CASP14 Refinement Assessment

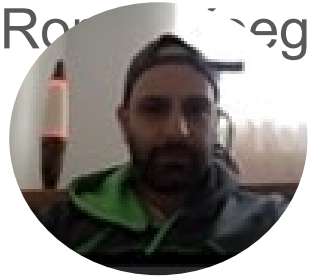
Dan Rigden plus...

Filomeno Sanchez Rodriguez

Adam Simpkin

Shahram Mesdaghi

Ro egan



Marcus Hartmann and Joana Pereira

Andriy K htafovych

Overview

Refinement target selection and properties

Group assessments, overall and by kinds of target

Refinability

Self-assessment of models and residues

Special targets - extended and NMR

Applications - Structure-based function prediction and Molecular Replacement

Conclusions

Target selection and properties

Refinement target selection: Andriy with my input

30 targets. Size <~280 residues, GDT_HA in range 28-80

Often aimed for best server model

Check structural context of errors in multidomain/complex proteins i.e. at least some refinement seemed plausible without knowledge of position of another domain or chain. 7 were domains deriving from multi-domain targets

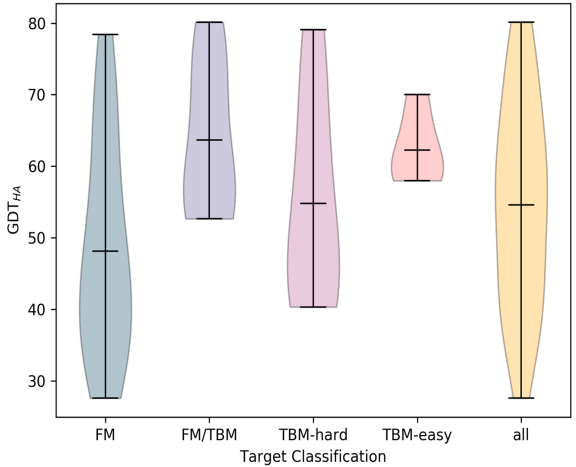
Initially selected few 427 (later = AlphaFold2) group models (too good!) but then decided that perfecting 427 models is just as important a challenge as improving worse models. (Arguably more so if 427 methods become the norm)

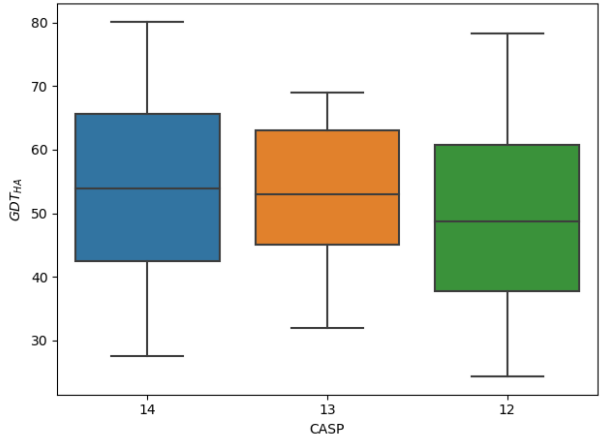
Seven double-barrelled targets (v1/v2) (GDT_HA group 427 53-80; non-427 30-53)

Seven extended targets (x1 or x2) - 6 weeks instead of usual 3 weeks

Refinement targets

Target class	Number of targets (CASP13)	Size in residues		
		min	max	mean
TBM-easy	5 (13)	103	246	165 (132)
TBM-hard	8 (5)	119	221	160 (130)
FM/TBM	6 (5)	75	171	107 (142)
FM	11 (6)	95	276	157 (137)
all	30 (29)	75 (77)	276 (204)	149 (134)





Group assessments, overall and by kinds of target

Standard rankings of group performance

Score comes from ML exercise in CASP12 paper.

"*To benefit from manual assessment while minimizing the pitfalls of subjectiveness and avoiding the definition of arbitrary weights for the different metrics, we used a machine learning approach to devise a linear combination of standard scores based on the visual inspection. Four assessors (LH, VO, HY, and GS) visually inspected all "model 1" predictions for 14 targets (33%) and each independently scored them.*"

$$S_{CASP12} = 0.46\, Z_{RMSD} + 0.17\, Z_{GDT\_HA} + 0.2\, Z_{SphGr} + 0.15\, Z_{QCS} + 0.02\, Z_{MolPrb}$$

Cα positional accuracy

Quality Control Score

Molprobity Score

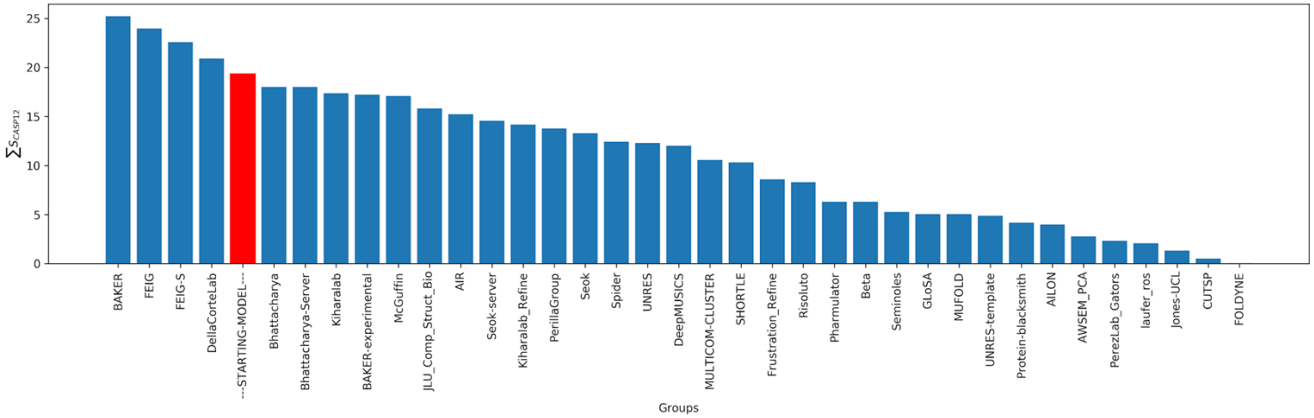
Andriy kindly updated the CASP page to allow analysis on different size and different quality targets

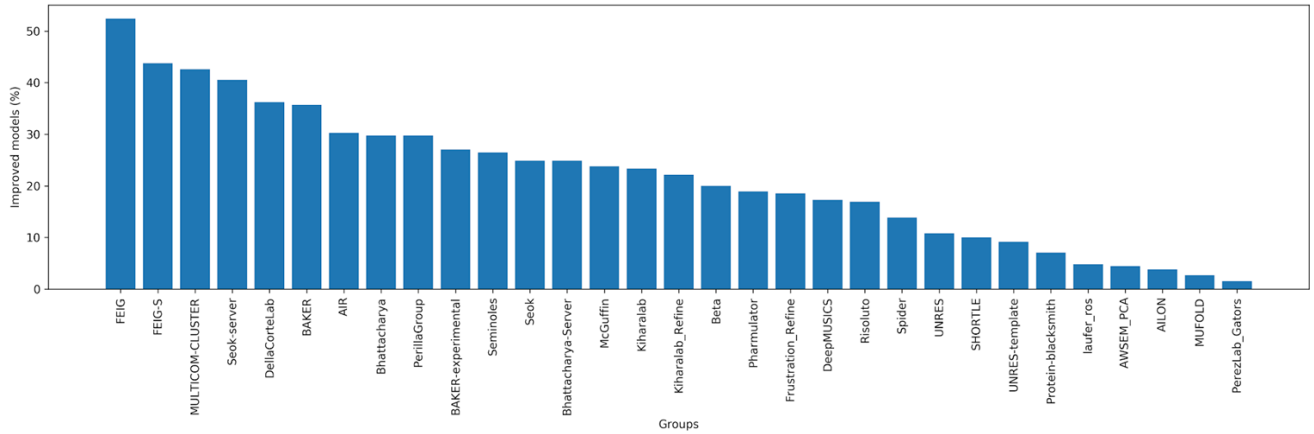
Standard rankings of group performance (model_1)

Only four groups - BAKER, FEIG, FEIG-S, DellaCorte outperform the naive predictor

Same groups high on %improved models, but joined by two more servers MULTICOM-CLUSTER and Seok-server.

Only FEIG group improved more than half

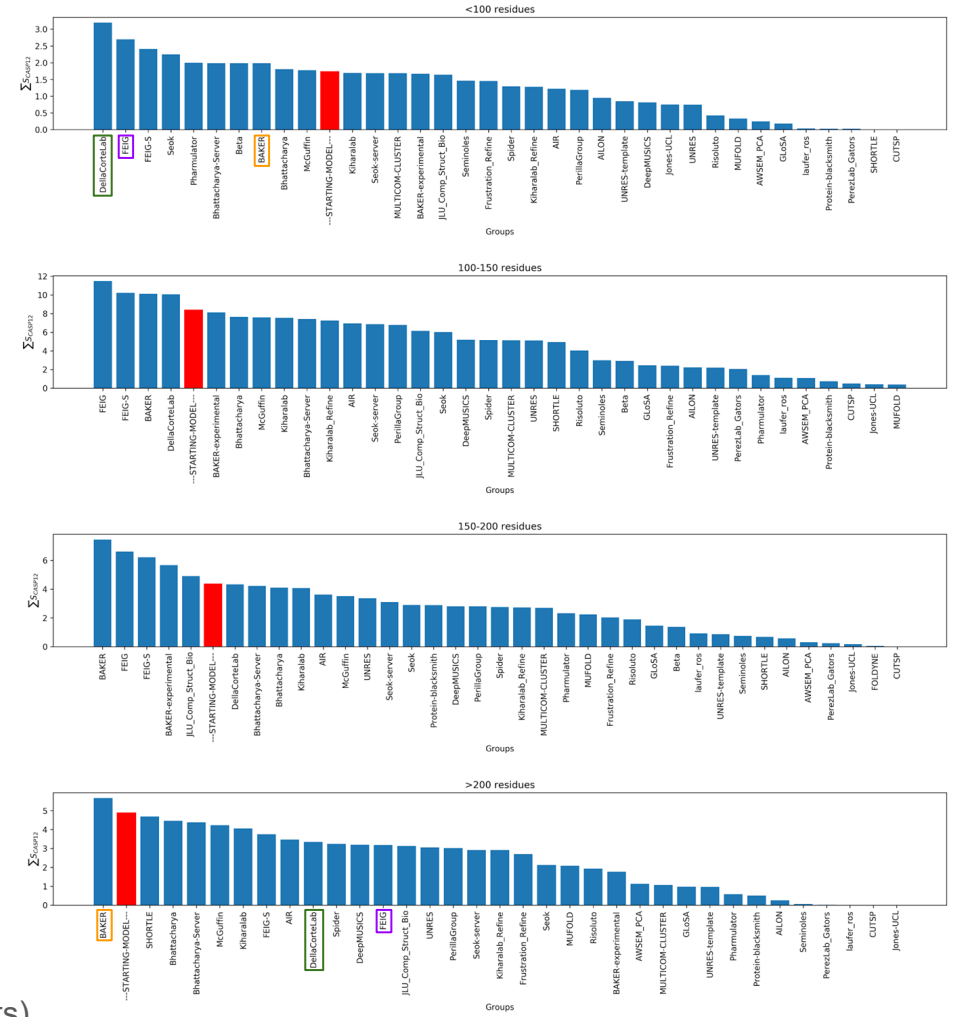




More groups consistently improve small targets. DellaCorteLab, FEIG ahead of BAKER

But only one, BAKER, beats the naive predictor on the largest targets when DelleCorteLab, FEIG well down

(double-barrelled count 2; excludes extended targets)



<100 res, n=4

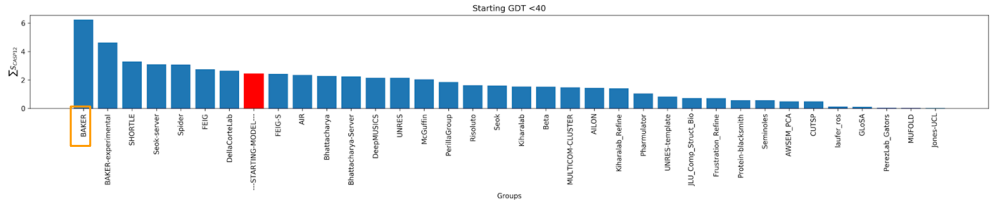
100-150 res, n=16

150-200 res, n=9

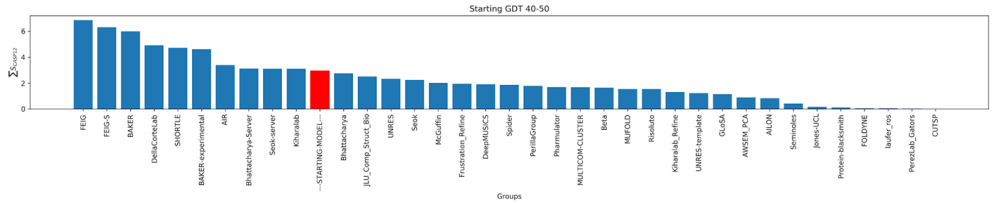
>200 res, n=8

More groups can consistently beat naive for worst starting structures. BAKER is the standout performer on the worst, followed by BAKER-experimental

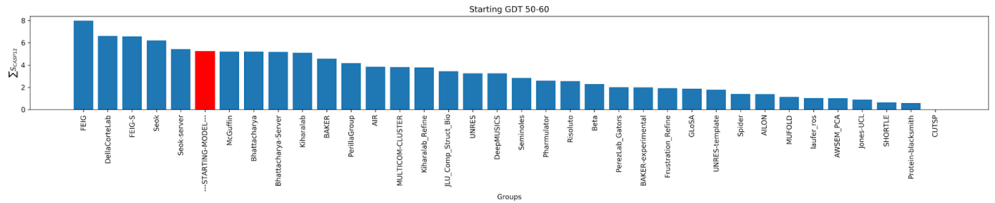
No group consistently beats naive for the best quality targets



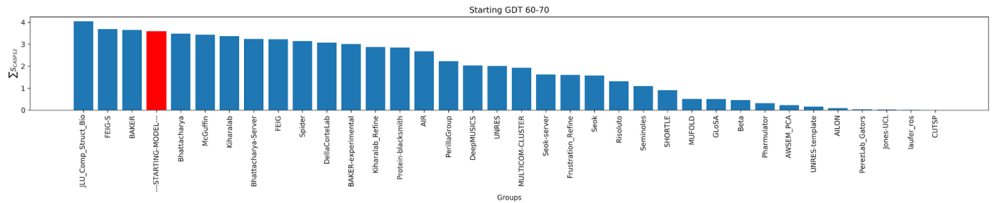
GDT<40, n=6



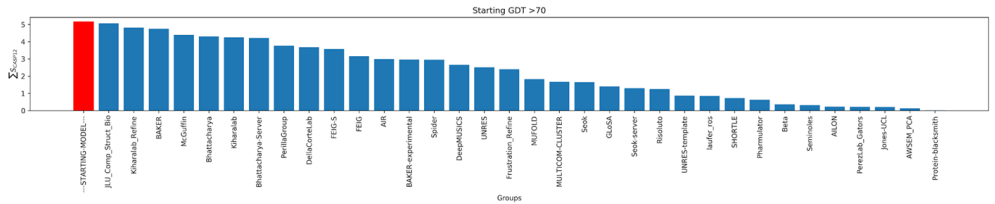
40 < GDT< 50, n=8



50 < GDT< 60, n=10

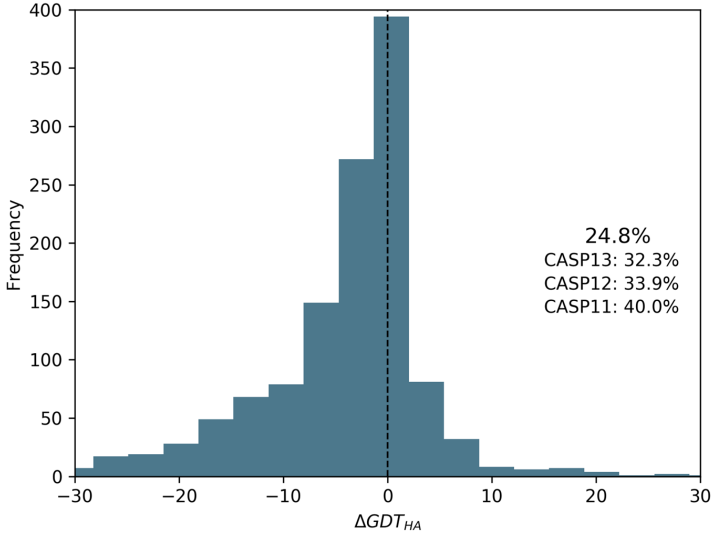


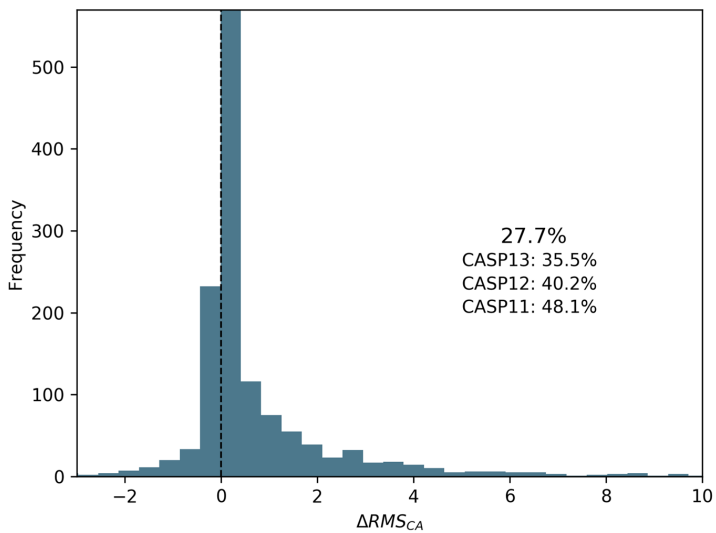
60 < GDT< 70, n=6



GDT>70, n=7

CASP on CASP analysis

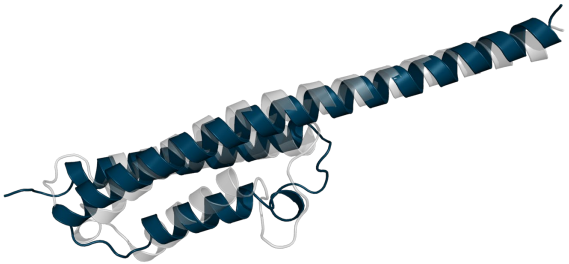




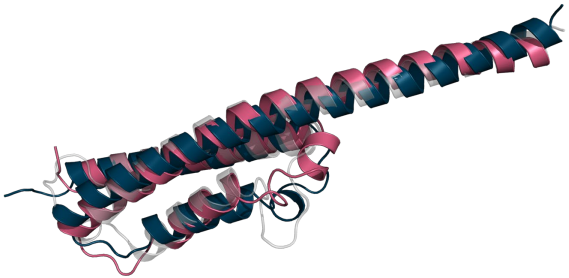
% improved: in line with or arguably worse than previous years

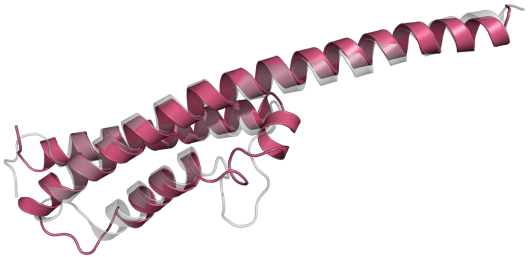
(excluded double-barreled predictions for one group)

Visualisation of the best* refinement (Beta group)



Starting model with target





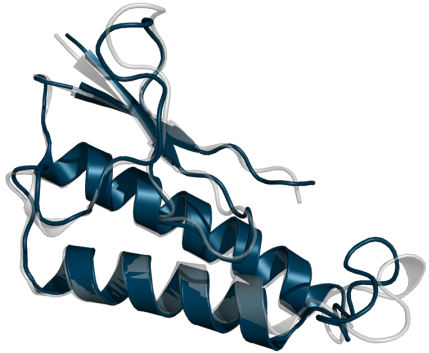
Refined model with target

#	Models	160 170 180 190 200 210 220 230 240 250 260 270	GDT_HA
-	target ss: C E H		-
-	starting model		40.34
1	R1030-D2TS270_5		63.86

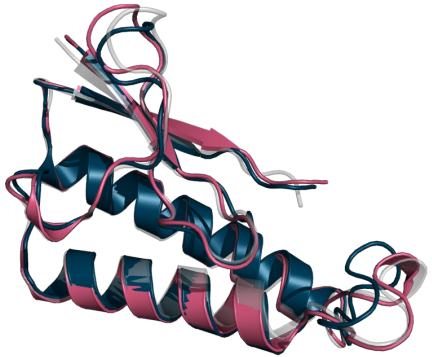
*that we're permitted to show

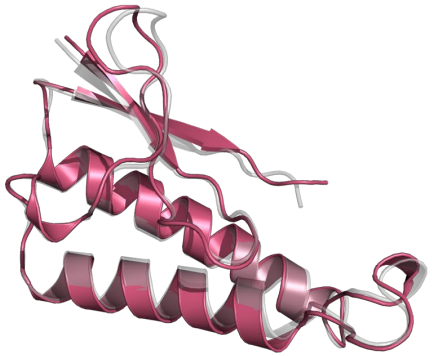
R1030-D2, Beta, deltaGDT_HA = 23.52

Visualisation of a BAKER group refinement



Starting model with target



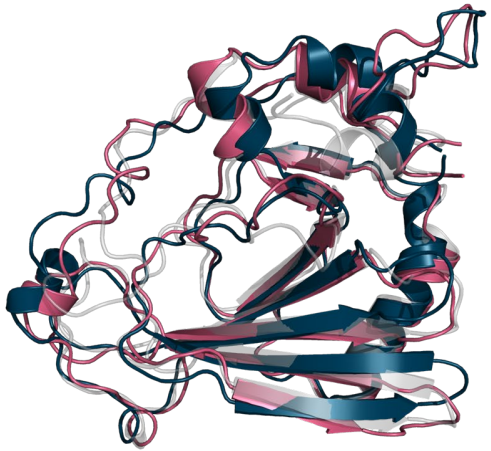


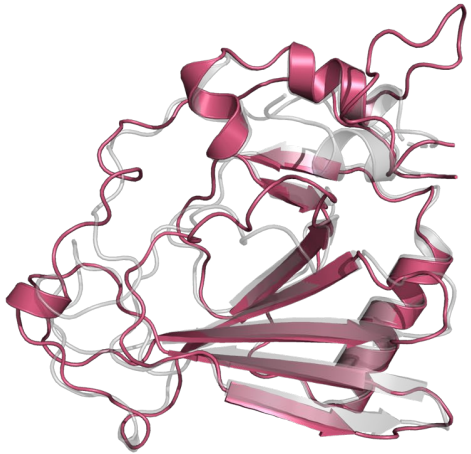
Refined model with target

#	Models	10 20 30 40 50 60 70 80 90	GDT_HA
-	target ss: C E H		-
-	starting model		74.75
1	R1065s2TS473_1		87.76

R1065s2, BAKER group, deltaGDT_HA = 12.99

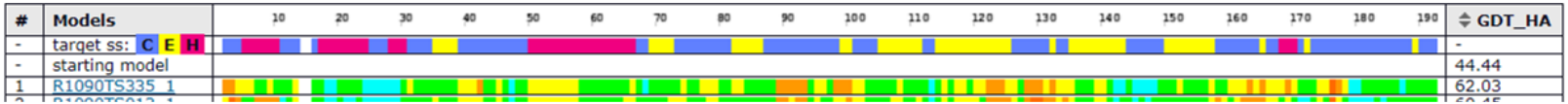
Visualisation of a group FEIG-S refinement





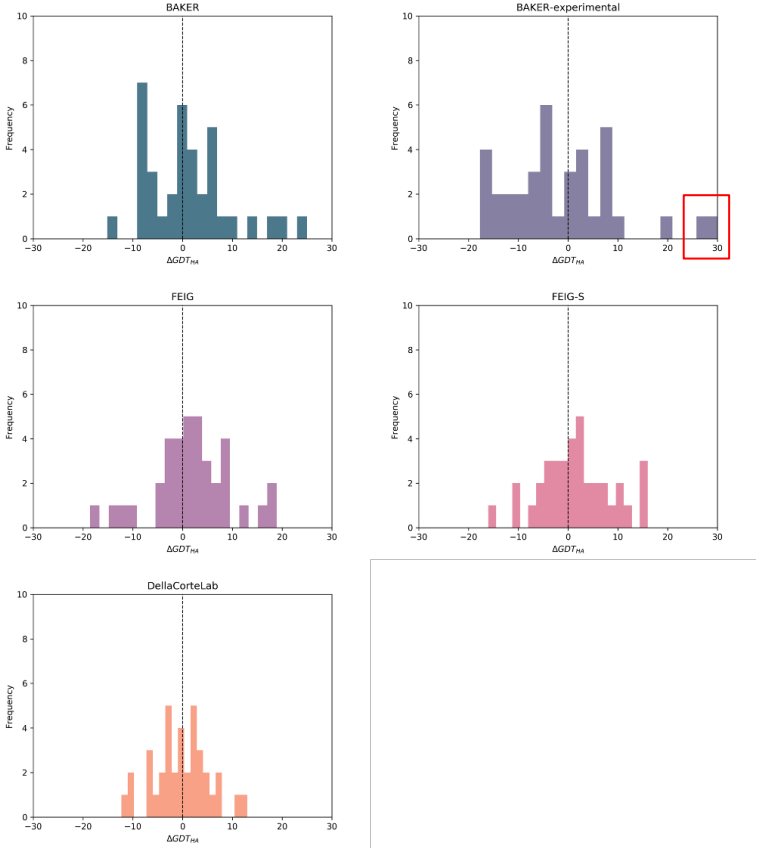
Starting model with target

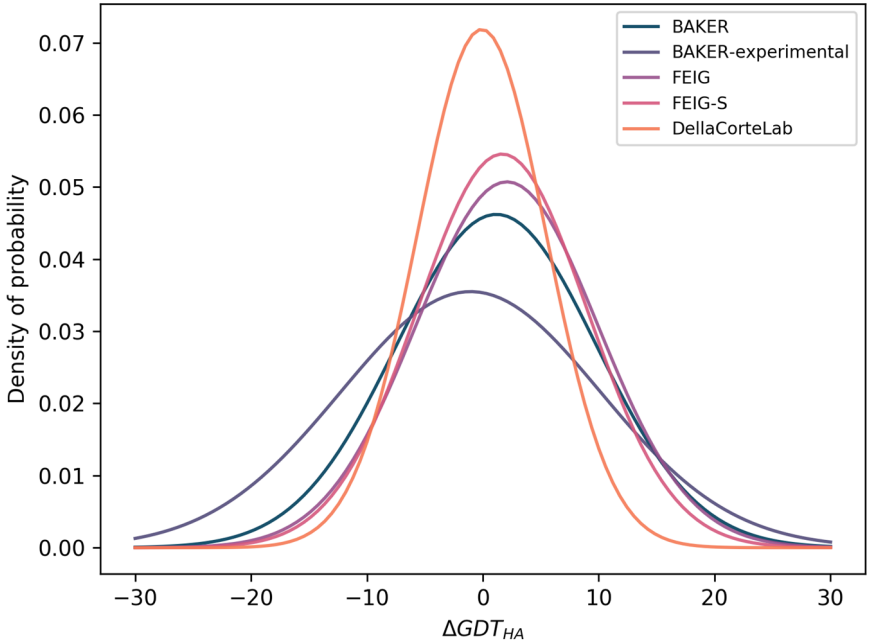
Refined model with target



R1090, FEIG-S group, deltaGDT_HA = 16.01

ΔGDT_{HA} distributions for individual groups





Refinability

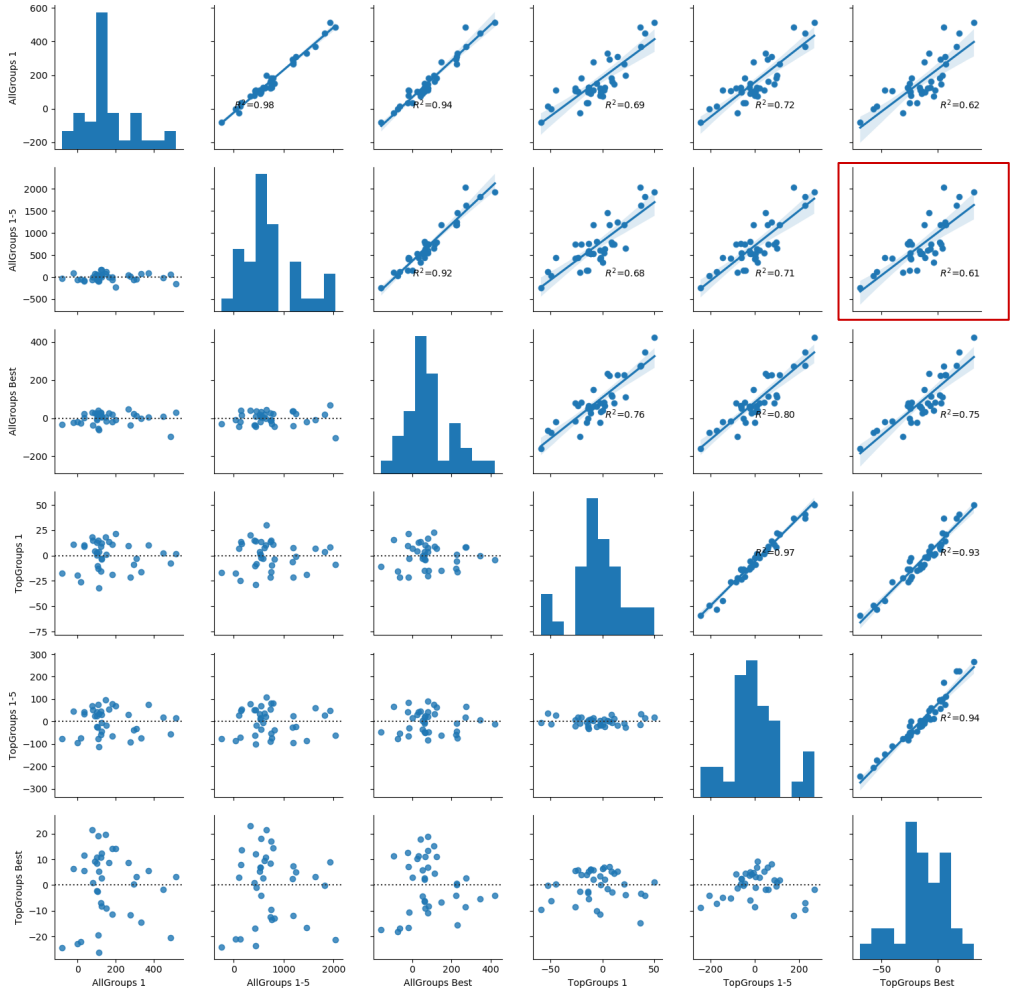
Defining refinability

∑ΔGDT_HA

Six potential refinability all groups or top four x _1 alone of _1 to _5 or the best correlate well

Therefore looked first at all groups, all models

Then at least correlated measure top groups, best model

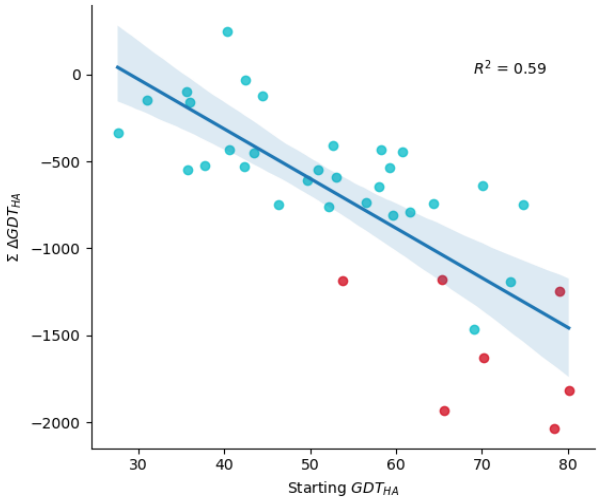


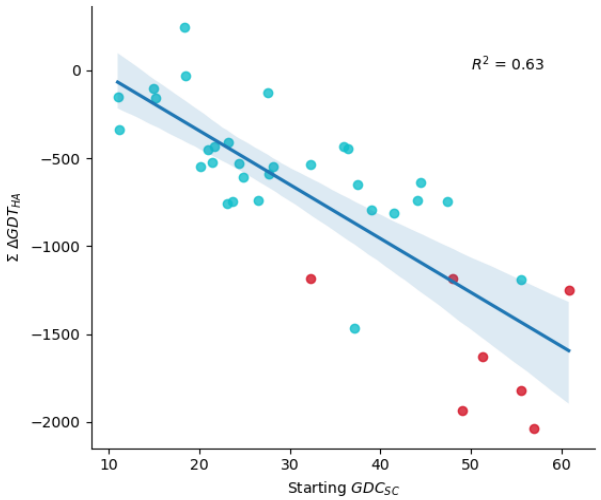
Refinability - all groups, all models.

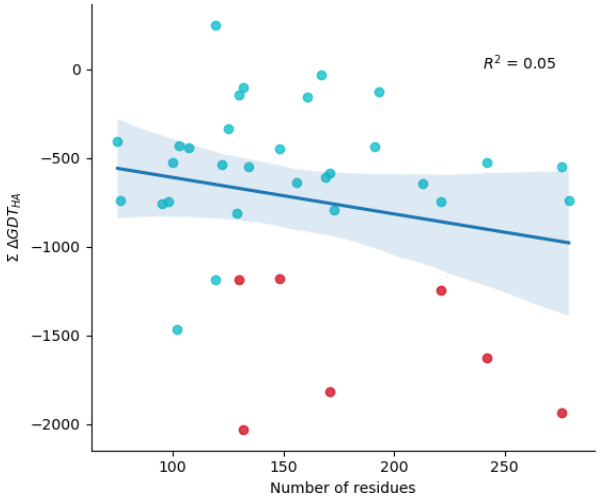
Size and %regular secondary structure not correlated

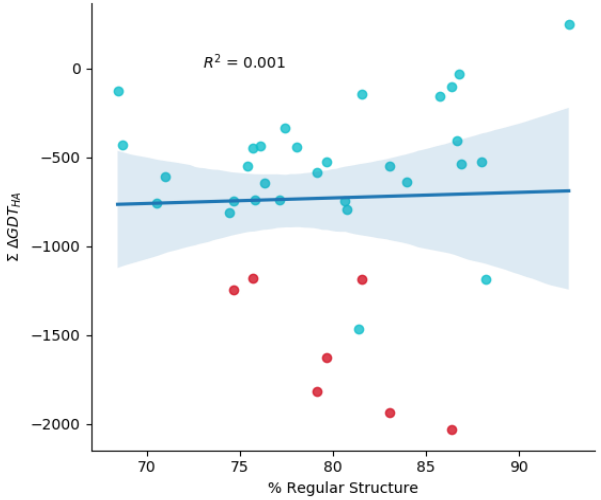
Starting model quality GDT_HA and GDT_SC clearly correlated

AlphaFold2 models are special







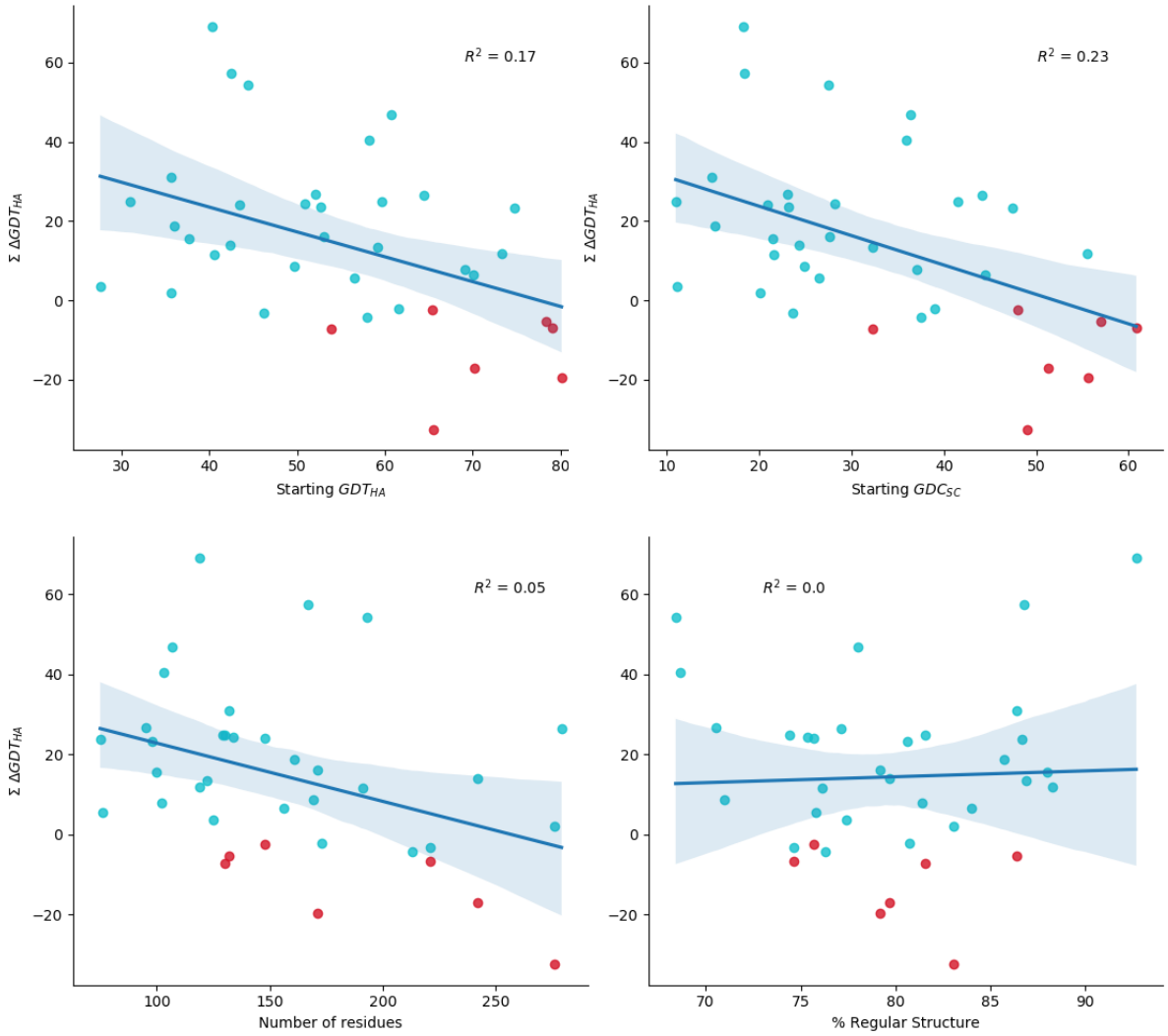


Refinability - top groups, best models.

For top groups best models, the correlation of refinability and starting GDT_HA is much weaker i.e. the best groups do almost as well with good targets as with poor ones

But, AlphaFold2 models, unrefinable!

High-quality models by other groups **are** refinable

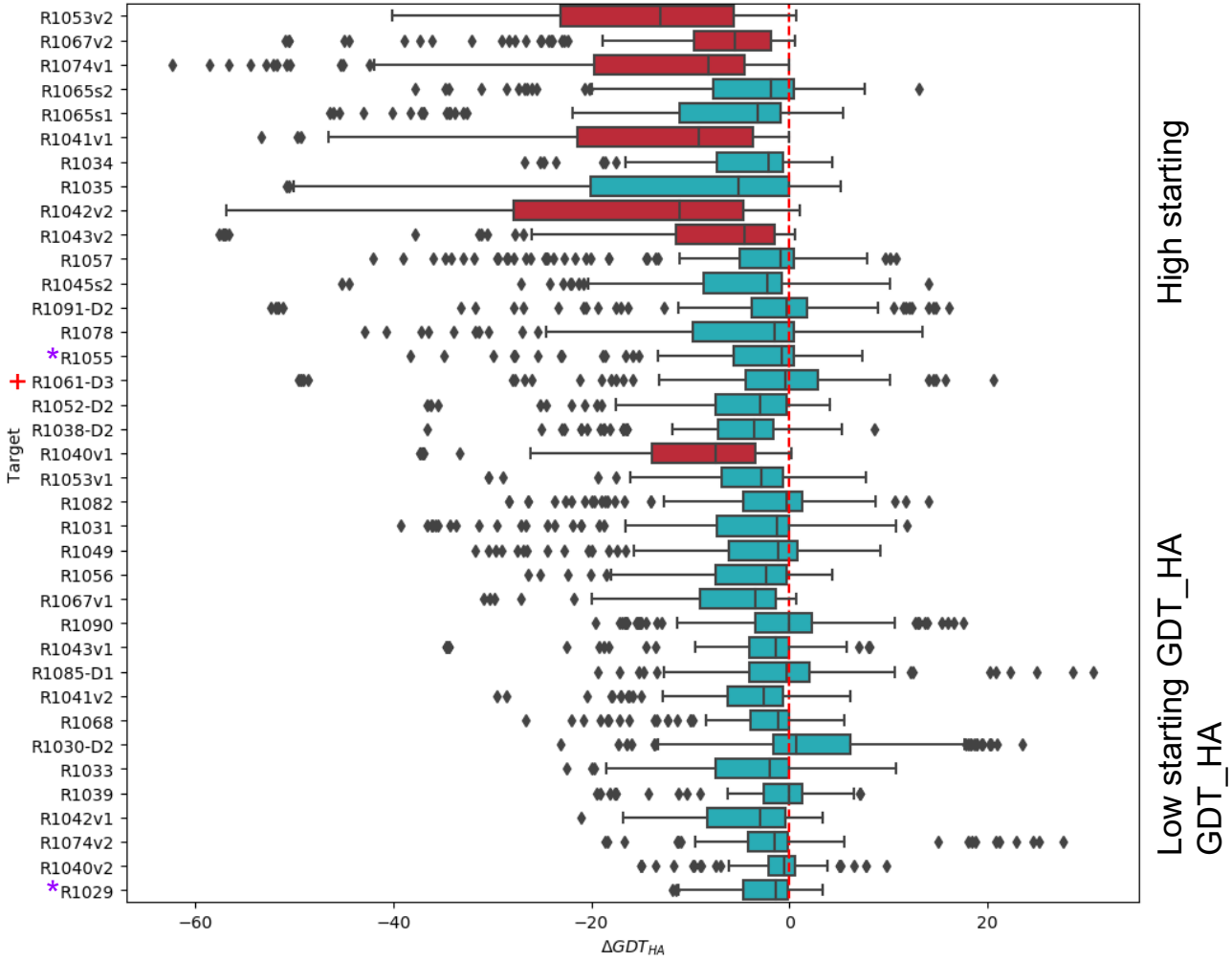


AlphaFold2 models have anomalously low refinability

AlphaFold2 refinement targets can barely and rarely be improved.

Other targets of similar quality **can** be refined

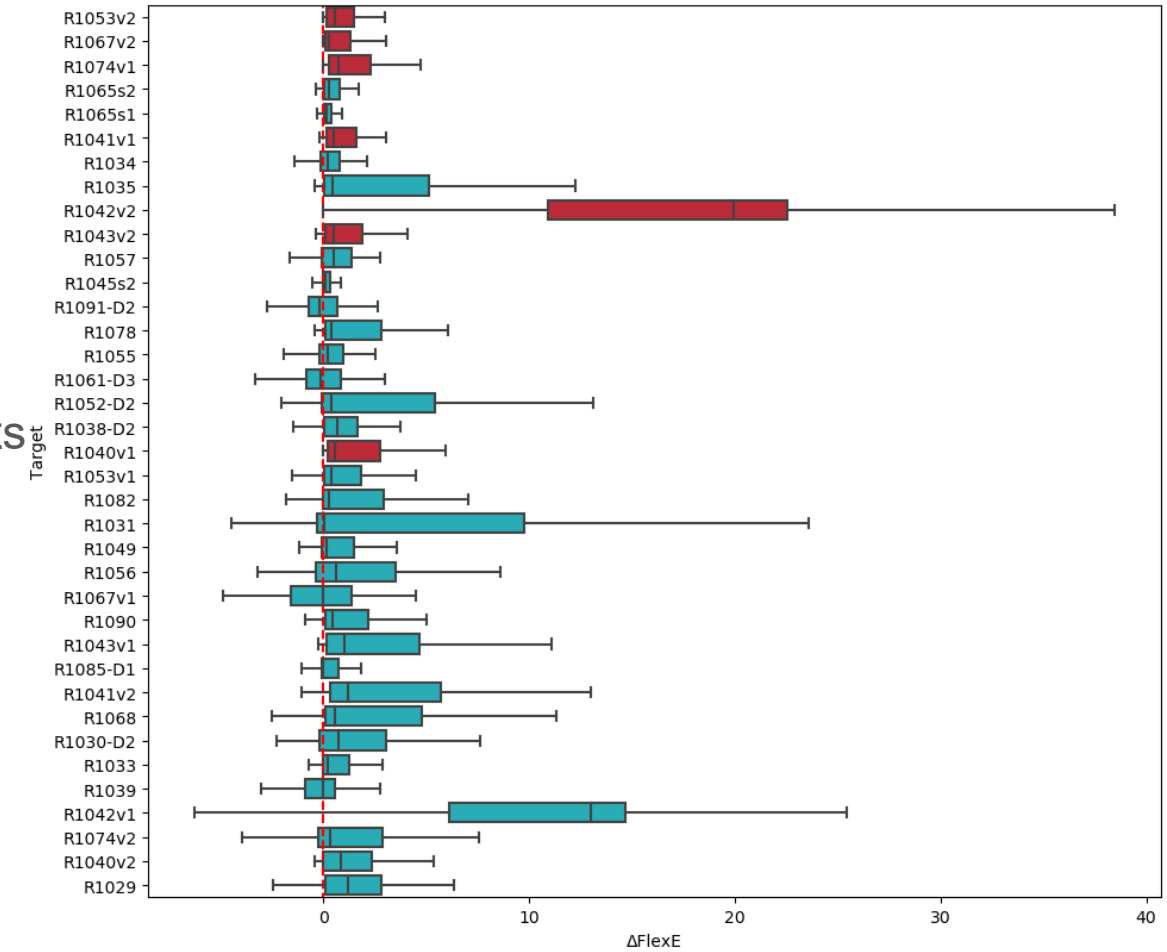
*=NMR +=CryoEM



AlphaFold2 models have anomalously low refinability

Same goes when AlphaFold2 and other targets are compared for FlexE

Measures energy of deformation between model and crystal structure. Somewhat orthogonal to coordinate accuracy

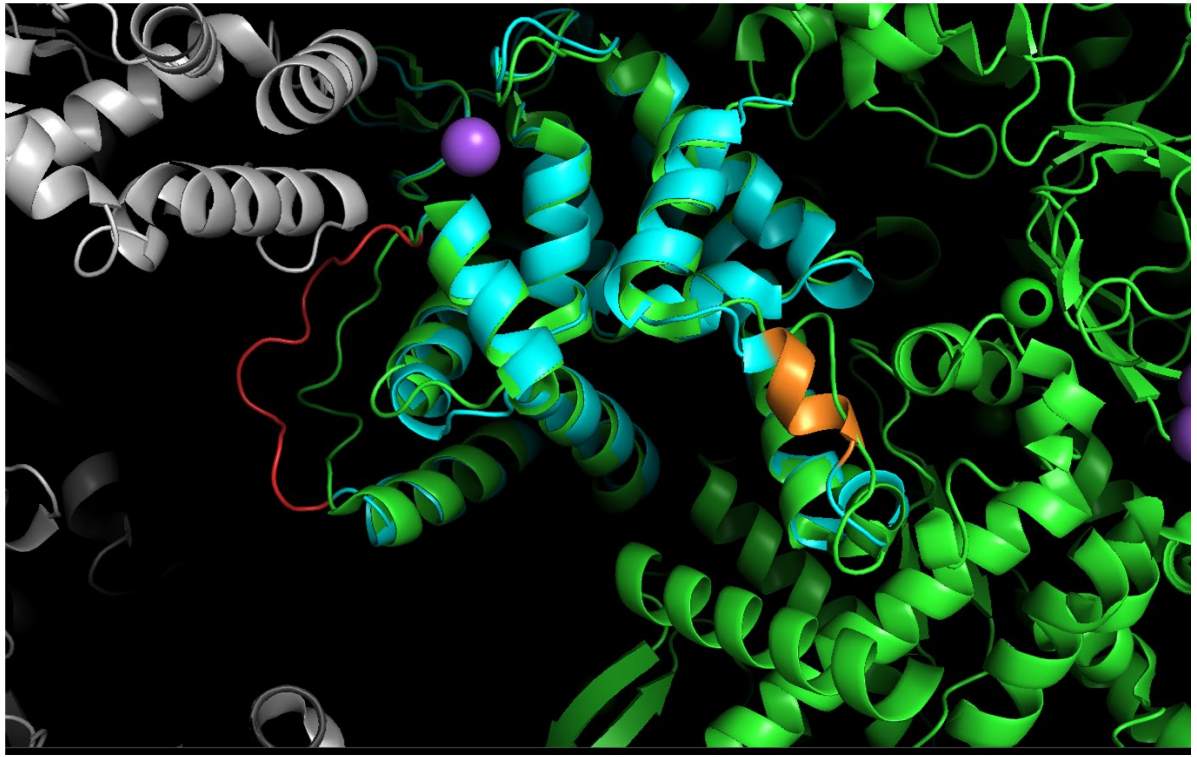


High starting GDT_HA

Low starting GDT_HA

Perez et al (2012) J. Chem. Theory Comput. 8, 3985

Most Alphafold2 ‘errors’ are at lattice contacts

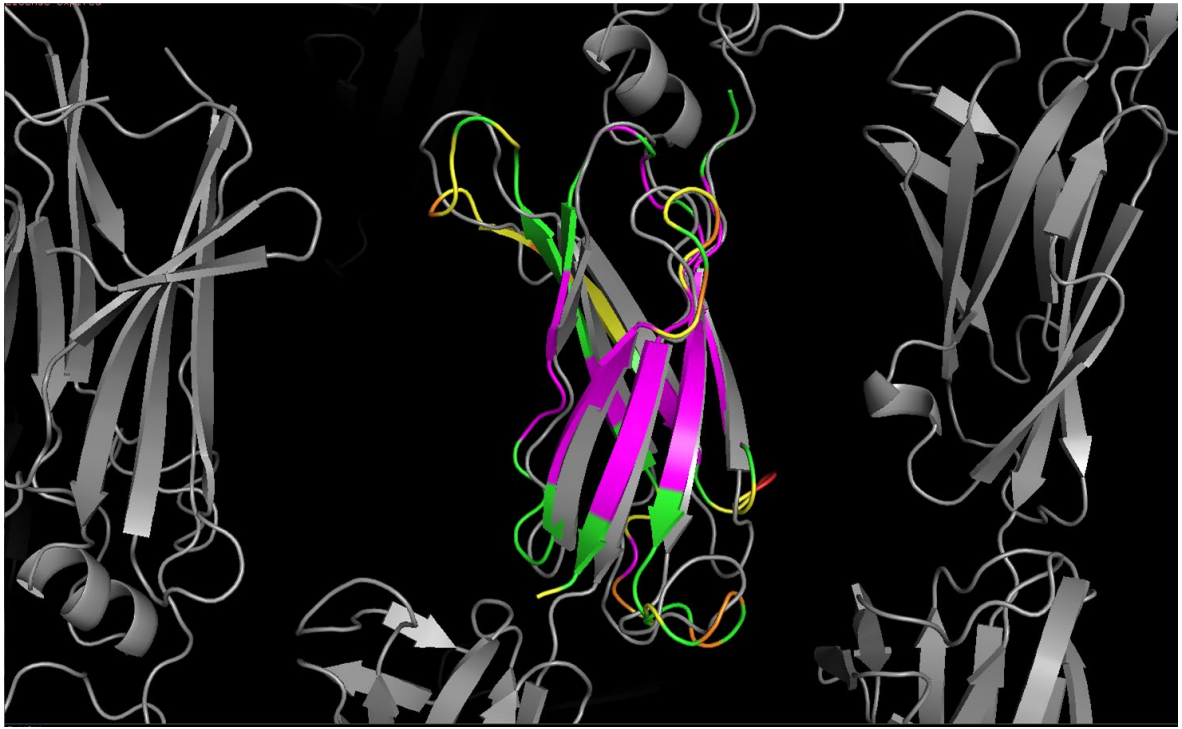


T1041 (GDT_HA = 70)

Most Alphafold2 ‘errors’ are at lattice contacts

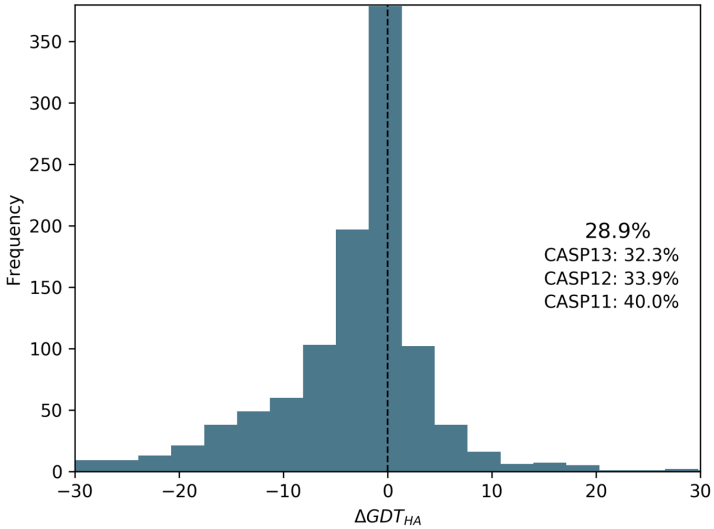
Target	***Errors near lattice contacts***	***Errors near domain contacts***	***Errors near chain contacts***	***Uncomplicated errors***
1040	1 (16 residues)			
1041	1 (12)	1 (5)		
1042	1 (6)	1 (6)		1 (3)
1043	3 (8,3,4)			
1053				1 (6)
1067	1 (20)			
1074	1 (6)			
Total regions	***8***	***2***		***2***
Total residues	***75***	***11***		***9***

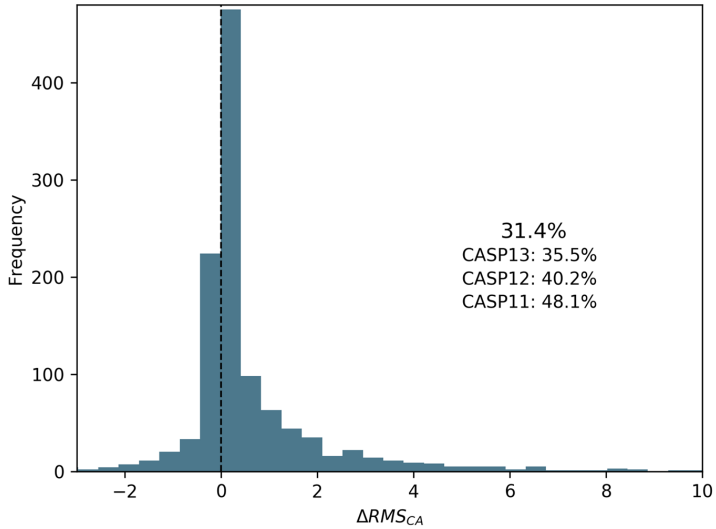
Other refinement targets contain refinable errors



T1091 (GDT_HA = 61)

CASP on CASP analysis without AlphaFold2





Excluding ‘unrefinable’ AlphaFold2 models improves these stats, but still comparable to previous years

Self-assessment of models and residues

Ability of groups to rank their predictions

Most groups have positive CC between assigned _1 to _5 ordering and actual order of quality

19/26 put best as _1 more than 20% of the time

Top four groups vary

	Only targets where 5 unique models submitted		
Group Name	# targets	Spearman CC	% correct model 1
Seok	31	0.57	35.48
BAKER-experimental	34	0.45	55.88
Spider	19	0.42	31.58
Frustration_Refine	28	0.41	35.71
FEIG-S	32	0.37	50.00
FEIG	33	0.32	30.30
DeepMUSICS	13	0.29	38.46
Kiharalab_Refine	33	0.24	33.33
laufer_ros	23	0.18	43.48
PerezLab_Gators	25	0.17	40.00
DellaCorteLab	27	0.13	11.11
Seok-server	22	0.10	18.18
BAKER	29	0.09	34.48
AWSEM_PCA	25	0.07	20.00
Bhattacharya	21	0.05	19.05
UNRES-template	27	0.05	18.52
Kiharalab	23	0.01	65.22
Bhattacharya-Server	28	0.00	25.00
UNRES	24	0.00	29.17
AILON	30	-0.07	13.33
AIR	19	-0.09	21.05
Beta	19	-0.11	5.26
MUFOLD	11	-0.18	18.18
McGuffin	6	-0.23	0.00
MULTICOM-CLUSTER	4	-0.38	25.00
Protein-blacksmith	29	-0.45	0.00
SHORTLE	0	Na	Na
Risoluto	0	Na	Na
Pharmulator	0	Na	Na
Seminoles	0	Na	Na
PerillaGroup	0	Na	Na

Ability of groups to estimate residue level errors

ASE

ASE (Accuracy Self Estimate) score is based on the accuracy of every residue position in the model reported in the temperature factor field.
The ASE score is calculated by formula:

$$ASE = 100.0*(1 - Mean(|S(tf_i/d_0)-S(d_i/d_0)|))$$

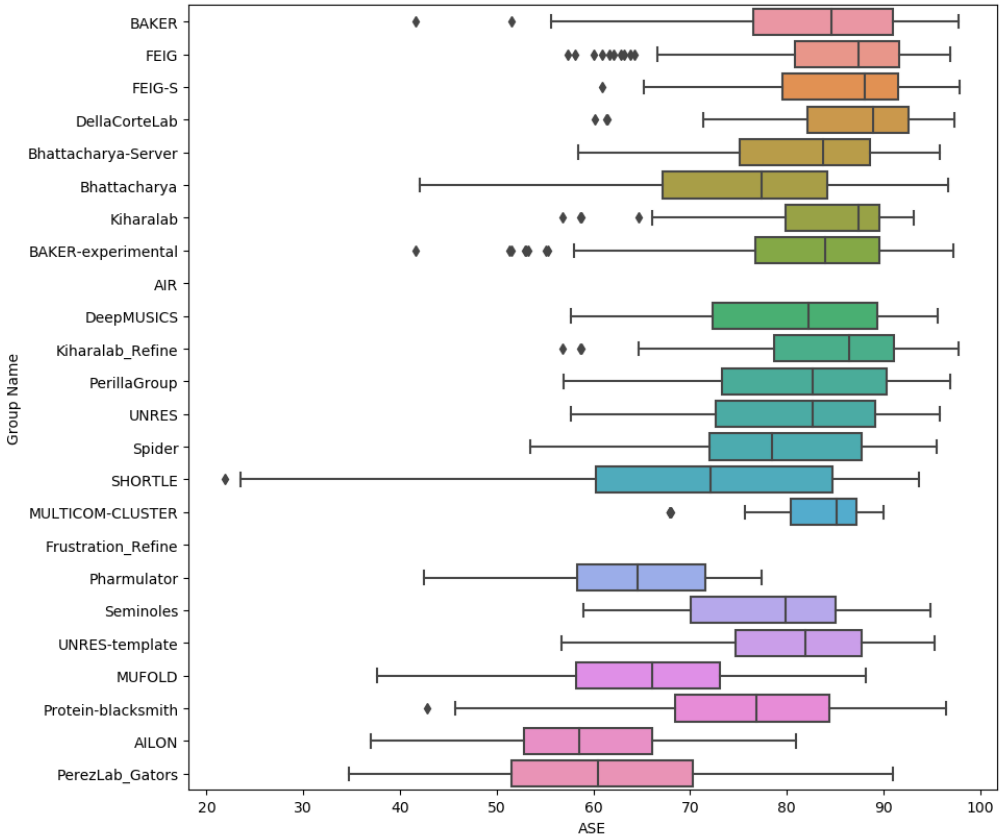
where tf_i - temperature factor of the i-th residue in the model
d_i - distance between i-th residues in lga alignment (sequence dependent mode)
S(x) = 1/(1+x^2) - S-function
d_0 - scaling factor, set d_0=5.0

Groups ordered by overall z-score ranking

No error estimates from two groups and excluded three groups since quality predictions looked backward

The best refinement groups are among the best in self-assessment of error too. Probably no coincidence…
Includes FEIG-S

Calculated on all submissions _1 to _5

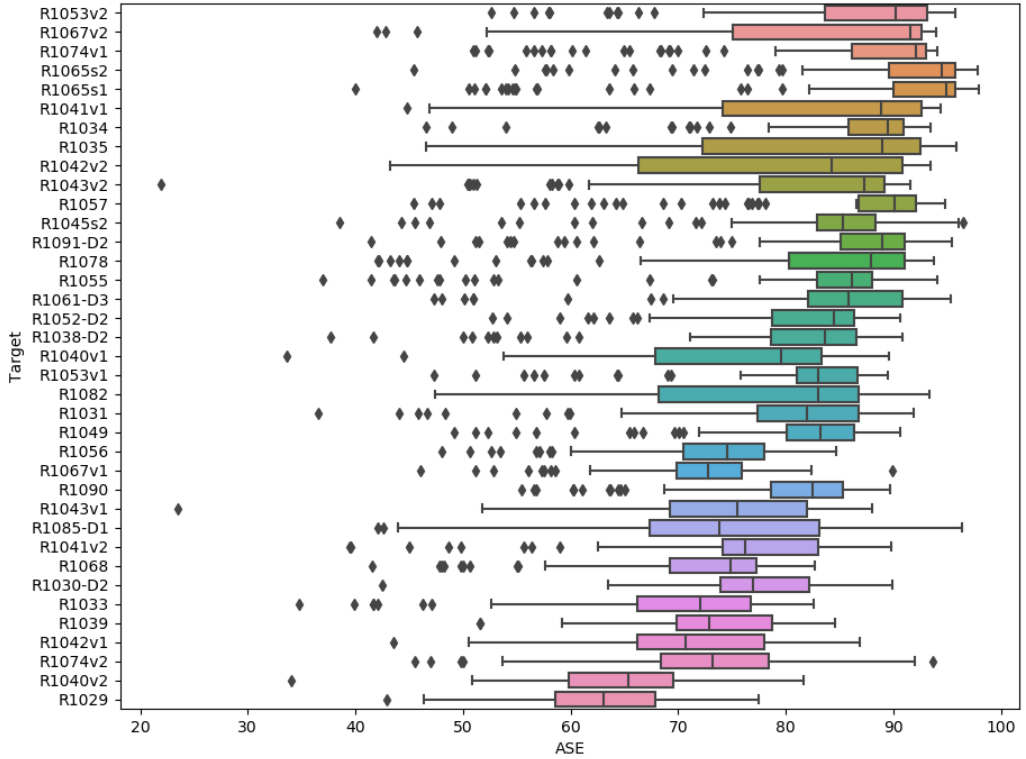


Some targets are harder than others

Ordered by starting GDT_HA

Calculated on all submissions _1 to _5

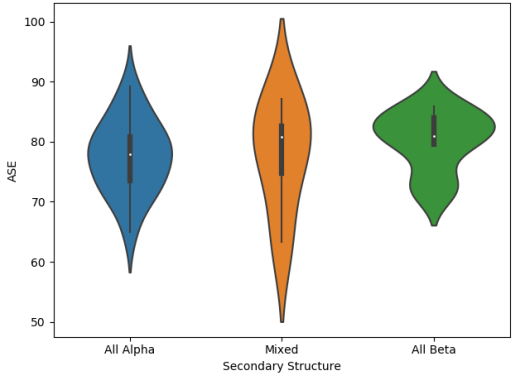
Generally harder to predict residue error on results from poorer quality refinement targets

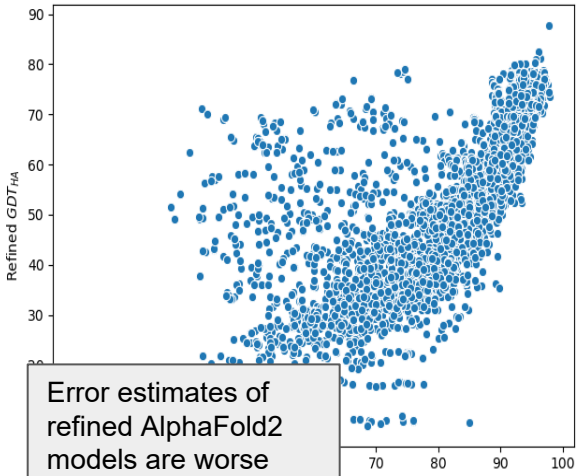


Factors (not) correlated with ASE

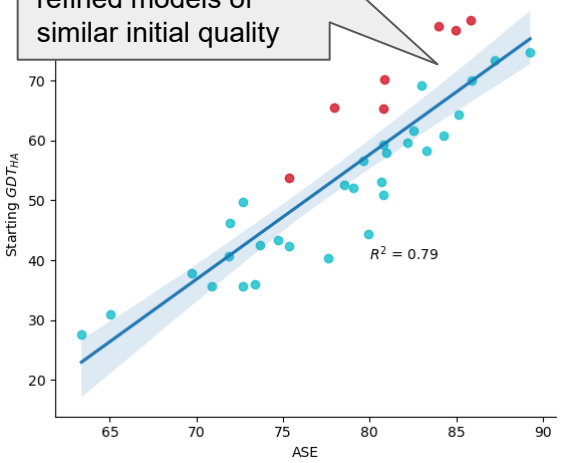
The better the model, the better the accuracy of per-residue error estimates

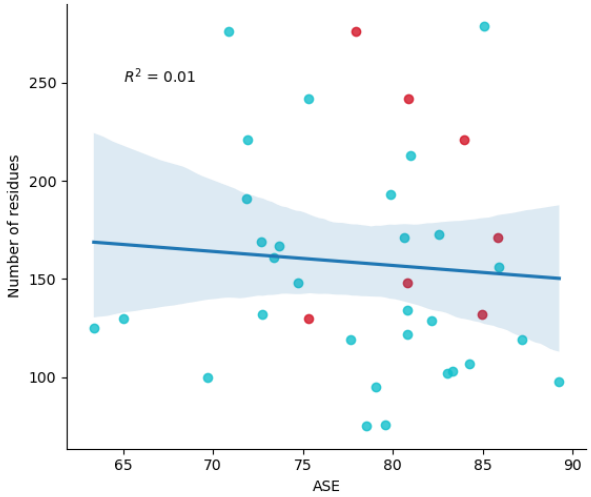
No relationship with size, regular structure or class

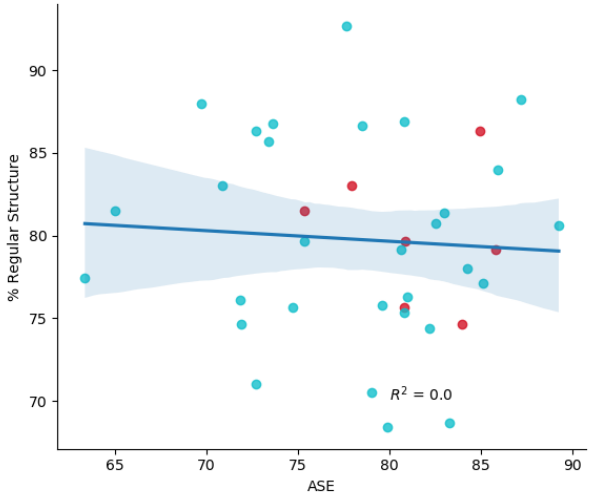




Error estimates of refined AlphaFold2 models are worse than those of other refined models of similar initial quality

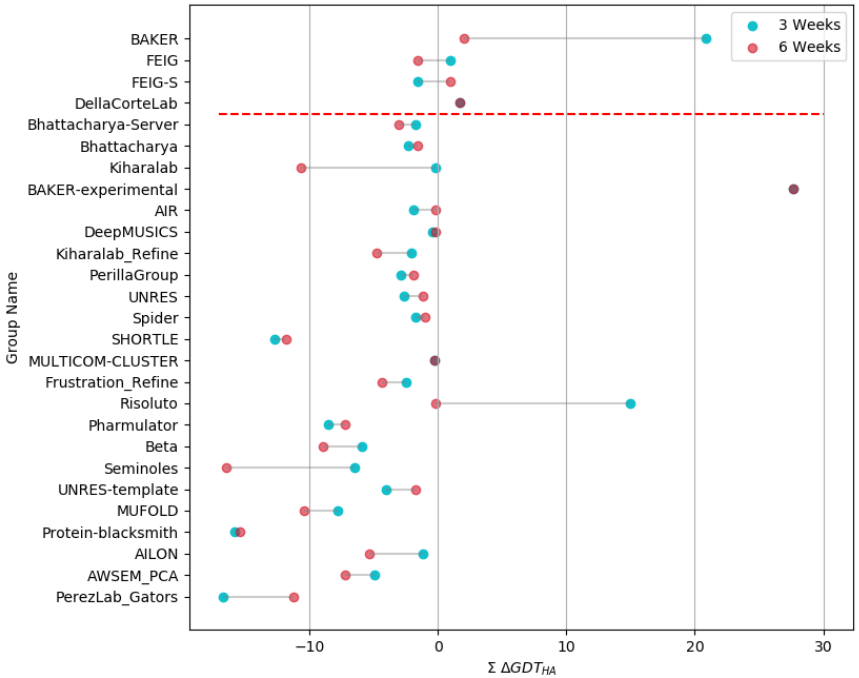


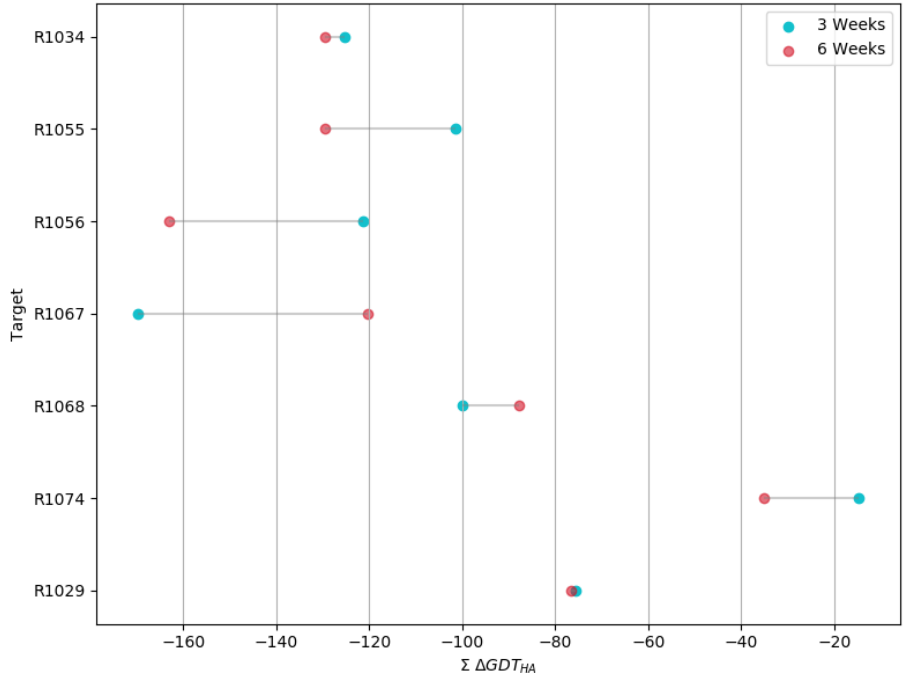




Special targets - extended and NMR

Extended targets



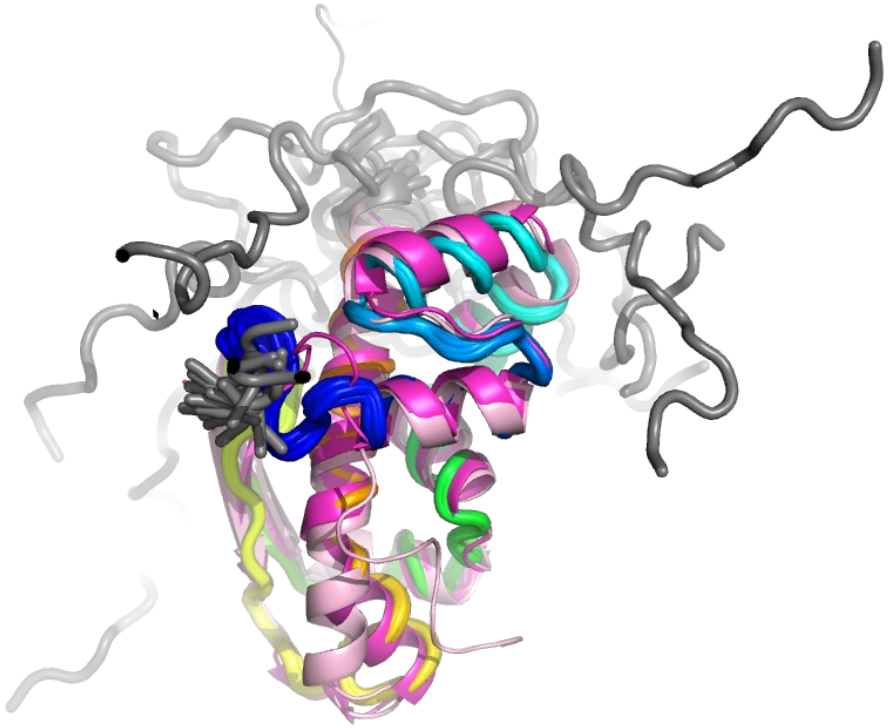


6 weeks results are worse than 3 weeks as often as they're better

Only best server FEIG-S, benefits overall among top four groups

NMR structures

These lack the complications eg crystal packing of other targets, and have extra information on multiple ‘correct’ conformations in the ensembles. Unfortunately, R1029 and R1055 ensembles were too tight to be very interesting

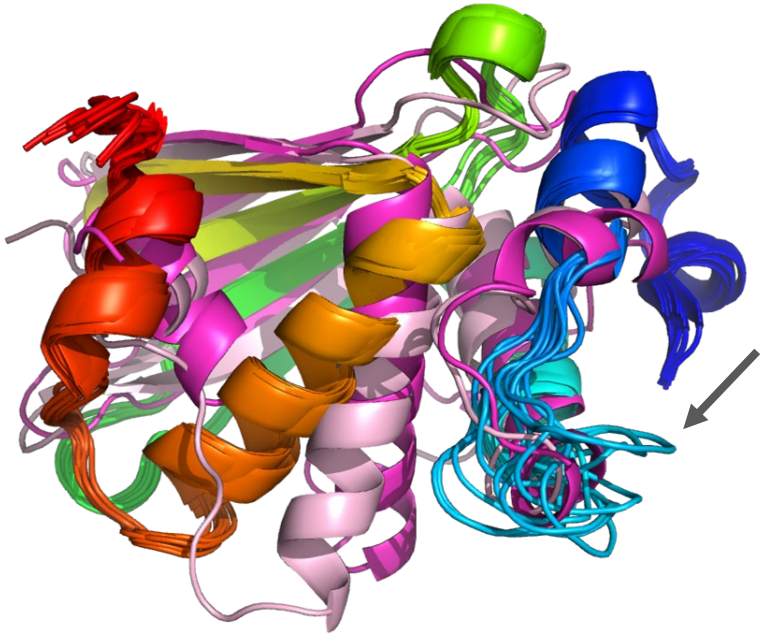


Refined region, but ensemble ~identical here

ensemble
unrefined
refined
not present in refined models

NMR structures

Unfortunately, R1029 and R1055 ensembles were too tight to be very interesting



8 residue loop where ensemble members vary. Very little change in structure on refinement

Major changed region, but not really improved and, in any case, ensemble ~identical here

ensemble
unrefined
refined
not present in refined models

Applications - Structure-based function prediction and Molecular Replacement

Structure-based function prediction

We wanted to assess whether refinement made a real-world difference to the ability to infer function from a structure. If crystal structure predicted a function, did refined versions out-perform the original refinement target?

Four enzymes, one double-barrelled

Catalytic site motifs sought using **ProFunc** and **Catsid**

Two DNA-binding proteins

Nucleic acid binding predicted with **DNA_BIND** and **BindNA**

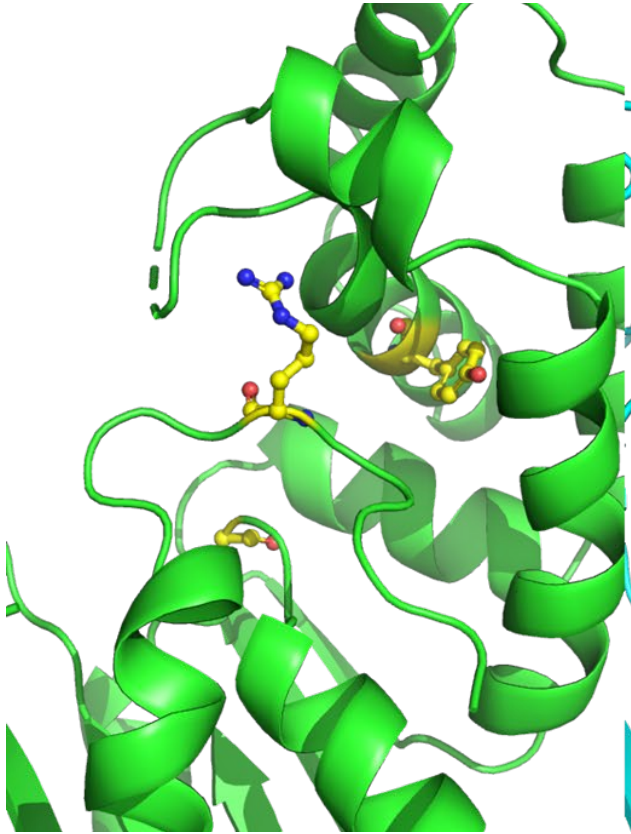
Three protein-protein interactions

Protein-protein docking done with **ClusPro**

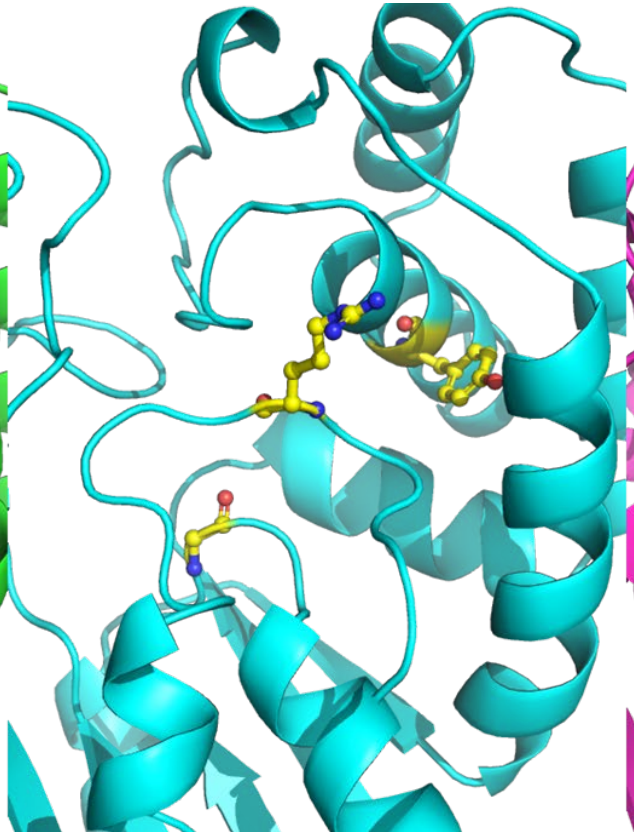
T1057 N4-cytosine methyltransferase

			CastID Methyltransferase hit score	Profunc Methyltransferase Active site template score
T1057	crystal		0.004	82.039
	unrefined		No hits	81.141
	13	refined 1	No hits	0
		refined 2	No hits	0
		refined 3	No hits	0
		refined 4	No hits	0
		refined 5	No hits	80.445
	323	refined 1	No hits	82.953
		refined 2	No hits	0
		refined 3	No hits	86.852
		refined 4	No hits	82.953
		refined 5	No hits	81.141
	335	refined 1	0.005	0
		refined 2	No hits	90.141
		refined 3	No hits	0
		refined 4	0.004	0
		refined 5	0.005	0
	473	refined 1	No hits	82.953
		refined 2	0.004	82.953
		refined 3	0.004	123.938
		refined 4	0.004	0
		refined 5	No hits	81.141

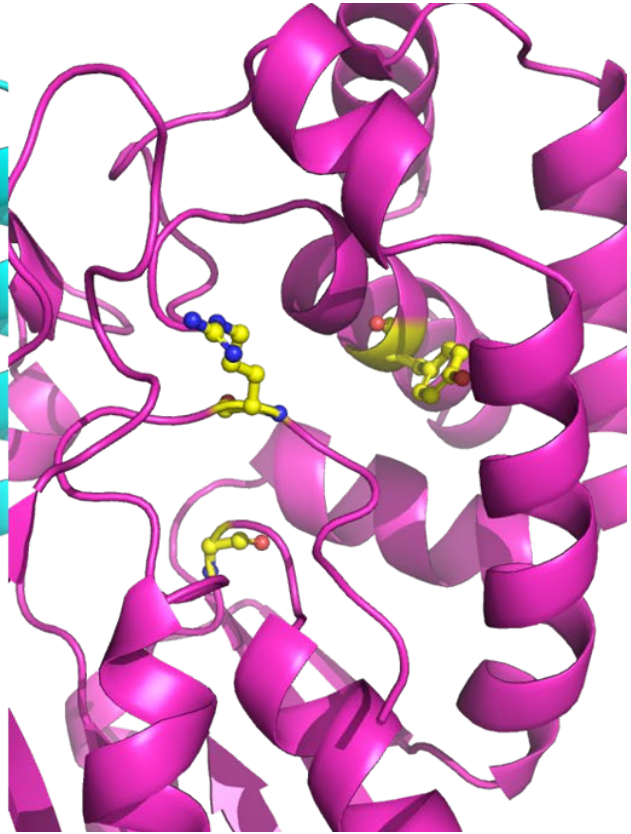
			DNAbind (0.5313)	bindup
T1057	crystal (original target processed)		YES 0.6628	YES
	unrefined		NO 0.4760	NO
	13	refined 1	0.5354	NO
		refined 2	0.4965	NO
		refined 3	0.5307	NO
		refined 4	0.535	NO
		refined 5	0.5266	NO
	323	refined 1	0.5148	NO
		refined 2	0.5089	NO
		refined 3	0.5285	NO
		refined 4	0.506	NO
		refined 5	0.5466	NO
	335	refined 1	0.068	NO
		refined 2	0.5245	NO
		refined 3	0.065	NO
		refined 4	0.4956	NO
		refined 5	0.483	NO
	473	refined 1	0.511	NO
		refined 2	0.4913	NO
		refined 3	0.4693	NO
		refined 4	0.4574	NO
		refined 5	0.5466	NO



T1057



T1057TS209_2
GDT_HA 64.4



R1057TS473_2
GDT_HA 67.8

ClusPro protein-protein interaction

T0145: Correct interface not identified, even from crystal structures.No helpful covariance. T1055: Was a crystal structure (5n2e) for partner and some site-directed mutagenesis on both sides. Nevertheless, plausible complex not found

T1065: Both partners were refinement targets

------ PPDbench -------

Receptor	Ligand	Cluster Size	Lowest Energy	Fnat	L_rms	I_rms	CAPRI Assesment	Eyeballing-Filo
T1065s1_Processed_Xtal	T1065s2_Processed_Xtal	126	-878.6	0.7	2.39	3.11	Medium	Identical
T1065s1_Processed_Xtal	T1065s2TS209_1	108	-616.1	0.7	4.91	4.84	Medium	Good
T1065s1_Processed_Xtal	R1065s2TS473_1	159	-713.2	0.08	31.61	29.25	Incorrect	Bad
T1065s1_Processed_Xtal	R1065s2TS335_1	169	-743.4	0.09	24.93	23.89	Incorrect	Bad
T1065s1_Processed_Xtal	R1065s2TS013_1	80	-676.8	0.1	26.36	25.08	Incorrect	Bad
T1065s1_Processed_Xtal	R1065s2TS323_1	173	-752.9	0.07	32.76	30.62	Incorrect	Bad
T1065s1_Processed_Xtal	R1065s2TS149_1	114	-655.9	0.75	3.88	3.47	Medium	Good
T1065s1TS351_4	T1065s2TS209_1	215	-640.4	0.47	9.61	9.1	Acceptable	Good
T1065s1TS351_4	T1065s2_Processed_Xtal	119	-574.2	0.13	24.37	22.79	Incorrect	Bad
R1065s1TS473_1	T1065s2_Processed_Xtal	99	-591.9	0.09	31.59	28.41	Incorrect	Bad
R1065s1TS335_1	T1065s2_Processed_Xtal	152	-629.1	0.84	1.82	1.89	Medium	Good
R1065s1TS013_1	T1065s2_Processed_Xtal	150	-630.5	0.84	3.78	3.08	Medium	Good
R1065s1TS323_1	T1065s2_Processed_Xtal	108	-535.5	0.62	22.83	6.06	Medium	OKish
R1065s1TS149_1	T1065s2_Processed_Xtal	105	-551.2	0.86	2.88	2.91	Medium	Good
R1065s1TS473_1	R1065s2TS473_1	113	-650.3	0.28	15.84	13.79	Incorrect	OKish
R1065s1TS335_1	R1065s2TS335_1	146	-590.9	0.1	25.4	24.04	Incorrect	OKish
R1065s1TS013_1	R1065s2TS013_1	127	-623	0.1	34.52	31.73	Incorrect	Bad
R1065s1TS323_1	R1065s2TS323_1	123	-627.4	0.08	26.91	25.69	Incorrect	Bad
R1065s1TS149_1	R1065s2TS149_1	235	-727.9	0.43	7.43	6.63	Acceptable	Good

Refinement of s2 tends to degrade results

Original models give decent result

But refinement of s1 tends to improve results

Refining both leans towards bad

Molecular replacement

Randy Read's scripts used to measure Phaser LLG of placed and rigid-body refined xtal/model, given rest of asu (thanks Marcus, Joana!)

"The LLG is the difference between the likelihood of the model and the likelihood calculated from a Wilson distribution, so it measures how much better the data can be predicted with your model than with a random distribution of the same atoms."

These can use (i) a constant B factor, (ii) interpret predicted residue error as B-factor or (iii) apply per-residue B factor based on the supplied per-residue error estimates

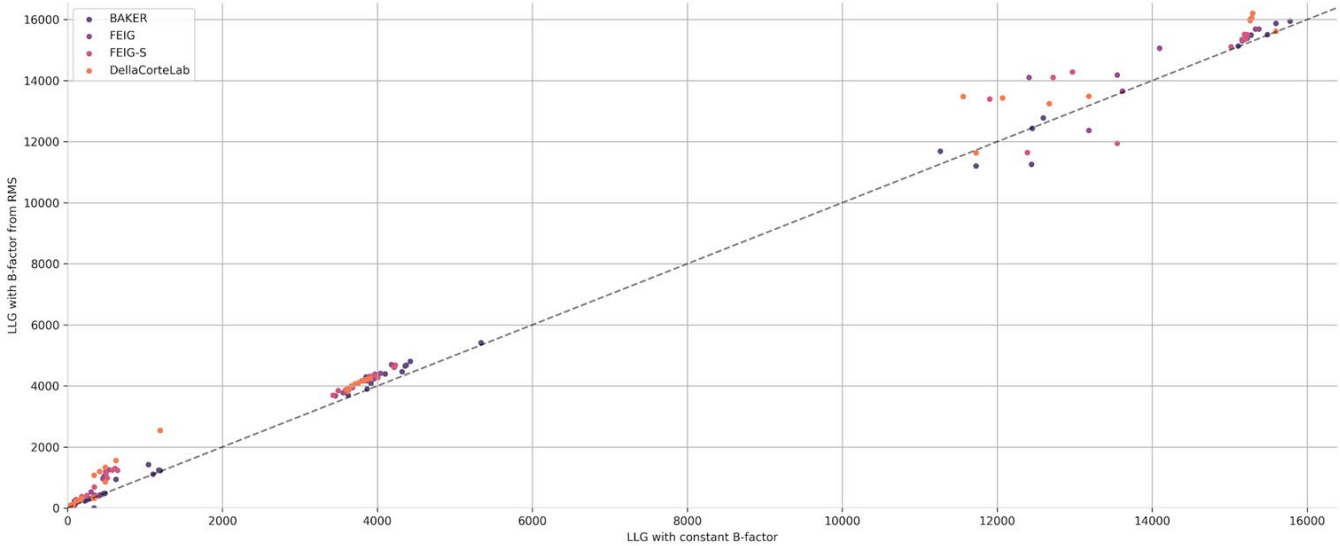
Low LLG scores, improved on refinement, assessed for real world impact

Phaser or Molrep used for MR via MrBUMP

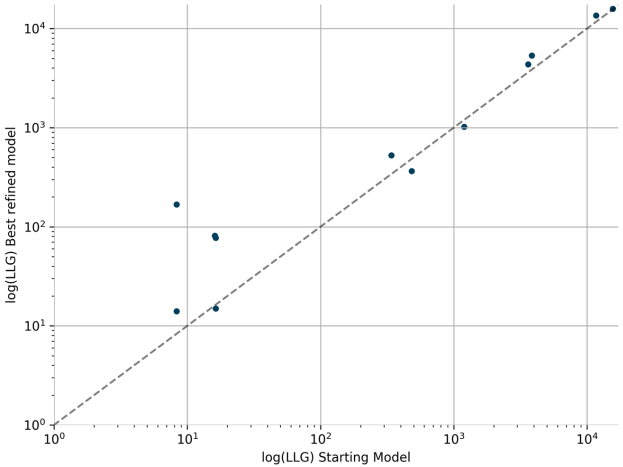
Correctness of placement checked by local and global map CC calculation in Phenix

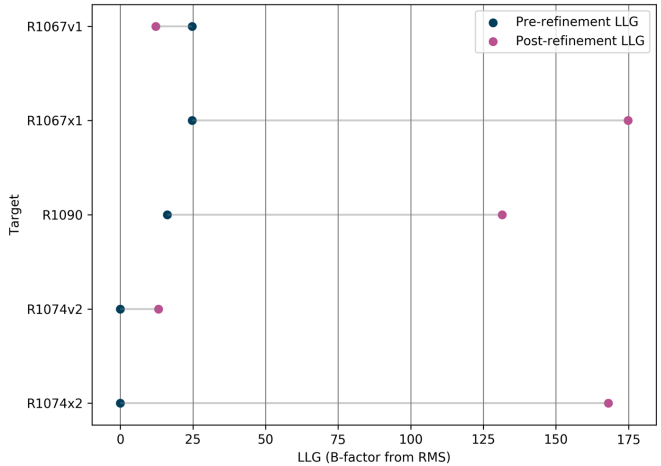
Constant B-factor vs B-factor from error estimate

Looking at the top 4 groups, the calculated LLG is consistently better when using the B-factors from error estimate



Refinement tends to improve best available LLG



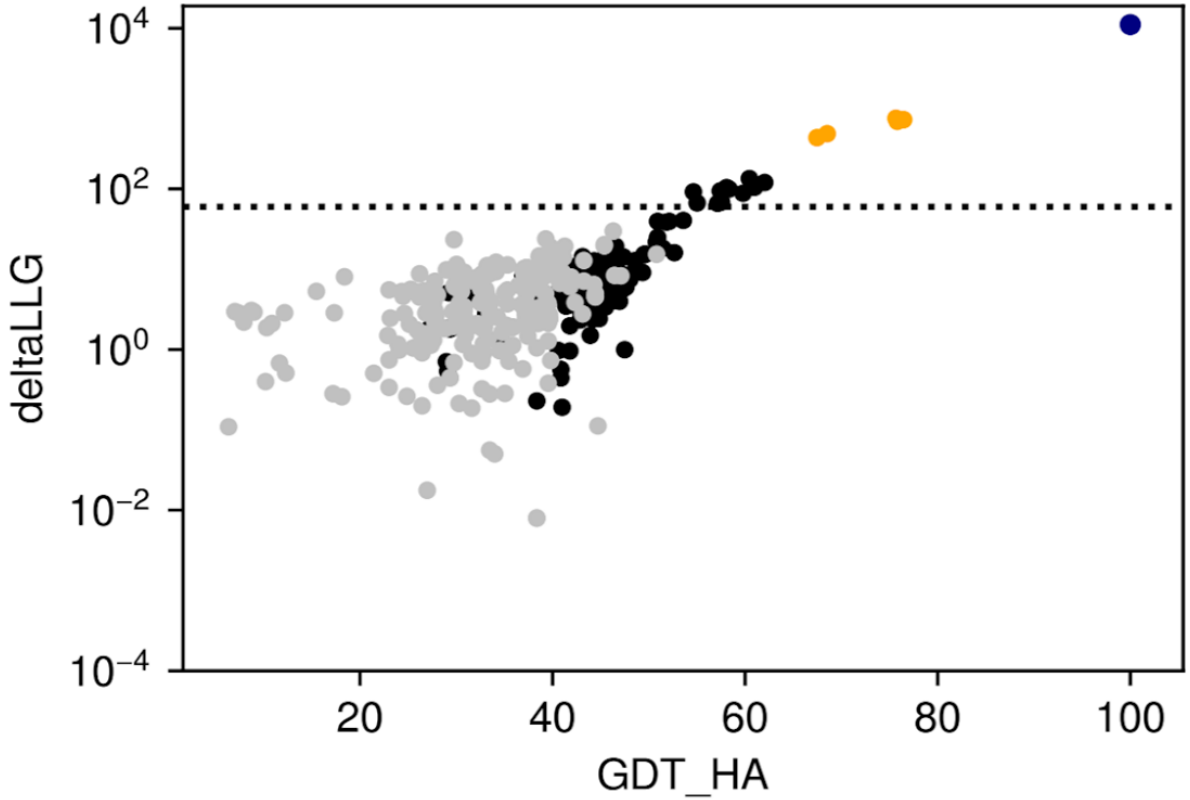


Best predictions across all groups for each target. For all but 3 the best refined model had a better LLG (B from predicted errors) than the starting model

Took a closer look at the starting models with an LLG < 120. Several large improvements

Refinement often takes predictions over LLG>60 threshold

T1090



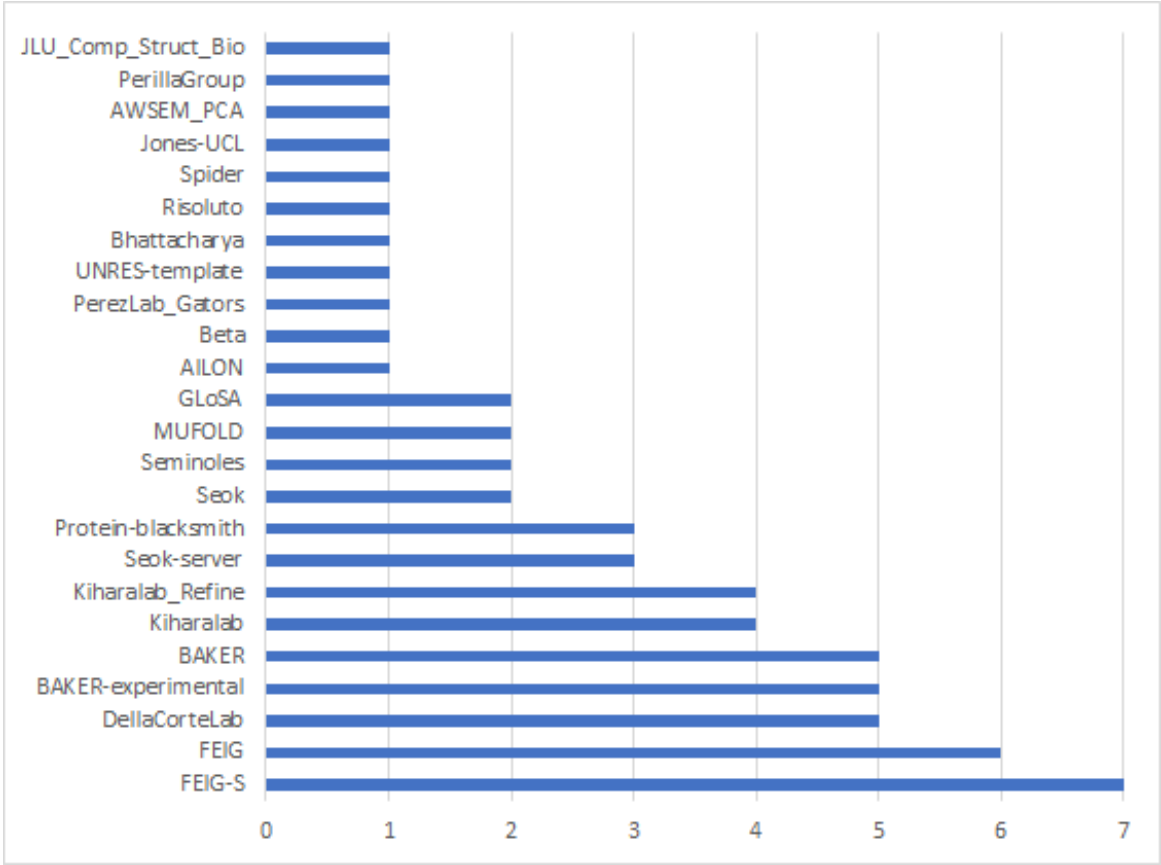
crystal structure
AlphaFold2 models
other models
refinements

MR and refinement of non-AlphaFold2 targets

Target	GDT_HA	mol/ asu	Res. (Å)	Refinement target solves?	Refined version solves?	Best refined version LLG	Best map CC	Phaser solutions	Molrep solutions	Number of groups producing successful refinements
T1030	40	1	3.03	no	yes	63	0.567	3	0	1
T1034	70	4	2.057	yes	yes	868	0.572	196	177	35
T1038	57	3	2.5	no	no	20	0.126	0	0	0
T1049	51	1	1.75	no	yes	53	0.275	10	4	6
T1052	58	1	1.976	no	yes	15	0.399	2	0	1
T1053	53	4	3.294	no	no	56	0.059	0	0	0
T1056	50	1	2.3	no	yes	58	0.416	11	38	14
T1067	46	1	1.44	no	yes	67	0.418	8	15	11
T1074	36	1	1.5	no	yes	132	0.501	15	18	7
T1082	53	3	1.147	no	no	45	0.06	0	0	0
T1085	43	1	2.491	no	yes	40	0.57	4	2	3
T1090	44	1	1.77	no	yes	83	0.4	12	14	7
T1091	61	1	2.994	no	no	-	0.079	0	0	0

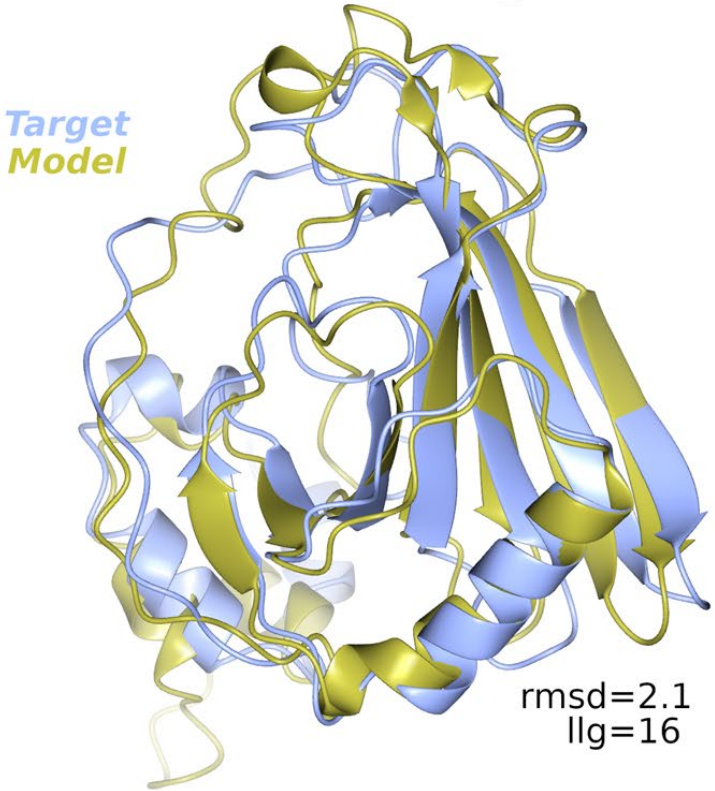
Many groups make some targets succeed in MR

FEIG-S refinement makes most targets succeed in MR



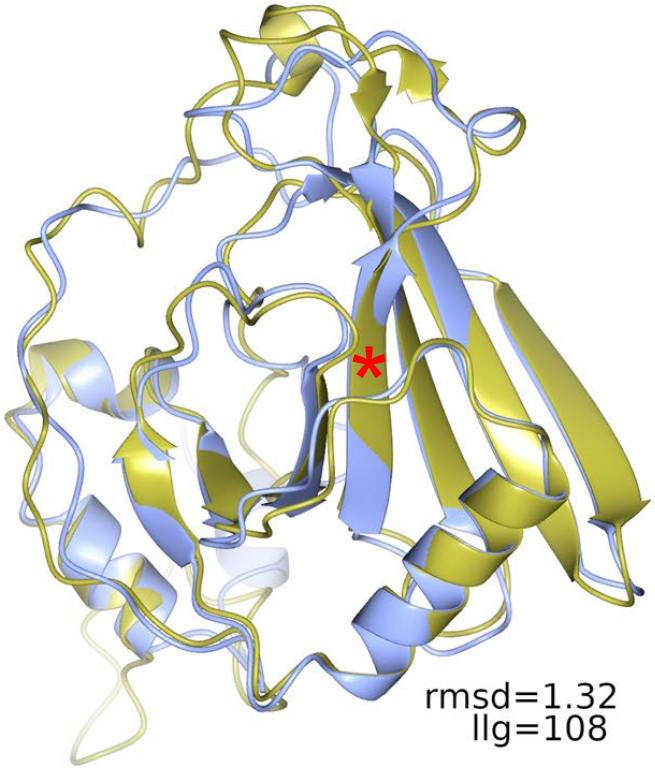
MR refinement example

Unrefined (T1090TS351_1.pdb)



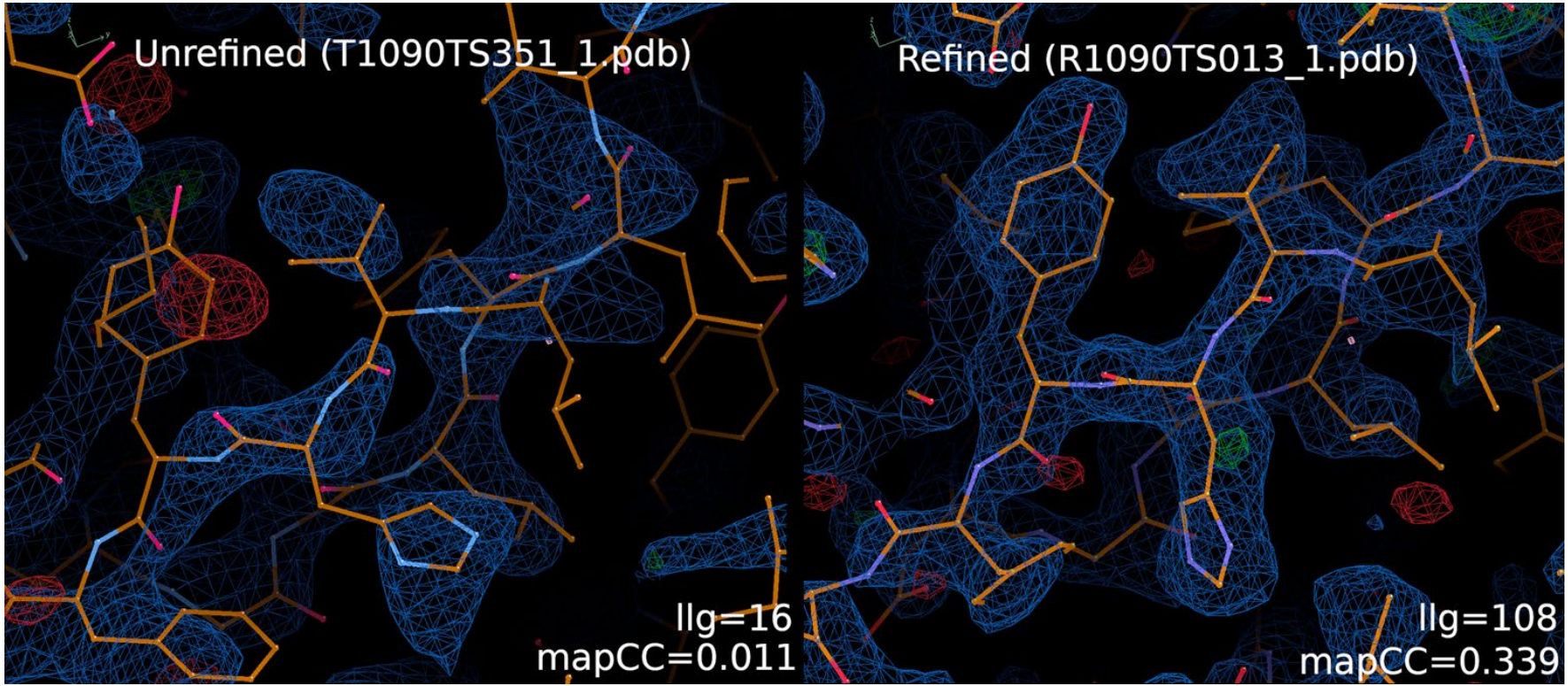
Original prediction (GDT_HA 44), cannot be placed by MR, here superposed artificially

Refined (R1090TS013_1.pdb)



FEIG-S-refined (GDT_HA 60), as placed by MR

MR refinement example



Original prediction, cannot be placed by MR, here superposed artificially

FEIG-S-refined, as placed by MR

Conclusions

vs previous CASPs, similar improvements on similar quality refinement targets. Suggests performance maintained but not really improved

Best groups have quite distinct approaches. MD-centred approaches more conservative, smaller range from best to worse results. Best for small proteins. Rosetta-based methods more of a gamble - bigger potential gains **and** losses. Can improve largest targets

Mixed results with extended targets

Structure-based function prediction results show small differences but often in the right direction

Refinement makes a big real-world difference to MR with poor models. Accurate residue error estimates further help