Klein algorithm[4] for ZC moment computations as the molecular surface is mapped to a grid directly. The surface sampling resolution and shell thickness, both affecting the performance of the shape representation in ZC space, can easily be controlled by specifying the grid size L (sampling resolution) the range of isovalues in the vicinity of $-sr_p$ (shell thickness). In the work presented here, we use a 64^3 grid and a thickness of 2 voxels for the molecular surface.

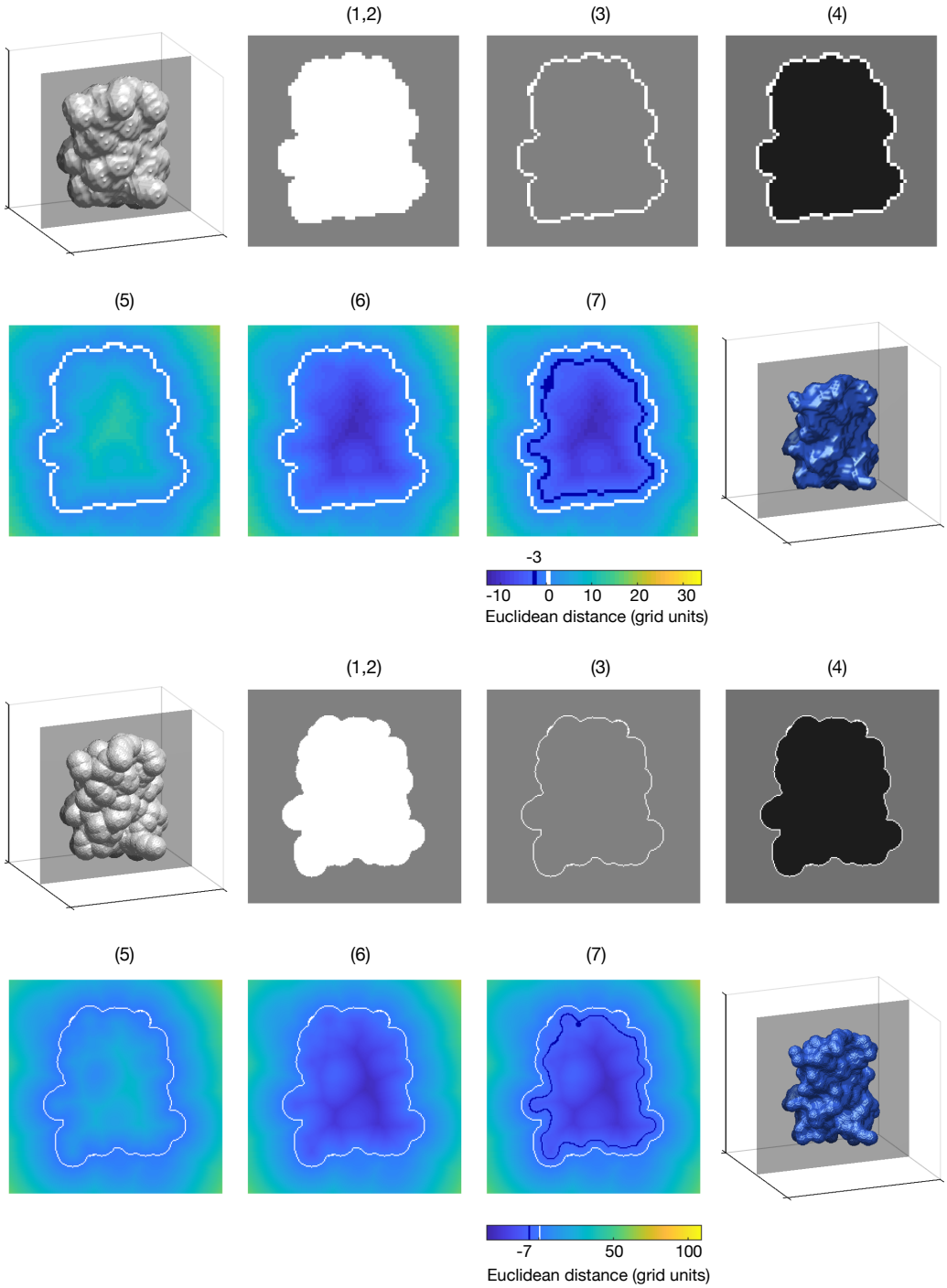

Figure S1. Steps 1-7 for computing the molecular surface using the Euclidean distance transform (EDT) on a 64^3 (above) and 200^3 (below) grid for BPTI (PDB ID code 5PTI)

.

S2. Shape alignment

Figure S2 shows the ZEAL score (A) and RSMD (B) landscape for the self-alignment of a rotated copy of 1B3T on a 2D search grid where the correct alignment is located.

We benchmark the robustness of the surrogate optimization algorithm implemented in ZEAL by performing self-alignment trials of five protein structures. Each structure is randomly rotated, over all possible Euler angles, and then aligned to an un-rotated copy using 1000 function evaluations with default settings (expansion order of 20; 64^3 grid; molecular surface representation and scaling factor $s = 0.7\mathrm{R}_{\mathrm{max}}$). Table S1 below shows statistics for the heavy-atom RMSD before and after the search for each structure. In all cases, the resulting RMSD is close to zero and we conclude that the algorithm is very likely to find the correct alignment.

Figure S3 A-D and figure S2 C show how the score correlates with RMSD for the five structures when one copy is systematically rotated over a dense grid of all possible Euler angles. For high scores, typically above 0.8, there is a strong linear correlation between the ZEAL score and RMSD.

Table S1. Self-alignment benchmark

ID+chain	RMSD before search (Å)	RMSD after search (Å)
1B3TA	23.2 ± 9.71	0.024 ± 0.025
3A8GA	27.9 ± 5.41	0.058 ± 0.123
5MOKA	24.8 ± 8.23	0.030 ± 0.025
2B3JA	14.9 ± 2.92	0.017 ± 0.013
4G9SB	20.7 ± 3.24	0.036 ± 0.032

Mean (± standard deviation) heavy-atom RMSD for the self-alignment of five structures (PDB ID code + chain ID code) before and after ZEAL-alignment. Statistics refer to 20 random rotations and the search was stopped after 1000 function evaluations using default settings.

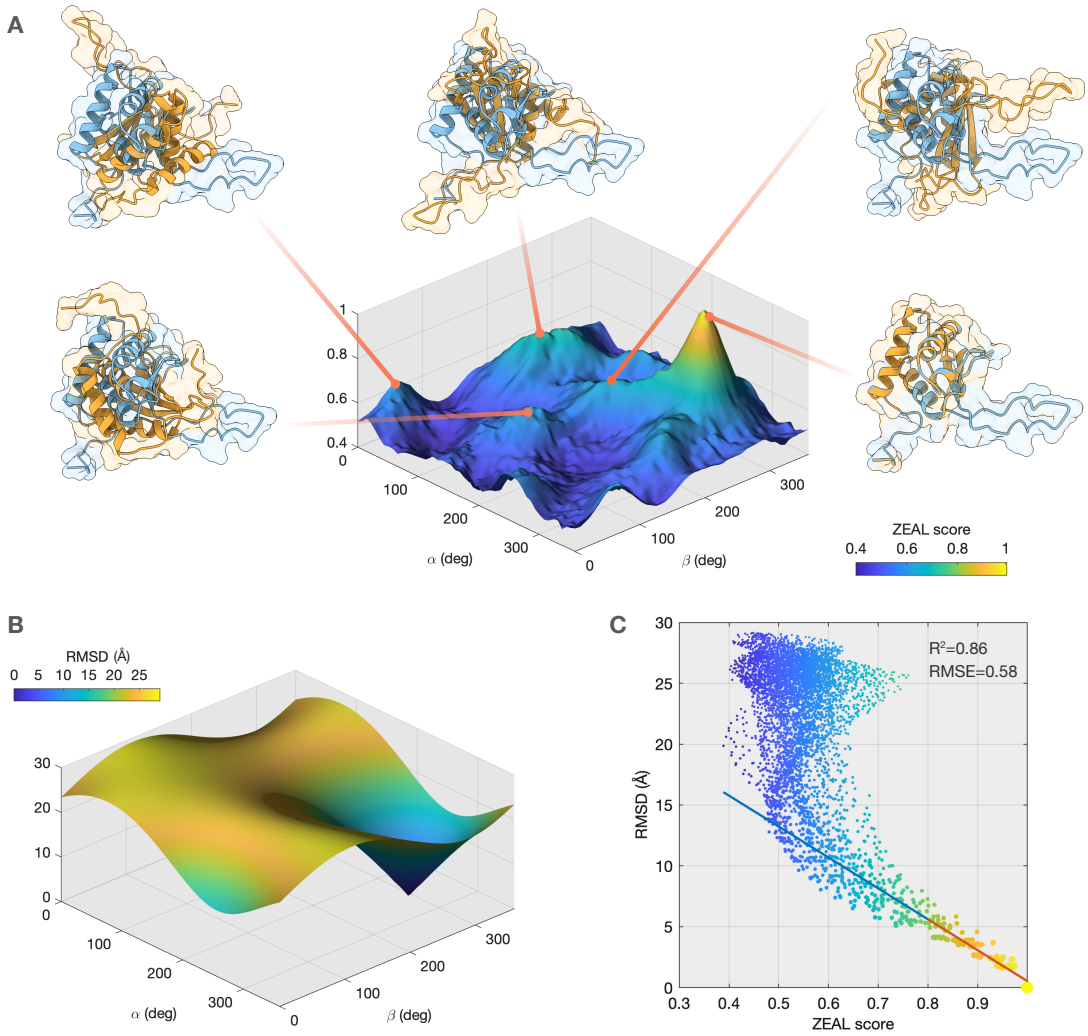

Figure S2. Self-alignment of EBNA1 nuclear protein (PDB ID code 1B3T) on a 2D search grid over Euler angles α and β (zyz convention). (A) Cross section through the 3D correlation-function landscape (ZEAL score) and (B) the heavy-atom RMSD. (C) The correlation between ZEAL score and RMSD with a linear model fitted for ZEAL scores $>$ 0.8 (red line). The adjusted R square and root mean square error (RMSE) are given in the figure.

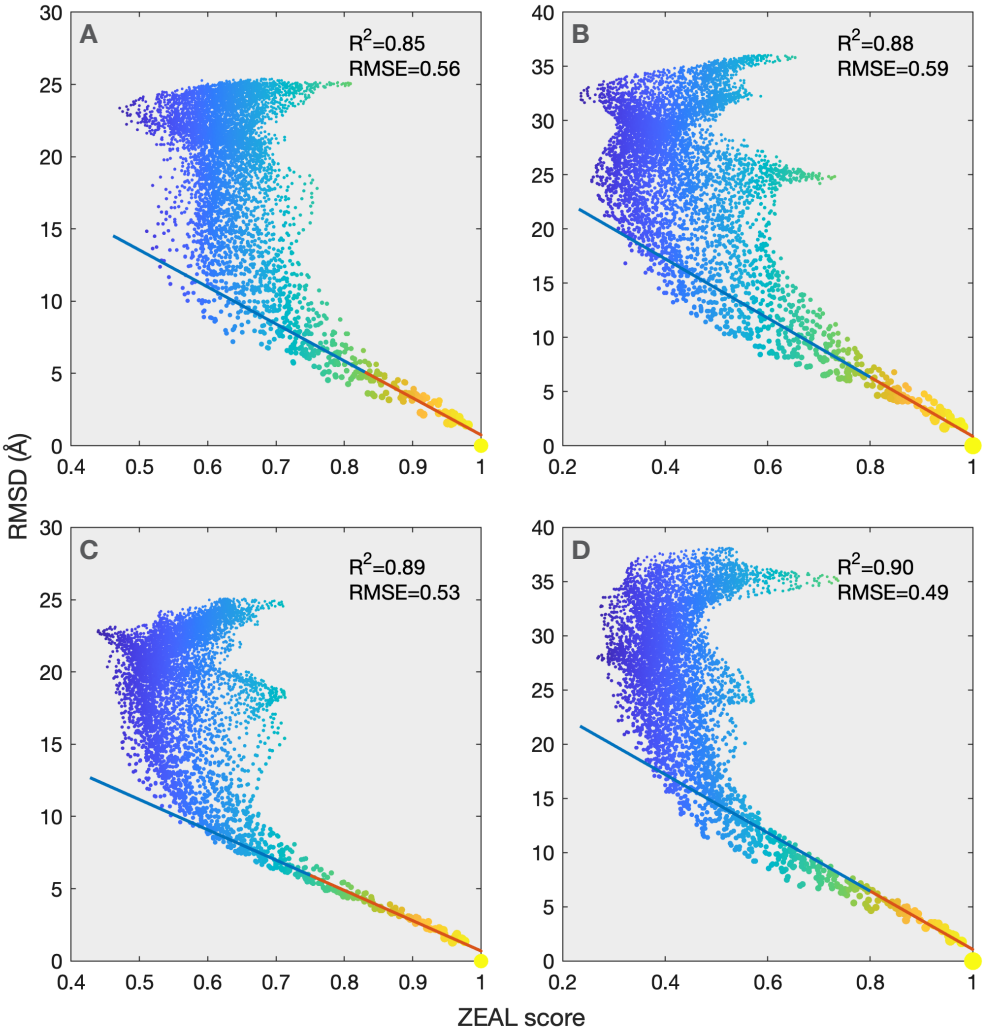

Figure S3. The correlation between ZEAL score and RMSD for the self-alignment of (A) 3A8GA, (B) 5MOKA (C) 2B3JA and (D) 4G9SB. The size of data points are scaled by the inverse RMSD and coloured by score. The red part of the line represent data that was used to create a linear fit, with the adjusted R^2 and root mean square error (RMSE) presented in each figure.

S3. Benchmark analysis

Table S2 lists the settings used for molecular matching in HEX version 8.0.0. Table S3 and S4 lists the two sets of structures (100 pairs in each) used for the benchmark analysis. All pairs have a Euclidean ZC shape descriptor distance $d_E < 0.025$, and a chain-length (length) and radius of gyration (Rg) difference < 10 %. Pairs in the high TM-score set (table S3) have TM-scores > 0.9, and pairs in the low TM-score set have TM-scores < 0.3 (table S4).

Table S2. Settings for molecular matching in HEX version 8.0.0.

Correlation Type	Shape		
Compute Device	CPU		
FFT Mode	3D		
Sampling Method	Receptor Box		
Grid Dimension	0.6	**Solutions**	10
Receptor Range	180	**Step Size**	20
Ligand Range	180	**Step Size**	10
Twist Range	360	**Step Size**	5.5
Distance Range	6	**Box Size**	6
Scan Step	0.8	**SubSteps**	0
Steric Scan	3		
Search Order	10		

Table S3: Protein structures in the high TM-score benchmark set

pair No	ID+chain A	ID+chain B	length A	length B	Rg A (Å)	Rg B (Å)	aligned length	seq.id (%)	RMSD (Å)	TMs A	TMs B	d_E
1	3UNVA	2QVEA	513	525	27.4	27.6	507	37.7	1.50	0.96	0.94	0.021
2	3HNYM	1CZTA	156	160	14.6	15.3	156	42.3	0.86	0.98	0.95	0.014
3	5NZBA	4ESPA	132	130	13.8	13.6	129	72.9	1.77	0.91	0.93	0.021
4	3H7RA	1ZGDA	307	308	19.0	18.7	296	40.9	1.51	0.93	0.92	0.024
5	3HGJA	1Z41A	348	327	19.5	19.4	326	50.9	1.09	0.92	0.97	0.021
6	4DI0A	1J30A	139	141	18.2	18.2	138	60.9	1.03	0.96	0.94	0.015
7	4M8SA	1CYDA	247	242	17.7	17.5	239	30.5	1.80	0.90	0.92	0.023
8	4K7XA	4JCIA	314	312	19.3	20.0	312	65.4	1.95	0.92	0.93	0.017
9	6NIBA	3H7CX	345	362	18.8	19.3	339	74.6	0.73	0.97	0.93	0.024
10	4G9SB	4DY5A	111	111	14.0	13.8	111	82.9	1.03	0.95	0.95	0.021
11	1CG5A	2W72A	141	141	14.9	14.8	140	42.1	1.24	0.93	0.93	0.023
12	4Q7ZA	6QFEA	233	227	16.7	16.6	224	33.0	1.25	0.92	0.95	0.024
13	4MVAA	3TA6A	255	255	17.5	17.6	250	41.6	1.24	0.94	0.94	0.021
14	5G6RA	4OQZA	288	283	23.6	23.2	282	28.4	1.26	0.94	0.96	0.024
15	1KGCD	6BJ8D	201	201	21.8	21.4	197	64.0	1.75	0.91	0.91	0.022
16	2GJDA	6GV3A	155	156	16.1	16.0	155	58.7	0.77	0.98	0.97	0.018
17	3GFVA	5AD1A	282	290	19.9	19.7	274	42.3	1.29	0.94	0.91	0.023
18	2EA3A	2PFEB	183	186	14.4	14.6	182	57.7	0.77	0.98	0.96	0.016
19	6JK4A	2PY2A	126	127	13.5	13.4	125	83.2	0.45	0.98	0.97	0.024
20	1MNGA	6EX5A	198	198	16.7	16.6	192	49.5	1.05	0.93	0.93	0.021
21	3BRSA	3KSMA	271	272	19.4	19.5	264	24.2	1.81	0.92	0.91	0.024
22	1O91A	6U66A	131	136	14.6	15.3	131	42.0	1.17	0.94	0.91	0.019
23	3V3WA	3QKEA	397	397	20.8	20.8	397	72.5	0.57	1.00	1.00	0.019
24	3F6DA	5F0GA	218	207	17.2	17.0	204	54.9	1.09	0.91	0.95	0.024
25	3RGKA	5YCEA	149	151	15.1	15.0	149	85.9	0.63	0.98	0.97	0.022
26	1NFJA	2BKYA	87	89	14.2	14.9	87	60.9	1.30	0.93	0.91	0.021
27	3WPCA	3WPFA	747	740	31.4	31.6	736	72.0	1.42	0.97	0.98	0.013
28	1PBYA	1JMXA	489	493	25.2	25.2	480	40.2	1.72	0.94	0.94	0.023
29	4XV0A	1I1WA	301	302	18.4	18.2	301	64.5	0.62	0.99	0.99	0.017
30	4M9DA	3HIDA	417	424	22.0	22.4	409	46.7	1.70	0.94	0.92	0.021
31	3GZBF	3LZAD	130	133	14.9	15.1	129	72.1	0.65	0.97	0.95	0.021
32	3H81A	2PBPA	256	255	19.4	19.7	255	45.5	1.11	0.96	0.97	0.020
33	3AV3A	3P9XA	184	195	16.3	16.7	183	60.7	0.99	0.96	0.91	0.025
34	5XVJA	5Y53A	134	132	15.6	15.4	130	76.2	1.31	0.92	0.93	0.016
35	3HH8A	5JPDA	280	275	18.9	19.1	271	53.1	1.14	0.94	0.96	0.025
36	4AOHA	1AGIA	122	125	14.5	14.7	122	65.6	1.34	0.96	0.94	0.023
37	3OP4A	2PNFA	235	248	17.5	17.8	235	51.1	1.47	0.95	0.90	0.025
38	4FCUA	1VICA	253	255	18.9	19.0	251	42.2	1.67	0.94	0.93	0.016
39	3A8GA	4OB0A	195	202	19.3	19.5	193	42.5	1.80	0.93	0.90	0.024
40	4BVQA	5HWEA	364	370	19.5	19.7	361	43.8	1.79	0.94	0.92	0.011

pair No	ID+chain A	ID+chain B	length A	length B	Rg A (Å)	Rg B (Å)	aligned length	seq.id (%)	RMSD (Å)	TMs A	TMs B	d_E
41	1SPGB	1WMUB	147	146	15.2	15.1	146	51.4	1.01	0.95	0.96	0.024
42	5WS7A	5OTNA	156	156	15.3	15.1	156	69.9	1.40	0.94	0.94	0.017
43	1YARH	5LE5K	203	201	16.6	16.6	199	28.6	2.04	0.91	0.92	0.022
44	4O6RA	5EKCA	489	484	24.2	24.6	480	33.3	1.31	0.96	0.97	0.020
45	3NGJA	1UB3A	218	211	16.4	16.2	211	43.6	0.89	0.94	0.98	0.021
46	3FW9A	5D79A	495	498	22.4	22.5	487	42.5	1.82	0.94	0.94	0.014
47	4YIOA	1XUQA	202	200	16.6	16.5	198	57.6	0.95	0.95	0.96	0.019
48	3W9SA	1ZH2A	116	120	13.0	13.1	115	40.9	0.98	0.94	0.91	0.023
49	4F40A	4MHBA	280	274	18.1	18.3	260	45.8	1.00	0.91	0.93	0.022
50	4IO1A	6EEPA	218	218	17.4	17.2	217	53.0	1.07	0.96	0.96	0.024
51	4KX4A	3IR4A	217	211	17.3	17.5	211	84.8	0.61	0.96	0.99	0.016
52	5O0DA	3NBKA	157	158	15.5	15.5	157	78.3	1.39	0.95	0.94	0.024
53	1ZUWA	2DWUA	261	265	17.9	18.0	259	52.9	1.14	0.96	0.95	0.020
54	2CYGA	1AQ0A	312	306	18.9	18.6	306	55.2	0.89	0.96	0.98	0.016
55	1CS1A	3ACZA	383	386	21.9	21.9	377	35.8	1.55	0.95	0.94	0.014
56	4KC7A	3LV4A	449	431	21.8	21.5	424	56.4	0.97	0.93	0.97	0.024
57	3JU8A	4KNAA	480	487	24.3	24.7	480	60.0	0.71	0.99	0.98	0.024
58	2V3ZA	5WZEA	439	442	25.3	25.2	438	49.3	1.32	0.98	0.97	0.018
59	5T5XA	4MLPC	470	492	24.1	24.2	470	79.1	1.31	0.98	0.94	0.025
60	1JD1A	1X25A	126	126	13.7	14.0	125	40.8	1.12	0.94	0.94	0.023
61	4HM8B	2BMOB	192	194	18.1	18.4	192	77.6	0.80	0.98	0.97	0.016
62	1ARBA	4NSVA	263	264	16.5	16.6	263	77.9	0.44	0.99	0.99	0.016
63	4EIEA	1GDVA	82	85	11.7	11.7	82	39.0	0.74	0.96	0.92	0.024
64	2BFDB	1UMDB	332	323	19.8	19.6	322	46.3	0.89	0.95	0.98	0.022
65	3E7PA	3KKZA	253	251	18.3	18.2	245	71.8	1.15	0.94	0.95	0.017
66	2C7SA	3I36A	289	286	19.0	18.5	281	42.0	1.63	0.93	0.94	0.021
67	4YETA	3H1SA	198	190	16.1	16.3	190	47.9	1.23	0.92	0.96	0.022
68	2ZBXA	3ABAA	399	397	21.1	21.0	392	43.6	2.10	0.92	0.93	0.018
69	4WEOA	5T5QA	238	237	17.0	17.3	233	34.3	1.90	0.91	0.91	0.024
70	1WU4A	5YXTA	374	379	19.3	19.2	372	75.8	0.69	0.99	0.97	0.025
71	5V0ZC	2P2OA	189	184	17.2	16.9	184	42.4	1.08	0.94	0.96	0.024
72	4AK9A	3B9QA	302	298	21.2	21.6	294	69.7	1.30	0.94	0.95	0.024
73	5AOVA	2DBQA	334	327	22.4	22.5	327	84.7	0.48	0.97	0.99	0.024
74	1H5QA	1GEEA	260	261	18.2	18.3	251	29.9	1.52	0.92	0.91	0.020
75	5VN5B	6DXSA	332	342	19.1	19.4	324	35.5	1.66	0.93	0.90	0.019
76	3UWCA	3NYTA	362	359	21.6	20.5	350	32.6	1.75	0.91	0.92	0.024
77	6DXBA	3OITA	388	367	21.0	20.6	366	54.1	1.20	0.92	0.97	0.018
78	4DNXA	1JHDA	396	396	21.9	21.8	396	80.8	0.78	0.99	0.99	0.018
79	5XAOA	4RSLA	431	436	21.7	21.8	429	72.3	1.06	0.98	0.97	0.024
80	1DM5A	2ZHJA	315	315	21.5	21.6	314	47.5	1.25	0.97	0.97	0.019
81	2OSAA	5IRCB	196	195	16.6	16.6	191	36.6	1.77	0.90	0.91	0.018

pair No	ID+chain A	ID+chain B	length A	length B	Rg A (Å)	Rg B (Å)	aligned length	seq.id (%)	RMSD (Å)	TMs A	TMs B	d_E
82	2PFYA	2PFZA	295	301	18.9	18.9	294	42.5	0.96	0.97	0.96	0.021
83	1XE0A	4N8MA	108	108	14.4	14.3	106	80.2	1.30	0.94	0.94	0.024
84	4E98A	2NUHA	110	104	15.8	14.5	104	32.7	0.90	0.90	0.95	0.024
85	2INCB	3GE3B	322	304	21.6	22.4	302	60.6	0.88	0.92	0.98	0.019
86	1MI3A	5Z6UA	319	318	19.1	19.2	316	74.1	0.93	0.98	0.98	0.023
87	5JGYA	2BGSA	306	308	18.9	18.9	306	88.2	0.49	0.99	0.99	0.023
88	2D5KA	1JIGA	148	146	15.7	15.6	146	50.7	0.54	0.97	0.99	0.016
89	2IWXA	6CJIA	214	214	17.1	17.0	213	87.3	0.55	0.99	0.99	0.012
90	1RURH	1KCVH	218	217	21.4	21.1	214	55.1	1.66	0.91	0.91	0.023
91	5EWOA	3QSQA	213	216	16.8	17.1	213	53.1	0.82	0.98	0.97	0.021
92	4UPIA	4UPLA	547	555	25.6	27.3	530	54.7	1.60	0.94	0.93	0.022
93	2C0RB	6CZXA	361	361	20.9	21.4	360	47.8	1.28	0.97	0.97	0.022
94	1MTZA	3WMRA	290	296	18.2	18.2	290	27.9	1.59	0.95	0.93	0.015
95	5J8NA	2JC4A	258	256	17.3	17.0	250	27.6	1.76	0.90	0.91	0.022
96	3WC3A	2XFGA	435	446	20.4	20.8	426	43.9	1.57	0.95	0.93	0.016
97	4JCPA	2CMTA	170	164	14.4	14.4	163	70.6	0.68	0.94	0.98	0.025
98	5ICEA	6I73A	351	353	23.3	22.8	339	36.0	1.80	0.91	0.91	0.020
99	2OX4A	2GL5A	393	401	20.6	20.7	392	51.3	1.19	0.98	0.96	0.014
100	4UO0A	4FNKA	325	318	31.5	28.9	318	83.6	0.75	0.97	0.99	0.012

PDB ID code and chain ID code (ID+chain) for structural homologs A and B in the high TM-score benchmark set. The structures have TM-scores (TMs) > 0.9, ZC shape descriptor (ZCD) Euclidean distance $d_E < 0.025$, chain-length (length) and radius of gyration (Rg) difference < 10 %.

Table S4: Protein structures in the low TM-score benchmark set

pair No	ID+chain A	ID+chain B	length A	length B	Rg A (Å)	Rg B (Å)	aligned length	seq. id. (%)	RMSD (Å)	TMs A	TMs B	d_E
1	1R8NA	3IHWA	185	167	15.6	14.9	96	11.5	5.84	0.27	0.29	0.023
2	2ZYZB	3VP5A	183	185	17.6	18.5	100	7.0	6.21	0.27	0.27	0.022
3	3BX4A	5N07A	132	137	16.8	17.5	52	3.8	4.61	0.24	0.23	0.022
4	2ZPMA	1UCRA	79	74	12.1	11.8	43	7.0	4.12	0.28	0.29	0.024
5	5MOKA	2HO1A	208	219	20.3	21.1	91	1.1	6.01	0.25	0.24	0.025
6	3ZS3A	2VYOA	222	206	16.8	16.6	99	5.1	5.67	0.26	0.27	0.021
7	1YWFA	3HHIA	241	258	16.9	17.7	105	8.6	5.70	0.26	0.25	0.024
8	1MGQA	5D4SA	74	70	12.5	11.8	33	3.0	3.05	0.28	0.29	0.021
9	5URPA	3C37A	201	217	17.2	17.4	90	8.9	5.75	0.26	0.25	0.025
10	4BF5A	3G5BA	396	383	24.2	24.1	158	8.2	7.43	0.22	0.23	0.024
11	5HXLA	5MKWB	123	117	15.9	15.6	66	6.1	4.91	0.28	0.29	0.023
12	1GV9A	3P8AA	223	245	16.1	17.4	122	5.7	6.29	0.29	0.28	0.024
13	5XMZA	3IU5A	122	114	14.4	14.6	53	11.3	5.04	0.23	0.24	0.023
14	2PEZA	2O99A	176	170	15.7	15.6	93	10.8	5.64	0.28	0.29	0.024
15	3FANA	2GFNA	192	190	17.1	18.1	108	4.6	6.02	0.30	0.30	0.023
16	5O9WA	1VKNA	356	331	20.6	19.9	155	7.7	6.72	0.26	0.27	0.020
17	3CNUA	4HPLA	110	113	14.7	15.3	62	8.1	5.42	0.27	0.26	0.024
18	1WHIA	6BEVA	122	113	13.9	13.4	58	8.6	5.40	0.23	0.24	0.025
19	3E0EA	6J6VA	89	94	12.5	12.6	42	7.1	4.64	0.24	0.23	0.021
20	3C3JA	4QTCA	361	330	20.4	20.3	161	6.8	7.36	0.24	0.26	0.023
21	3VPGA	5X5OA	310	291	19.8	19.6	145	11.0	6.15	0.29	0.30	0.024
22	3P85A	5BY8A	224	221	16.9	18.1	97	7.2	6.20	0.24	0.24	0.019
23	4YS6A	1MNNA	313	290	20.2	20.3	135	3.7	6.47	0.25	0.27	0.025
24	1AZ5A	4A8XA	95	88	13.6	13.3	56	8.9	4.92	0.28	0.29	0.024
25	2DURA	6QWOA	251	252	17.7	17.4	122	5.7	6.61	0.26	0.26	0.024
26	5F86A	5N7QA	365	337	21.5	20.0	159	5.7	6.54	0.27	0.28	0.021
27	6CW3G	2P1MA	92	90	18.1	19.2	49	2.0	5.29	0.24	0.24	0.023
28	3JS4A	5M6QA	207	224	16.8	17.8	119	5.0	6.49	0.30	0.28	0.023
29	4I0WA	3L7HA	91	86	13.1	13.3	52	7.7	4.70	0.27	0.27	0.025
30	3PXXA	3HISA	282	257	18.0	18.1	129	6.2	6.74	0.25	0.27	0.025
31	1EVLA	3FK4A	401	388	24.5	23.3	174	5.7	7.05	0.25	0.26	0.025
32	4XTBA	2VPKA	109	115	14.8	14.2	59	6.8	5.34	0.24	0.23	0.024
33	6EKKA	2YXXA	375	375	21.5	22.5	162	7.4	7.22	0.24	0.24	0.023
34	1VCHA	3EFYA	168	183	16.1	16.7	81	6.2	5.29	0.29	0.27	0.025
35	2RJ2A	6R1HA	185	184	17.3	16.9	78	6.4	6.03	0.22	0.22	0.023
36	3FUCA	6AE9A	274	251	18.3	18.0	132	9.1	6.41	0.27	0.29	0.022
37	1T5IA	1VKBA	159	147	15.1	14.5	66	10.6	5.13	0.25	0.26	0.023
38	3G3TA	1WF3A	274	296	20.3	19.9	119	4.2	5.75	0.28	0.26	0.024
39	2PIIA	1BUOA	112	121	17.1	16.3	49	4.1	3.97	0.26	0.25	0.024
40	5F6QA	2ONFA	140	137	17.4	17.8	69	7.2	5.15	0.28	0.28	0.024

pair No	ID+chain A	ID+chain B	length A	length B	Rg A (Å)	Rg B (Å)	aligned length	seq. id. (%)	RMSD (Å)	TMs A	TMs B	d_E
41	6U54B	5GJYB	95	100	14.9	14.2	45	4.4	4.07	0.26	0.25	0.025
42	5EU0B	2Z7FI	55	50	11.5	11.5	27	0.0	3.87	0.23	0.24	0.025
43	3A8GA	4GUNA	195	205	19.3	17.9	99	5.1	5.83	0.28	0.27	0.024
44	4AUCA	3QJ3A	323	319	19.9	20.0	141	6.4	6.57	0.26	0.26	0.025
45	1LLNA	1FX4A	248	231	18.3	17.5	106	8.5	6.40	0.24	0.25	0.021
46	1DL5A	3PVDA	317	299	21.0	20.4	145	4.1	7.11	0.25	0.26	0.019
47	4ZJHA	1U02A	208	222	19.0	18.2	94	4.3	6.02	0.24	0.23	0.022
48	2SGAA	2FCRA	181	173	14.3	14.7	86	7.0	5.08	0.29	0.30	0.024
49	5NCRB	1XHDA	174	169	15.6	15.6	76	3.9	5.01	0.27	0.28	0.024
50	3KEPA	6OE6A	140	127	16.4	15.2	74	5.4	5.35	0.27	0.28	0.025
51	5K8CA	6DGMB	358	382	20.4	20.6	155	7.1	6.96	0.25	0.24	0.024
52	2H7ZA	6BN0A	75	79	13.1	13.1	42	14.3	4.15	0.28	0.28	0.023
53	3BIOA	3BL9A	280	286	21.1	21.6	114	7.0	6.11	0.24	0.24	0.024
54	5M2YA	5D3XB	130	141	14.1	15.3	60	5.0	4.52	0.28	0.26	0.024
55	2P7SA	5AYQB	110	121	13.7	14.7	57	5.3	4.70	0.29	0.27	0.018
56	4R37A	1MRZA	255	266	20.2	21.3	116	9.5	6.06	0.27	0.26	0.024
57	5NJOA	4NWYA	121	128	14.4	15.4	62	6.5	5.38	0.25	0.24	0.021
58	4IYKA	5E5LA	207	190	18.8	18.0	107	9.3	6.09	0.28	0.29	0.023
59	2F62A	3KBYA	151	138	17.0	16.2	75	5.3	5.38	0.27	0.28	0.022
60	3EQCA	3D4PA	312	307	19.8	19.5	142	6.3	6.22	0.28	0.29	0.023
61	3BQ9A	4CQBA	446	402	21.8	21.2	212	6.6	7.31	0.28	0.30	0.024
62	6EVNA	2V6UA	95	103	13.1	13.8	47	2.1	4.16	0.28	0.27	0.021
63	4OVSA	4FE3A	300	291	19.7	18.9	143	7.7	6.50	0.28	0.28	0.022
64	3O5NA	4MNOA	100	95	13.7	13.0	58	6.9	4.95	0.28	0.29	0.021
65	6HIUB	5GVDA	143	153	15.5	15.9	72	6.9	5.31	0.26	0.25	0.021
66	2WEIA	1LC0A	278	290	19.9	19.9	147	10.9	6.71	0.29	0.28	0.022
67	4N5MA	5Z37A	246	246	17.7	17.4	117	0.9	6.04	0.27	0.27	0.025
68	5L8LA	1E7WA	263	267	19.2	18.5	121	5.0	5.51	0.30	0.29	0.023
69	4LUNU	2E8VA	311	292	23.6	22.0	138	5.1	6.37	0.26	0.27	0.023
70	1PWAA	1MVOA	123	121	13.4	13.2	72	11.1	5.38	0.29	0.29	0.021
71	4Y2LA	1TFEA	147	142	17.7	18.0	66	7.6	5.12	0.25	0.26	0.025
72	3ZBVA	2OR7A	118	107	14.2	13.6	50	10.0	4.26	0.26	0.28	0.023
73	4LHSA	1IG0A	329	317	20.3	20.7	143	7.0	6.62	0.26	0.27	0.025
74	6MX6C	3K7IB	143	157	14.8	14.9	82	3.7	5.84	0.27	0.26	0.024
75	2GJ3A	2QHQA	119	111	14.5	14.0	66	10.6	5.74	0.25	0.26	0.023
76	5WCJA	4DOOA	222	203	17.4	16.9	96	9.4	5.88	0.25	0.26	0.025
77	3MVCA	3K9WA	157	163	15.8	15.8	79	13.9	5.19	0.27	0.26	0.024
78	2V2PA	1ZK5A	170	172	18.1	17.0	83	3.6	5.80	0.26	0.26	0.023
79	2RDQA	1RO7A	265	255	17.6	17.8	120	3.3	6.44	0.25	0.26	0.024
80	1U6RA	3AUPA	380	377	21.9	21.4	180	10.6	7.01	0.28	0.28	0.025
81	5A2DA	2P3YA	493	447	24.7	24.1	196	3.1	7.48	0.24	0.25	0.023

pair No	ID+chain A	ID+chain B	length A	length B	Rg A (Å)	Rg B (Å)	aligned length	seq. id. (%)	RMSD (Å)	TMs A	TMs B	d_E
82	6RLTB	2UY1A	456	438	28.7	31.1	188	4.3	6.39	0.27	0.28	0.023
83	3VZ9B	6QZLA	103	103	15.2	14.3	56	7.1	4.93	0.29	0.29	0.024
84	1DUVG	6T2WA	333	304	19.3	19.4	140	3.6	7.00	0.23	0.25	0.025
85	1MBMA	3GMFA	198	196	17.2	17.0	100	6.0	6.00	0.28	0.28	0.024
86	5M9NA	1O63A	202	200	18.7	19.4	103	9.7	6.13	0.27	0.27	0.023
87	4JEMA	5EWYA	157	165	14.7	15.7	80	2.5	5.44	0.28	0.27	0.024
88	4NYHA	5JBLA	182	165	15.9	15.9	79	10.1	5.49	0.25	0.27	0.025
89	5K6DA	1DSZA	77	75	11.9	11.7	43	11.6	4.15	0.29	0.29	0.024
90	4WF5A	5LT5A	188	204	16.8	16.8	86	7.0	4.79	0.30	0.28	0.024
91	1X8BA	3FUCA	259	274	18.9	18.3	116	6.9	6.08	0.27	0.26	0.022
92	1GWYA	1D4OA	175	177	14.9	15.1	89	2.2	5.60	0.29	0.28	0.020
93	3BQ9A	2QN0A	446	429	21.8	21.9	201	5.0	7.02	0.28	0.28	0.021
94	2SAKA	2O16A	121	132	15.8	16.3	63	7.9	4.81	0.27	0.26	0.020
95	3FD5A	2X3EA	339	325	19.6	18.9	175	7.4	7.14	0.29	0.30	0.023
96	4ATMA	3MVSA	221	210	31.9	29.1	87	5.7	5.29	0.24	0.25	0.017
97	2WSBA	1Y4JA	254	266	18.0	17.8	101	5.9	6.28	0.23	0.22	0.025
98	2XLGA	6AVXA	229	208	16.8	15.7	99	7.1	5.31	0.28	0.30	0.024
99	4HJIA	6NLFE	154	156	16.6	17.1	72	5.6	5.10	0.27	0.27	0.015
100	3PVIA	1SS4A	156	143	18.1	17.3	71	4.2	4.93	0.26	0.28	0.018

PDB ID code and chain ID code (ID+chain) for shape homologs A and B in the low TM-score benchmark set. The structures have TM-scores (TMs) < 0.3, ZC shape descriptor (ZCD) Euclidean distance $d_E < 0.025$, chain-length (length) and radius of gyration (Rg) difference < 10 %.

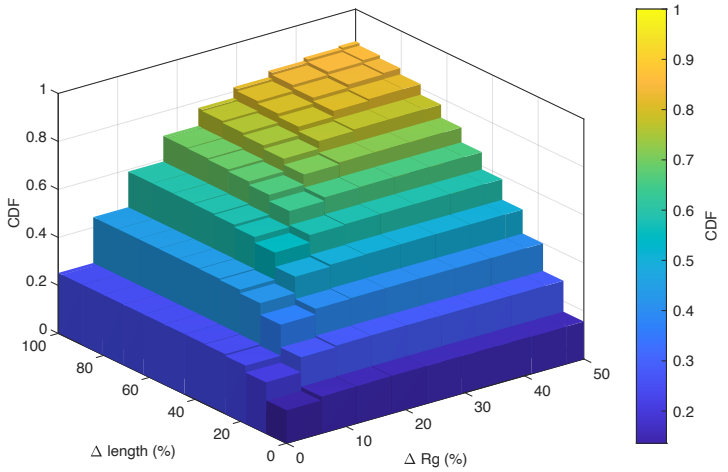

Figure S4. 2D cumulative distribution function (CDF) for the percent difference in residue length and radius of gyration (R_g) for shape matches in the S2 set.

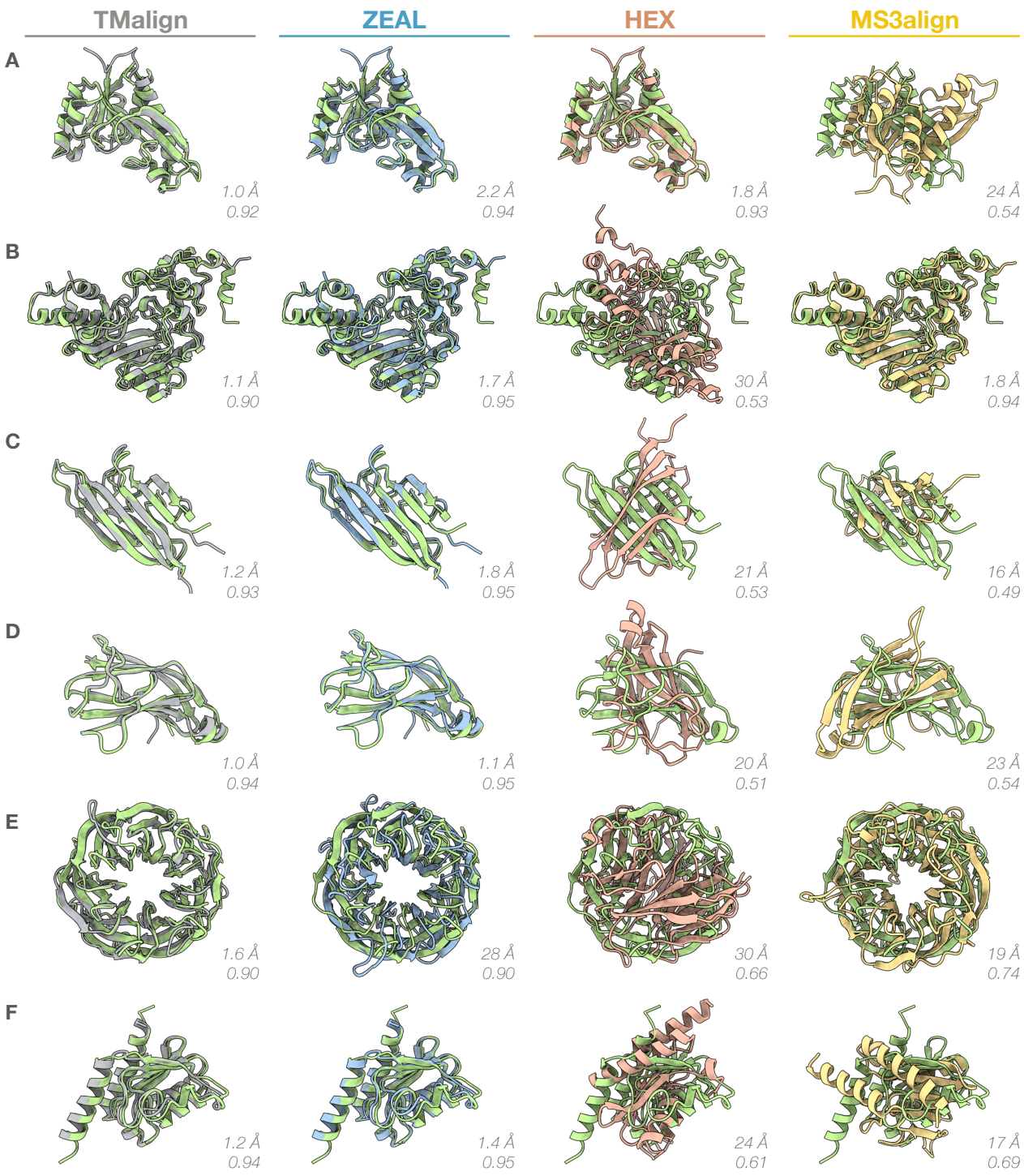

Figure S5. Six alignments computed by TM-align, ZEAL, HEX and MS3Align from the high TM-score benchmark set. The C_α $RMSD_{TM}$ and the ZEAL score for each alignment is also shown. A) 3AV3-A (green) & 3P9X-A B) 1U6RA (green) & 3JPZA C) 1O91A (green) & 6U66A D) 4G9S-B (green) & 4DY5-A E) 3SCY-A (green) & 6IGB-A F) 1WWR-A (green) & 2B3J-A.

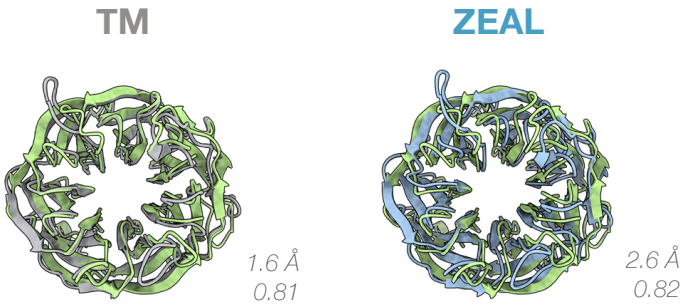

Figure S6. Alignment of 3SCY-A (green) and 6IGB-A using TM-align and ZEAL with ZC moments up to order 30. The C_α $RMSD_{TM}$ and the ZEAL score for each alignment is also shown.

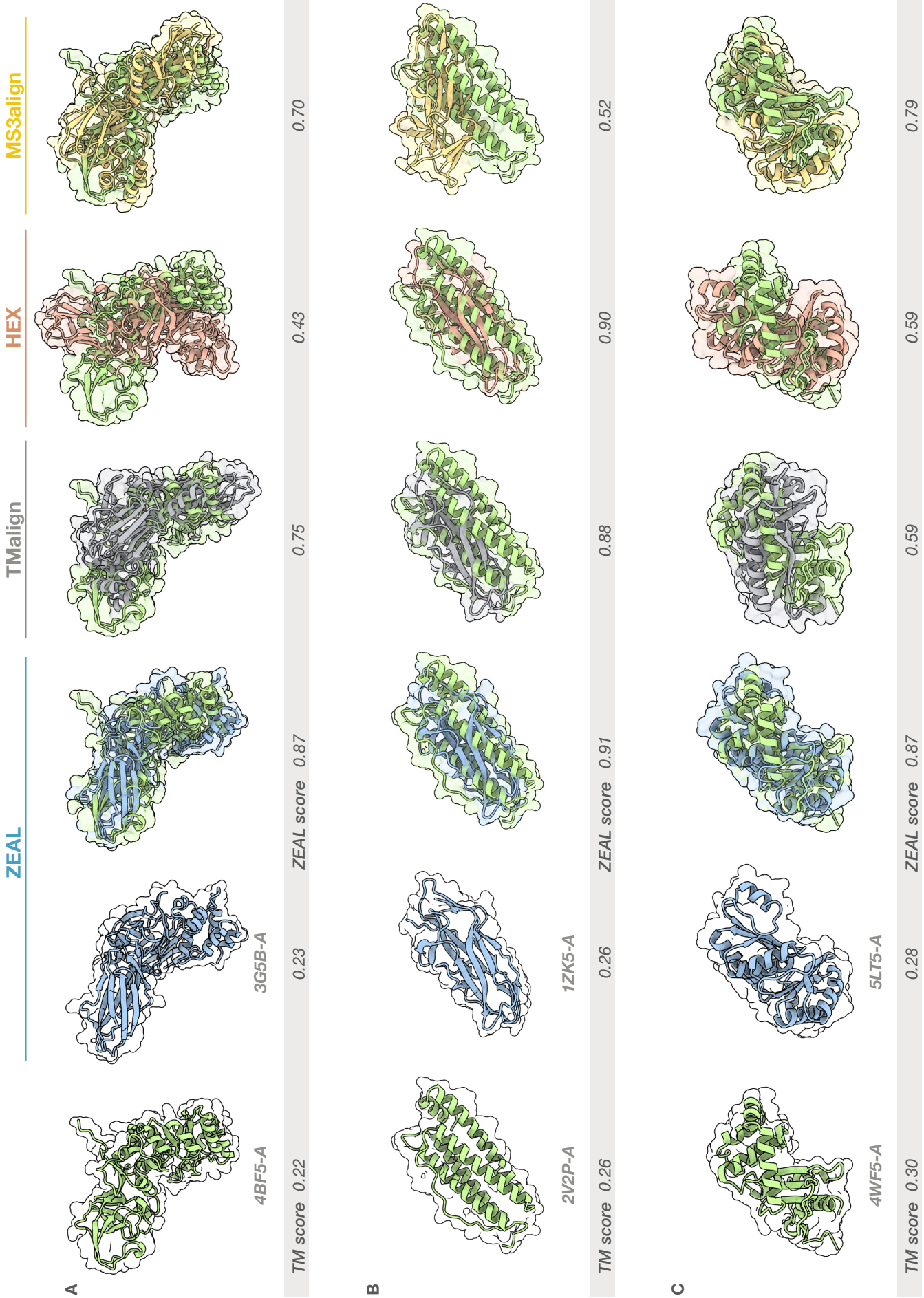

Figure S7 A-C. Superpositions from the low TM-score benchmark set.

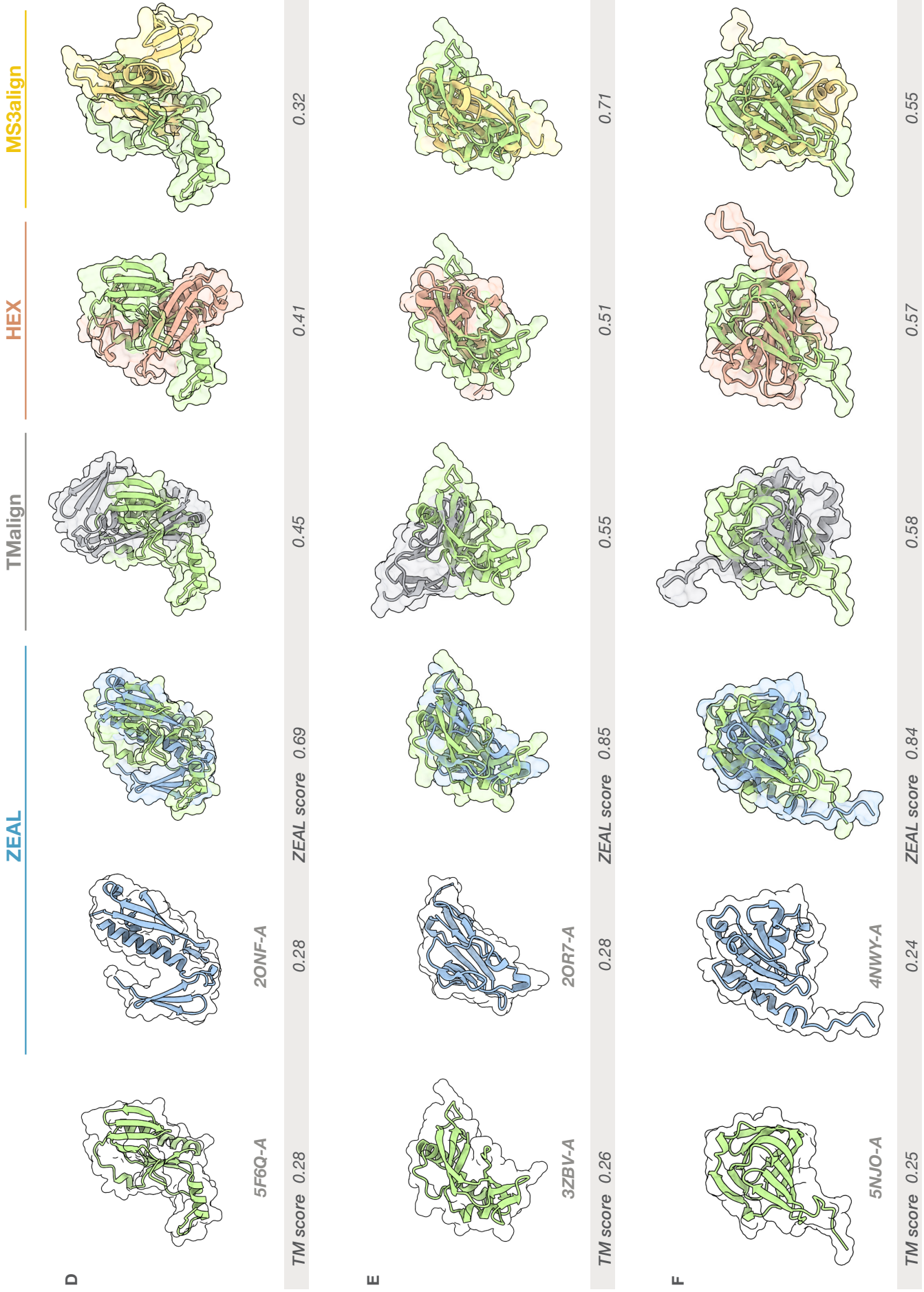

Figure S7 D-F. Superpositions from the low TM-score benchmark set.

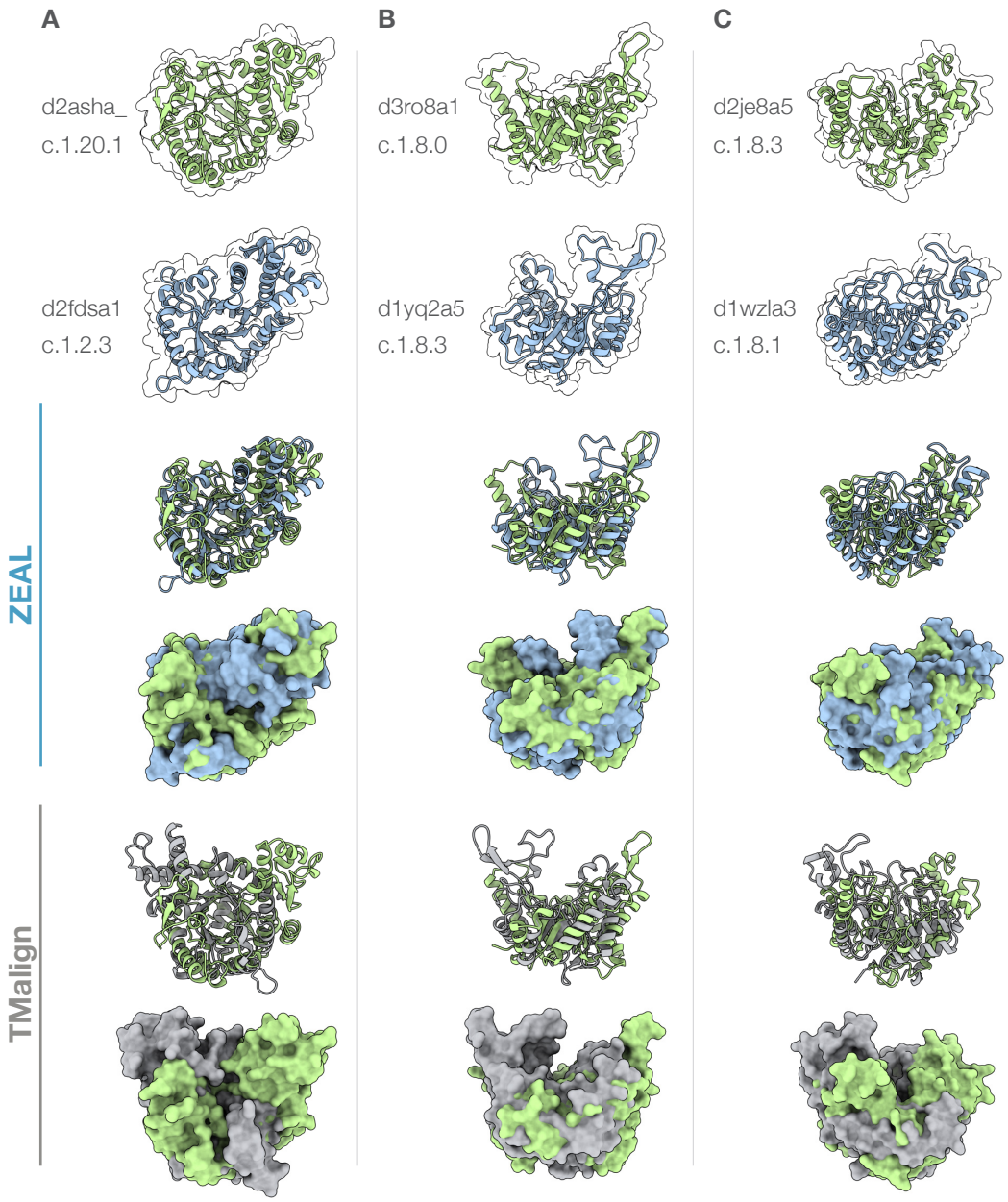

Figure S8. Examples of TIM-barrel proteins from the SCOPe (2.07-2021-01-20) database [2] where shape-alignments can be used to compare the relative placement of structural elements on the outside of the central beta-barrel. The labels left of each protein denotes the stable domain identifier (sid) and the SCOPe concise classification string (sccs).

Table S5. Benchmark statistics for the high TM-score data set

	TMalign	**ZEAL**	**HEX**	**MS3Align**[b]
mean (std) RMSD[a](Å)	1.25 (0.44)	1.97 (2.67)	10.67 (13.1)	11.26 (11.71)
median RMSD[a] (Å)	1.25	1.70	2.08	3.07
mean (std) ZEAL score	0.94 (0.03)	0.96 (0.02)	0.80 (0.21)	0.79 (0.20)
median ZEAL score	0.95	0.96	0.93	0.91

[a]RMSD computed using the same corresponding residues as those from TMalign.
[b]MS3Align crashed for 20 % of the alignments and data for these are omitted in the statistics.

Table S6. Benchmark statistics for the low TM-score data set

	TMalign	**ZEAL**	**HEX**	**MS3Align**[b]
mean (std) RMSD[a](Å)	5.81 (0.89)	21.13 (8.55)	22.06 (7.34)	22.19 (6.35)
median RMSD[a] (Å)	5.81	20.44	21.60	21.27
mean (std) ZEAL score	0.62 (0.10)	0.81 (0.04)	0.56 (0.15)	0.60 (0.11)
median ZEAL score	0.63	0.81	0.55	0.60

[a]RMSD computed using the same corresponding residues as those from TMalign.
[b]MS3Align crashed for 16 % of the alignments and data for these are omitted in the statistics.

S3.1. Order and scaling dependence

The performance of ZEAL depends on the expansion order. Figure S9 shows C_α RMSD distributions for the high TM-score set obtained from TM-align and ZEAL using moments computed up to order 6, 20 and 30 and with the same residue mapping as that of TM-align. Summarizing statistics are shown in table S7. Using order 30 results in a slight improvement in performance compared to order 20; 94 % of the alignments have RMSD within 1 Å from those obtained using TM-align when moments up to order 30 are used, compared to 90 % using moments uo to order 20. However, this comes at a much higher computational cost. Using low expansion orders results in faster alignments, but comes at the expense of worse performance; 57 % of the alignments have RMSD within 1 Å from those obtained using TM-align. As a trade-off between accuracy and computational speed, the default expansion order in ZEAL is set to 20.

In contrast to the expansion order, the alignment performance is much less sensitive to the scaling factor used in the normalization step required for the moment computations. Figure S10 shows C_α RMSD distributions for the high TM-score set obtained from TM-align and ZEAL, were protein surfaces are scaled such that the maximum distance from the geometric center, $R_{\max}$, corresponds to 70, 80 and 90 % of the unit sphere radius. Summarizing statistics are shown in table S8. At least for the scaling factors studied here, the performance is essentially the same. However, the shape matching performance from ZC shape descriptors are likely more sensitive to the type of normalization scheme employed.

Table S7. Benchmark statistics for the high TM-score data set: ZEAL order dependence

	TMalign	order 6	order 20	order 30
mean (std) RMSD[a](Å)	1.25 (0.44)	3.75 (5.93)	1.97 (2.67)	1.91 (2.64)
median RMSD[a] (Å)	1.25	2.14	1.70	1.61

[a]RMSD computed using the same corresponding residues as those from TMalign.

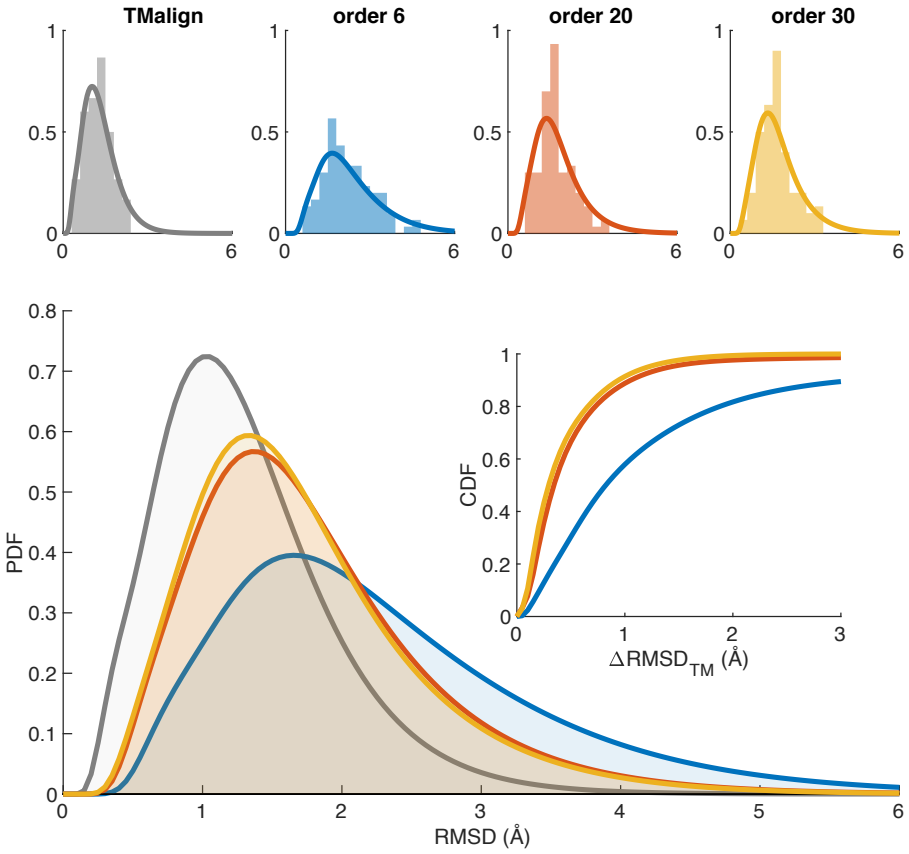

Figure S9. Benchmark results for the high TM-score data set using moment-expansions up to order 6, 20 and 30 in ZEAL. The probability density functions (PDF) of the backbone C_α RMSD were computed using the residue mapping from TM-align. The PDFs were estimated (kernel density) from the corresponding histograms shown individually for TMalign and each order used in ZEAL (top). The inset shows the associated cumulative distribution function (CDF) for the RMSD difference relative TM-align.

Table S8. Benchmark statistics for the high TM-score data set: ZEAL scaling dependence

	TMalign	**0.7·R$_{max}$**	**0.8·R$_{max}$**	**0.9·R$_{max}$**
mean (std) RMSD[a](Å)	1.25 (0.44)	1.97 (2.68)	1.97 (2.68)	1.80 (1.38)
median RMSD[a] (Å)	1.25	1.69	1.67	1.64

[a]RMSD computed using the same corresponding residues as those from TMalign.

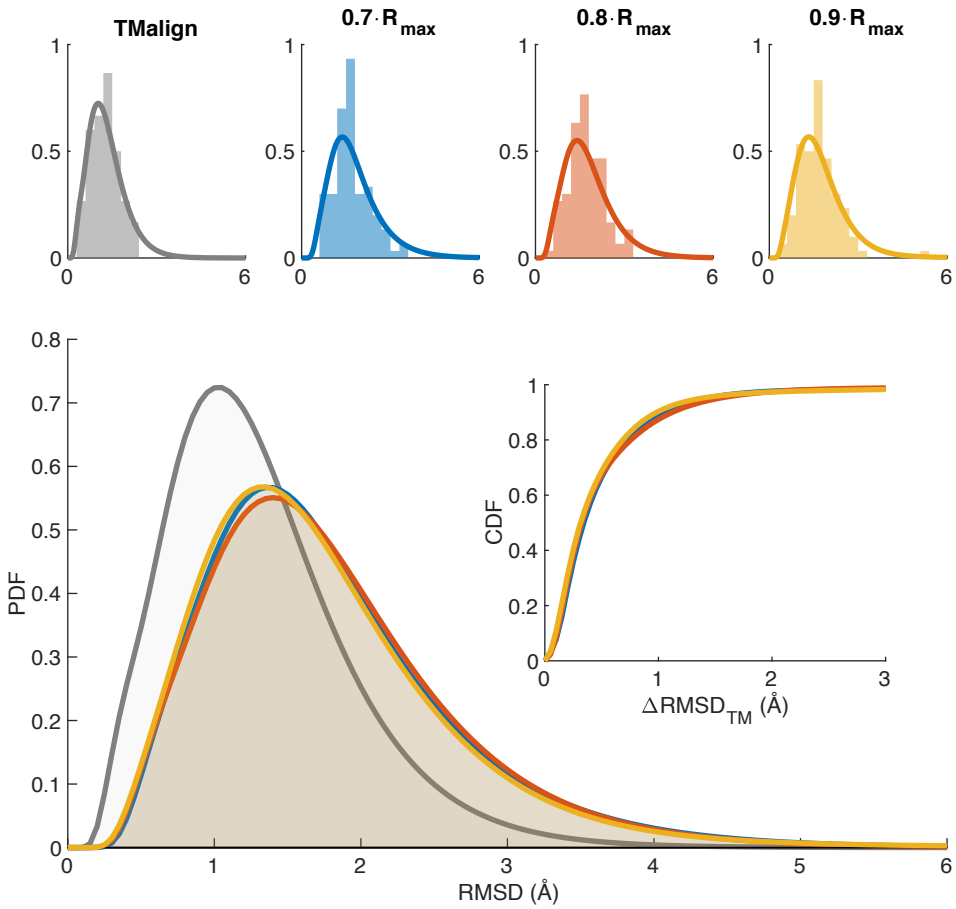

Figure S10. Benchmark results for the high TM-score data set using different scaling factors for the object normalization step. The protein surfaces are scaled such that the maximum distance from the geometric center, $R_{\max}$, corresponds to 70, 80 and 90 % of the unit sphere radius. The probability density functions (PDF) of the backbone C_α RMSD were computed using the residue mapping from TMalign. The PDFs were estimated (kernel density) from the corresponding histograms shown individually for TMalign and each scaling factor used in ZEAL (top). The inset shows the associated cumulative distribution function (CDF) for the RMSD difference relative TMalign.

S4. Coupling of shape and function

Figure S11 below shows the alignment from TM-align and ZEAL of two transferases with similar shape (PDB ID codes 5AN1 and 3UAW) that share 22 % sequence identity as determined by BLAST (query coverage 43 % and E-value 0.001). Without more information, it is difficult to say which alignment is biologically relevant and that captures the evolutionary relatedness, if any.

In general, we would like to probe the coupling between shape and function of two proteins, beyond sequence and secondary structure, as this could be the manifestation of divergent or convergent evolution. To do this, we start with the $S1$ data set with $N_{S1} = 18,965$ single-chain protein structures. As a proxy for function, we use the keyword (KW) annotations for each structure from UniprotKB[5]. We then form the $S1p$ set containing all $N_{S1}(N_{S1}-1)/2$ unique and non-identical pairs of proteins from the $S1$ set, and find the subset $A1$ which have similar shape and the subset $A2$ which have structures that are not similar. We determine the degree of shape similarity by computing the Euclidean distance of the ZCDs and find the $N_{A1} = 161,490$ pairs with distance < 0.025 - an empirically determined cut-off for structures having similar shape. As a proxy for structural similarity, we compute the TM-score for each pair in $S1p$ using TM-align (Version 20190818)[7] and find the $N_{A2} = 138,214,921$ pairs with TM-score < 0.3. We then construct the set A which is the intersection of set $A1$ and $A2$; these $N_A = 103,276$ pairs thus have similar shape and different secondary structures.

We define the following two events:

A: Two protein structures have similar shape, but are structurally dissimilar.

B: Two proteins are annotated with the same keyword

and the corresponding probabilities

$$P(A) = N_A \left(\frac{N_{S1}(N_{S1}-1)}{2} \right)^{-1} \tag{0.1}$$

$$P(B) = \frac{N_{KW}}{N_{S1}} \cdot \frac{N_{KW}-1}{N_{S1}-1} \tag{0.2}$$

If A and B are independent then

$$\kappa = \frac{P(A \cap B)}{P(A)P(B)} \tag{0.3}$$

will be equal to 1, where $P(A \cap B)$ is the joint probability of having a pair with similar shape, different secondary structure (as probed by the TM-score) and similar function (as probed by keywords).

Our hypothesis is that shape alone is independent from function, so we define

$$H_0 : \quad \frac{P(A \cap B)}{P(A)P(B)} = 1 \tag{0.4}$$

$$H_1 : \quad \frac{P(A \cap B)}{P(A)P(B)} > 1 \tag{0.5}$$

where H_0 is the null hypothesis and H_1 is the alternative hypothesis, i.e. that protein global shape intrinsically carries information about protein function. A significance test of κ, for a given KW, is performed using a permutation test. This is a resampling method where a permutation distribution - approximating the sampling distribution - is created consistent under the null hypothesis, that is, the distribution of κ due to the randomness in selecting our S1 sample data set. From the permutation distribution we then locate our observed κ and compute the P-value. This is the probability, if H_0 is true, that we would observe a value as extreme or more extreme than the one we did observe.

The permutation distribution is created by repeatedly computing κ from a permutation resample where the keyword annotations have been randomly reshuffled among the structures in the S1 set. For instance, if three proteins in the set {1 2 3* 4 5* 6* 7 8 9 10} have the * keyword in a sample, then a permutation resample could be {1* 2 3 4 5* 6 7 8* 9 10}. If the permutation distribution is consistent under H_0, then it should be centered at κ=1. Figure S12 shows permutations distributions from 10^5 permutation resamples for the keywords "kinase", "calcium", "DNA-binding" and "zinc-finger".

All distributions are centered at 1 and are approximately normal. With a significance level $\alpha = 0.1$ %, we cannot reject H_0 for proteins having the "kinase" or "calcium" keyword. But for "DNA-binding" and "zinc-finger" proteins we have P-values $\ll \alpha$ for which we can reject H_0 in favour of H_1. Thus, for these proteins, the global shape intrinsically correlates with function (i.e. the keywords) and the $A \cap B$ set of proteins pairs could be the result of functional convergence by evolution.

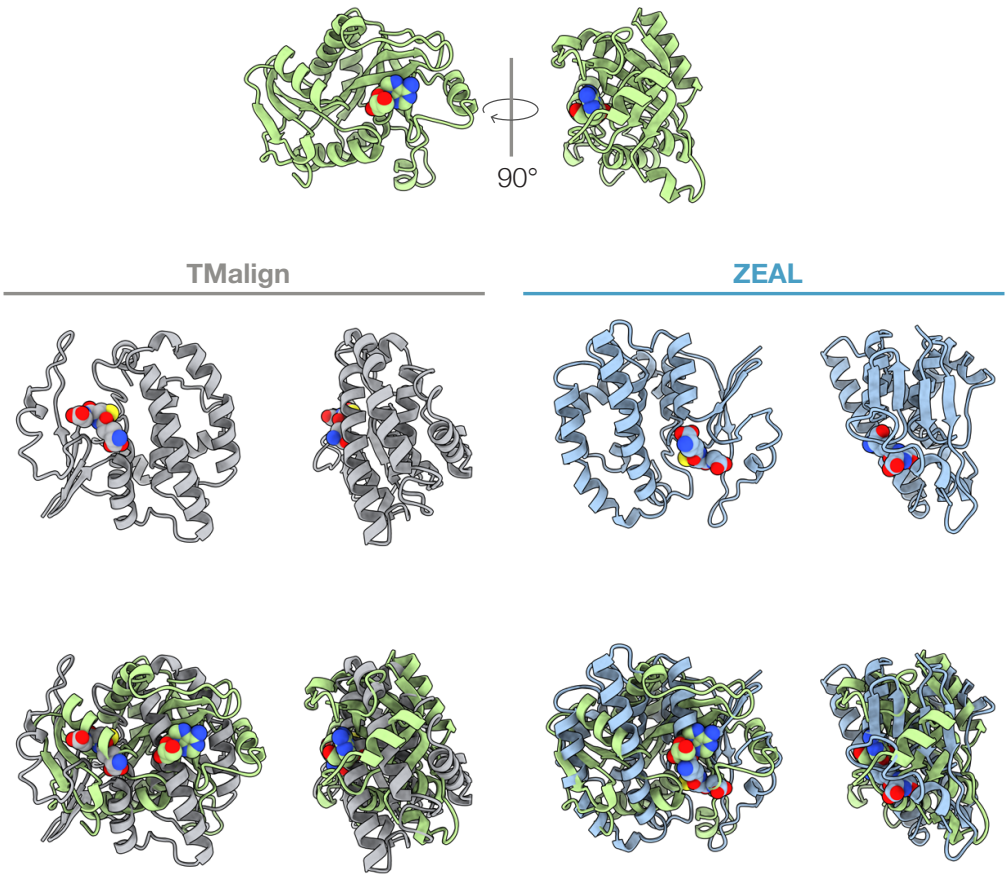

Figure S11. Structure and shape alignment of 4AN1 (chain A, green) and 3UAW (chain A) using TM-align (grey) and ZEAL (blue) respectively. The ligands are shown in vdW representation with carbon atoms shown in the same colour as the cartoon representations. Oxygen, nitrogen and sulphur atoms are shown in red, blue and yellow respectively.

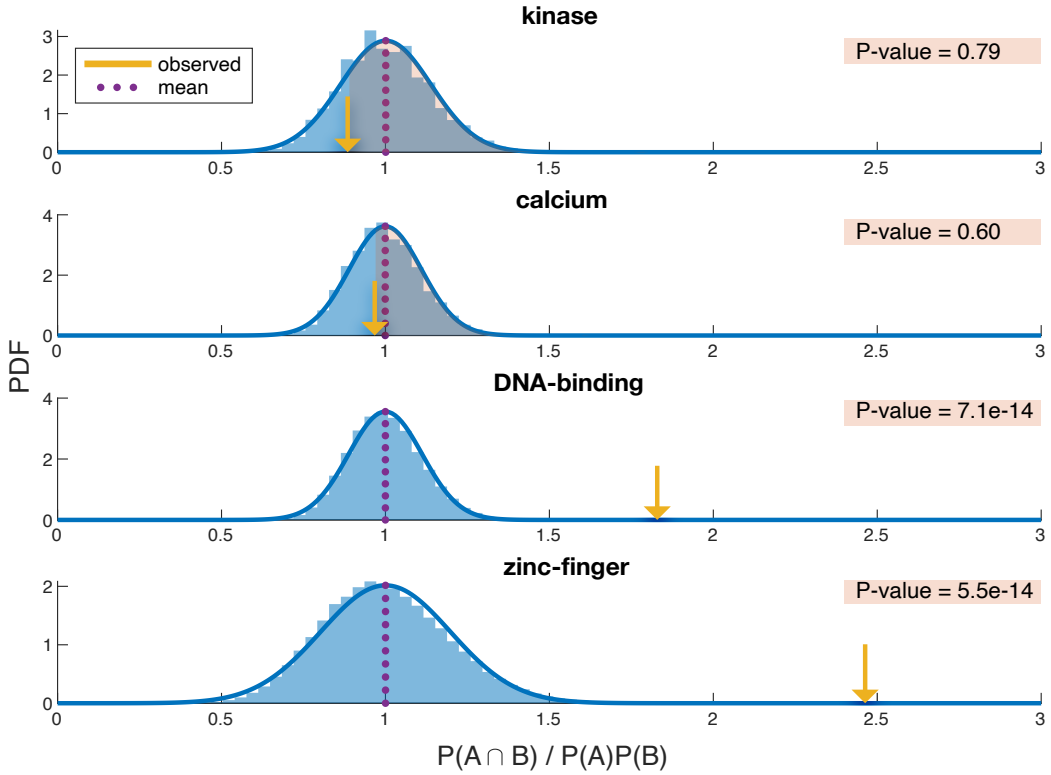

Figure S12. Permutation distribution of the κ statistic using 10^5 permutation resamples for the "kinase", "calcium", "DNA-binding" and "zinc-finger" keywords. The observed value in each distribution is shown by the yellow arrow together with the corresponding P-value, computed by integrating (shown as transparent red) the fitted normal distribution (blue line).

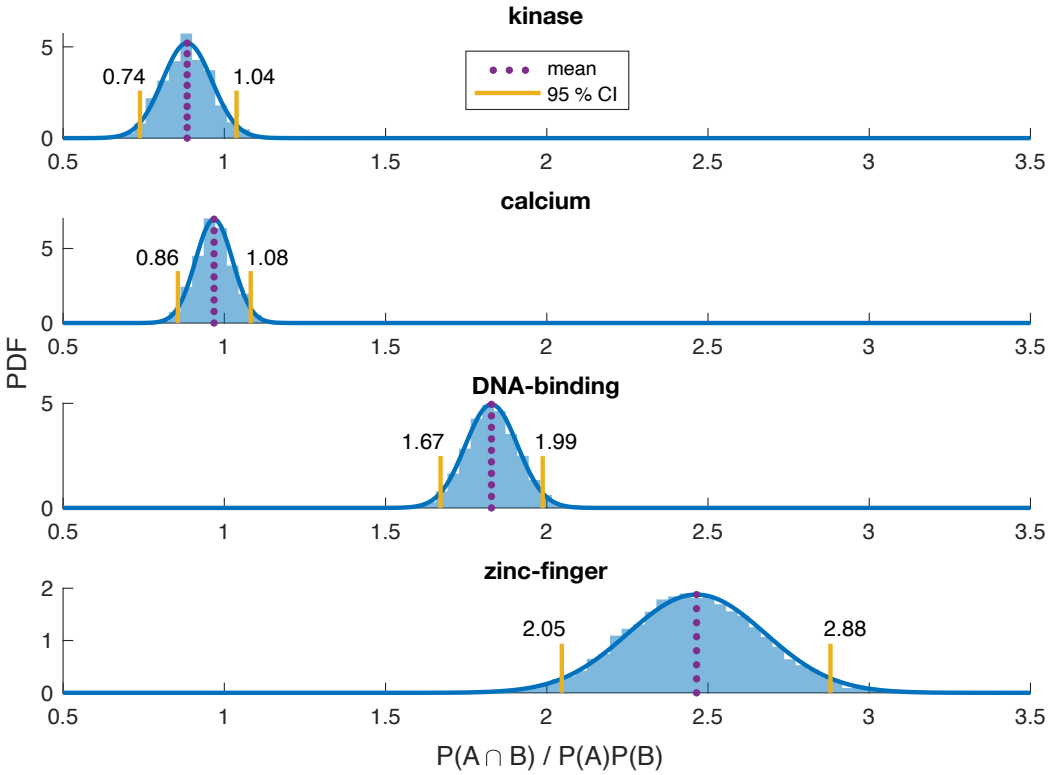

Figure S13. Bootstrap distribution of the κ statistic using 10^4 resamples taken (with replacement) from the S2 set. The 95 % boostrap percentile confidence interval is shown for each distribution.

Table S9. A selection of keywords with significant $\kappa > 1$.

keyword	instances[a]	fraction[a]	kappa	p-value[b]	95 % CI [c]	
ribonucleoprotein	120	0,006	4,64	6,44E-11	2,68	6,83
chromosome	202	0,011	3,33	2,36E-12	2,39	4,34
isopeptide bond	585	0,031	2,65	1,65E-25	2,34	2,98
differentiation	178	0,009	3,54	1,38E-10	2,32	4,75
zinc-finger	439	0,023	2,46	5,46E-14	2,05	2,88
ubl conjugation	874	0,046	2,20	3,98E-28	2,03	2,37
cleavage on pair of basic residues	142	0,007	3,29	1,18E-06	1,91	4,87
nucleus	1672	0,088	1,99	9,17E-39	1,89	2,08
chromatin regulator	179	0,009	2,95	5,55E-07	1,86	4,15
alternative splicing	2144	0,113	1,89	2,57E-40	1,82	1,97
cell adhesion	164	0,009	2,91	2,61E-06	1,82	4,12
hydroxylation	78	0,004	4,66	5,77E-06	1,74	8,12
dna-binding	984	0,052	1,83	7,08E-14	1,67	1,99
developmental protein	251	0,013	2,33	4,68E-06	1,66	3,05
repeat	1684	0,089	1,73	2,50E-21	1,65	1,82
phosphoprotein	2924	0,154	1,67	2,59E-34	1,62	1,72
transcription	1007	0,053	1,76	2,22E-12	1,61	1,91
transcription regulation	970	0,051	1,75	1,06E-11	1,60	1,91
cell junction	264	0,014	2,14	2,81E-05	1,52	2,81
lipid-binding	184	0,010	2,33	1,85E-04	1,49	3,35
activator	285	0,015	2,02	8,85E-05	1,46	2,58
membrane	2665	0,141	1,45	1,24E-14	1,40	1,50
innate immunity	183	0,010	2,31	4,61E-04	1,36	3,35
rna-binding	582	0,031	1,60	1,04E-04	1,35	1,84
transmembrane	1521	0,080	1,43	1,94E-07	1,34	1,52
lipoprotein	495	0,026	1,61	3,68E-04	1,32	1,91
methylation	345	0,018	1,73	8,90E-04	1,32	2,17
repressor	309	0,016	1,82	6,21E-04	1,32	2,34

[a] Total number of proteins in the S1 set with given keyword. [b] P-values computed using 10^5 permutation resamples from the S1 set. [c] 95 % boostrap percentile confidence interval using 10^4 resamples (with replacement) from the S2 set.

Section S5. ZEAL graphical user interface

The ZEAL graphical user-interface (Figure S14) showing a session of local shape-alignment of two DNA-binding proteins with PDB ID codes 4KIS (chain A, orange) and 1SKN (chain P, blue). Supplementary Figure S11 A and B show both proteins complexed with DNA. The region of interest (ROI), the zinc-ribbon domain of 4KIS (residues 270-310) that bind to DNA, is selected in a JSmol[3] window (top left). The shape-alignment search is monitored in the optimization window (top right), showing the best ZEAL score found vs. the number of iterations. The corresponding rotation of the shapes is updated in real-time in the main window (middle). The search stops when the maximum number of iterations have been completed, or when the user press a stop-button in the optimization window. The alignment before and after the search can be previewed in JSmol (bottom left and right) in the conventional ball-and-stick, van der Waals or cartoon representation. Supplementary Figure S11 shows the superposition including DNA structures.

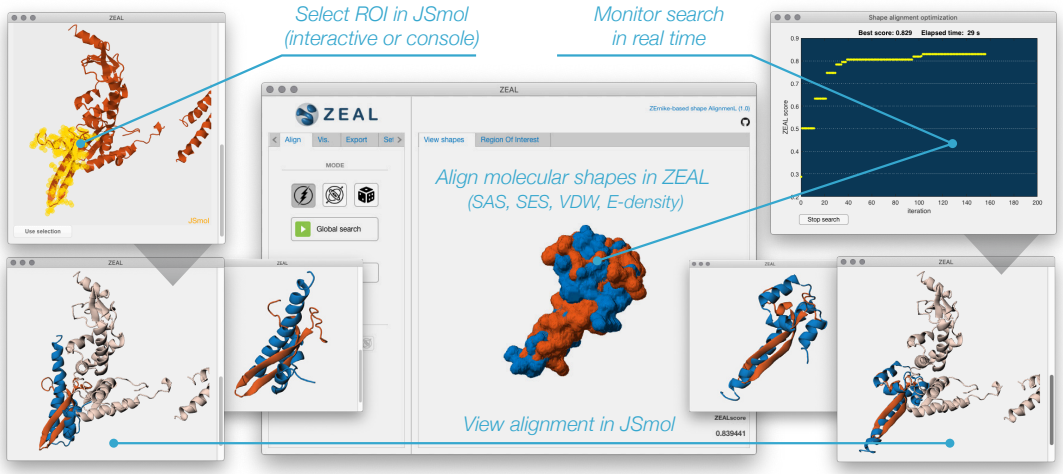

Figure S14. The ZEAL graphical user interface with screenshots from a local superposition session.

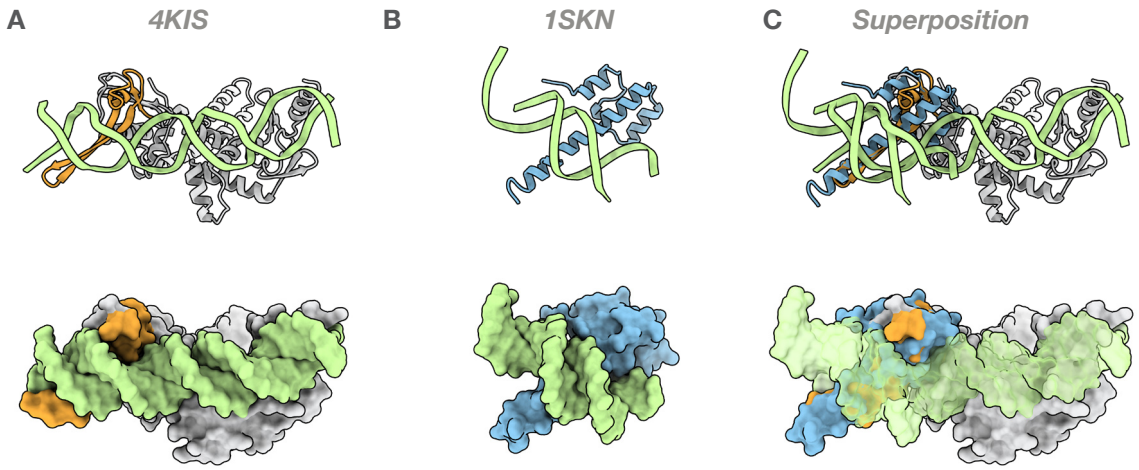

Figure S15. Cartoon and surface visualizations of the structures exported from ZEAL in the example shown in Figure S10. The region of interest of 4KIS is shown in orange, 1SKN is shown in blue and DNA in green.

References

(1) Bondi, A. *Journal of Physical Chemistry* **1964**.

(2) Fox, N. K.; Brenner, S. E.; Chandonia, J. M. *Nucleic Acids Research* **2014**.

(3) Hanson, R. M.; Prilusky, J.; Renjian, Z.; Nakane, T.; Sussman, J. L. *Israel Journal of Chemistry* **2013**, *53*, 207–216.

(4) Novotni, M.; Klein, R. *Proceedings of the eighth ACM Symposium on Solid Modeling and Applications* **2003**, 216–225.

(5) UniProt Consortium, T. *Nucleic Acids Research* **2018**, *46*, 2699–2699.

(6) Xu, D.; Zhang, Y. *PLoS ONE* **2009**, *4*, ed. by Buehler, M. J., e8140.

(7) Zhang, Y.; Skolnick, J. *Nucleic Acids Research* **2005**, *33*, 2302.