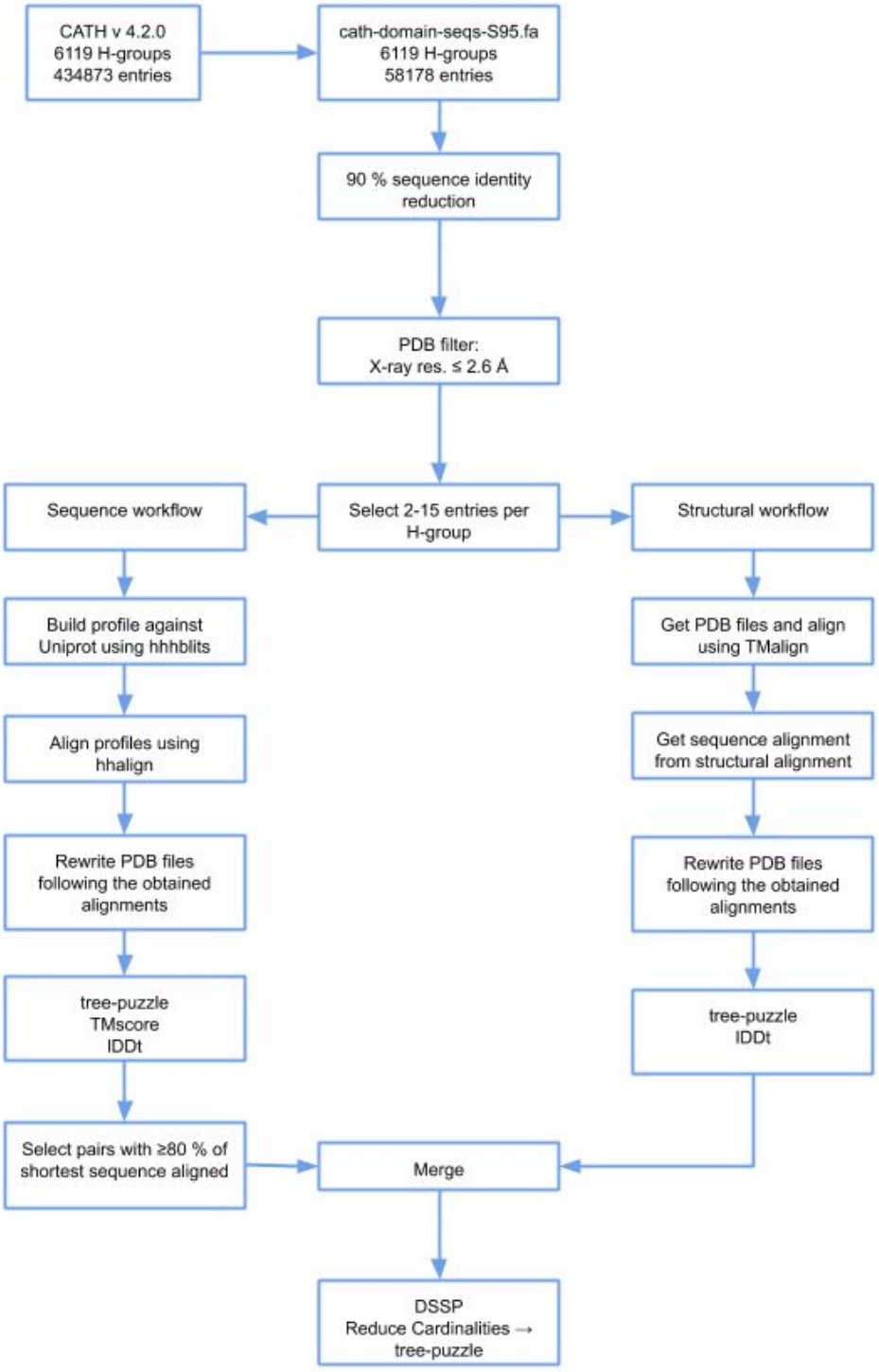

Figure S1. Workflow visualization.

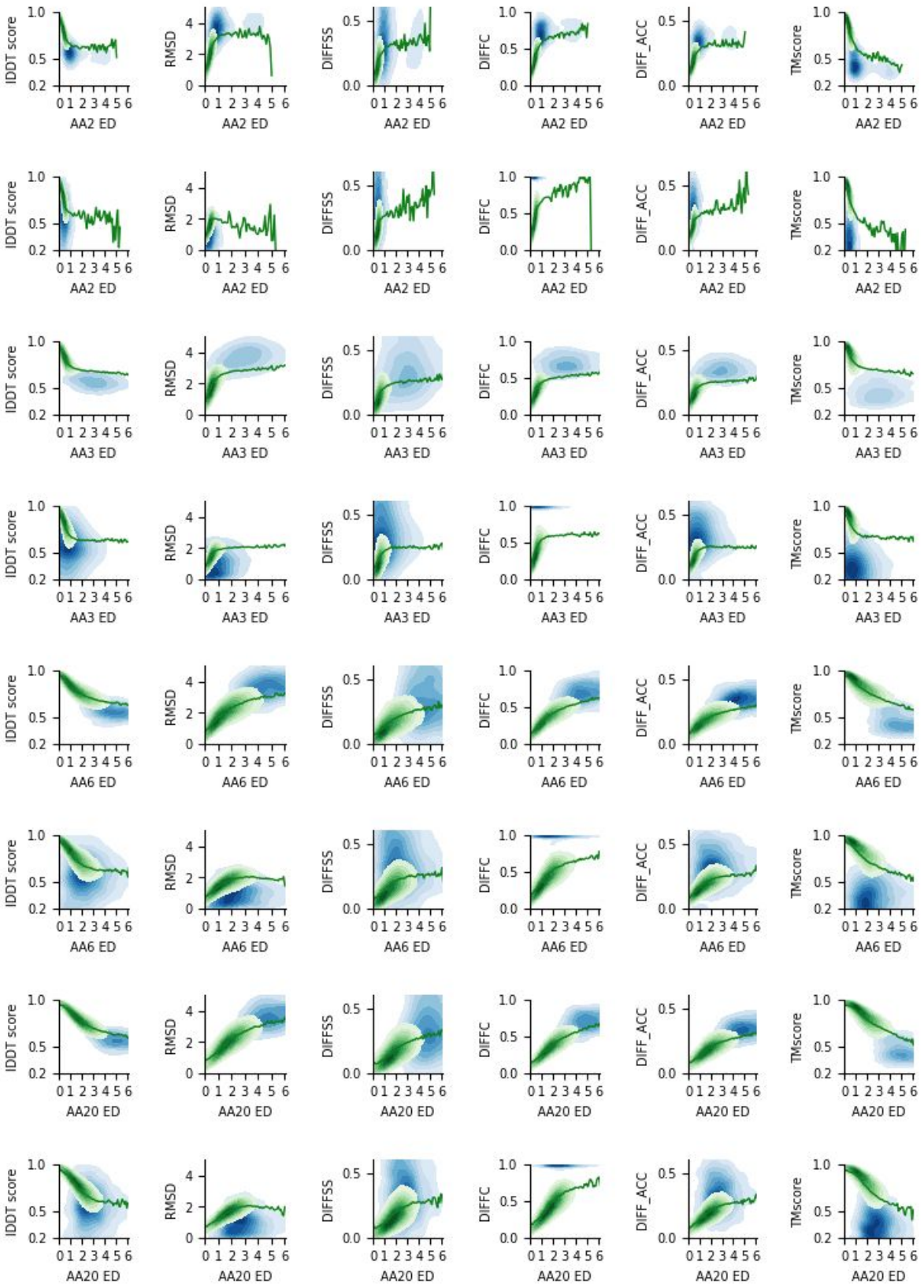

Figure S2. Running averages for ML distance against lDDT score, RMSD, DIFFSS, DIFFC, DIFF_ACC and TMscore for different cardinalities using the structural workflow and sequence workflow. The green kde plot represents the homology dataset and the blue the fold dataset. The thick green lines are the total running averages computed using the broad

dataset with 0.1 ED intervals for each cardinality. The first row for each score represents the results from the structural workflow and the second, those from the sequence workflow.

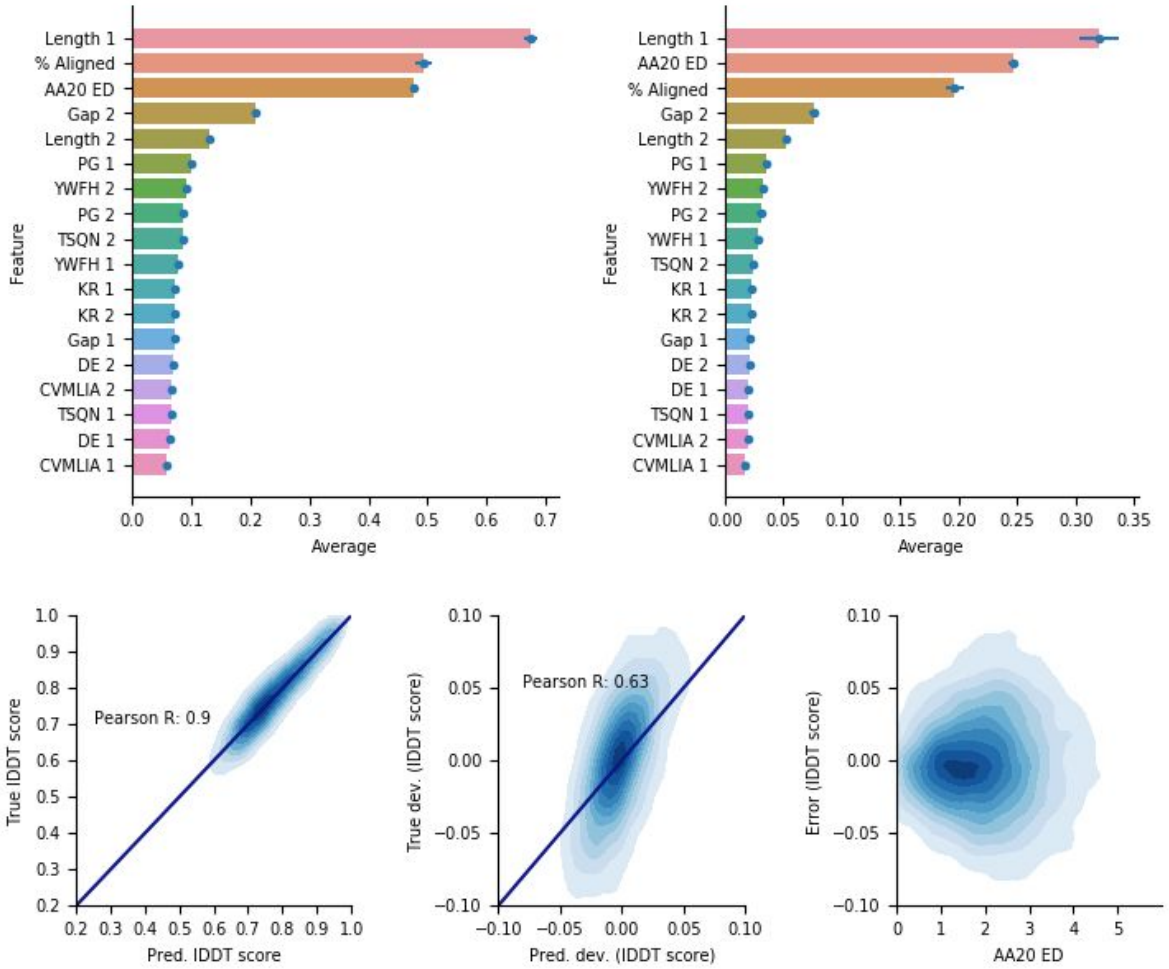

Figure S3. Feature importance by query domain (1) and subject (2) with standard deviations for predicting deviating pairs across a five-fold cross validation, using only sequence features. The left feature importances are calculated on the train set and the right on the test set, both are permutation importances (the decrease in predicted score when a single feature value is randomly shuffled). The true scores are compared with the predicted ones, so is the true deviation and the predicted deviation. The error in predicted score is also assessed for different EDs.

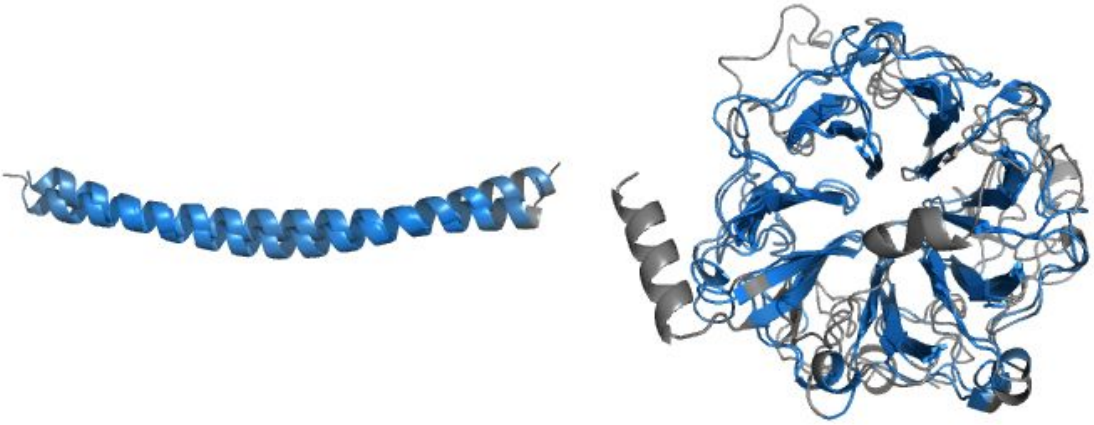

Figure S4. Representative outlier example from Set 1 (1urqA00 and 2ix7C00, fold 1.20.5, left) and Set 2 (3of7A00 and 3kciA00, fold 2.130.10, right). The superposed structures are colored by rmsd differences between their alpha carbons, with grey color representing large and blue small rmsd. As can be seen, the example from Set 1 are almost identical, while that from Set 2 mainly differs in loop regions.

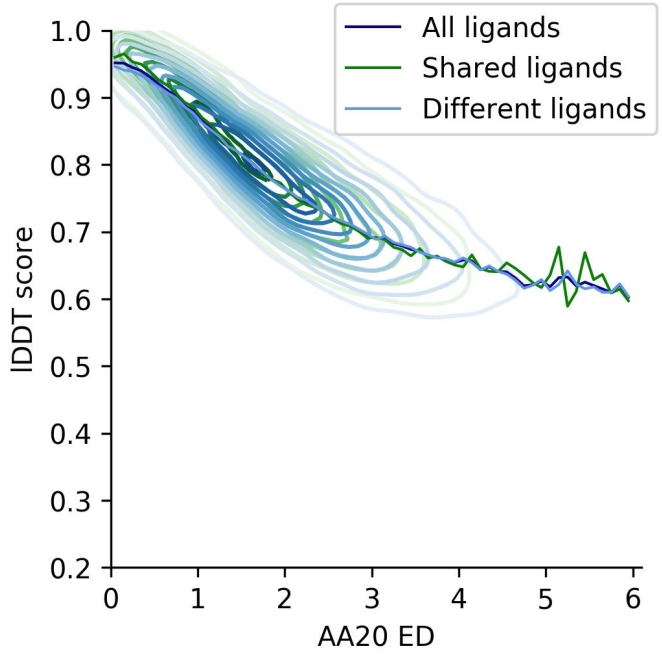

Figure S5. The impact of ligand binding displayed as a comparison between the sequence-structure relationships of the pairs with at least one shared ligand (Shared ligands) and those with no shared ligands (Different ligands). Both running averages (solid lines) and KDE plots are shown for the Shared and Different sets, while the complete Broad dataset (All ligands) is only represented by a running average. The Pearson correlation coefficient between the running averages is 0.99.

Table S1. Folds in outlier set 1 and 2, as well as the number of points from each fold. Fold 1.20.5 is the most represented in set 1, and is also the most deviating fold with at least 10 pairs.

Set 1, 142 pairs		**Set 2, 6 pairs**	
Fold	**Count**	**Fold**	**Count**
1.20.5	80	6	
1.10.1220	11	2.130.10	3
1.10.287	11	2.170.16	3
1.10.12	9		
4.10.410	7		
1.20.20	3		
4.10.260	3		
1.20.58	2		
4.10.280	2		
3.30.429	2		
2.30.110	1		
3.30.63	1		
2.40.170	1		
2.30.30	1		
3.30.70	1		
2.160.10	1		
1.20.1300	1		
3.10.450	1		
3.30.60	1		
2.70.130	1		
2.10.50	1		
2.60.450	1		

Table S2. Percent retained in each step of the workflow for the **Homology Dataset.**

Entries in CATH	#pairs	% pairs	#groups	% groups	# uids	% uids
			6119		434857	
95 % sequence identity reduction			6002	0,9808792286	54271	0,1248019464
PDB filter			2984	0,4971676108	41548	0,7655654033
At least two entries			2489	0,8341152815	39211	0,9437518051
At least two, but under 15 entries	**Number**	**Percent**				
Number of possible pairs	67445					
Number of possible H-groups	2489	1,000				
Number of possible uids	15127	0,386				
Sequence aln	#pairs	% pairs	#groups	% groups	# uids	% uids
hhalign	67445	1,000	2489	1,000	15127	1,000
TMscore	67259	0,997	2489	1,000	15105	0,999
lddt	67045	0,994	2486	0,999	15089	0,997
Merge	67037	1,000	2485	0,998	15087	1,000
80 % selection	49165	0,733	2429	0,977	14661	0,972
Structure aln	#pairs	% pairs	#groups	% groups	# uids	% uids
TMalign .aln	67445	1,000	2489	1,000	15127	1,000
TMalign .tsv	67445	1,000	2489	1,000	15127	1,000
lddt	67320	0,998	2487	0,999	15115	0,999
Merge	67320	1,000	2487	1,000	15115	1,000
Merge sequence (80 %) and structure	#pairs	% pairs	#groups	% groups	# uids	% uids
	49164	1,000	2428	1,000	14659	1,000
	#pairs	% pairs	#groups	% groups	# uids	% uids
DSSP	49103	0,999	2426	0,999	14636	0,998

Table S3. Percent retained in each step of the workflow for the **Fold Dataset.**

Topology workflow 1						
Entries in CATH	#pairs	% pairs	#groups	% groups	# uids	% uids
			1391		434857	
95 % sequence identity reduction			1385	0,9956865564	54271	0,1248019464
PDB filter			1266	0,9140794224	41548	0,7655654033
At least two entries	1933		447	0,3530805687	3866	0,09304900356
Sequence aln	#pairs	% pairs	#groups	% groups	# uids	% uids
hhalign	1931	0,999	445	0,996	3862	0,999
TMscore	1911	0,990	443	0,996	3822	0,990
lddt	1807	0,936	443	0,996	3614	0,936
Merge	1799	0,996	442	0,998	3598	0,996
Structure aln	#pairs	% pairs	#groups	% groups	# uids	% uids
TMalign .aln	1931	0,999	445	0,996	3862	0,999
TMalign .tsv	1931	0,999	445	0,996	3862	0,999
lddt	1919	0,994	444	0,998	3838	0,994
Merge	1919	1,000	444	1,000	3838	1,000
Merge sequence and structure dfs	#pairs	% pairs	#groups	% groups	# uids	% uids
	1799	1,000	442	1,000	3598	1,000
	#pairs	% pairs	#groups	% groups	# uids	% uids
DSSP	1799	1,000	442	1,000	3598	1,000