CASP 13

Chemical crosslink assisted modeling

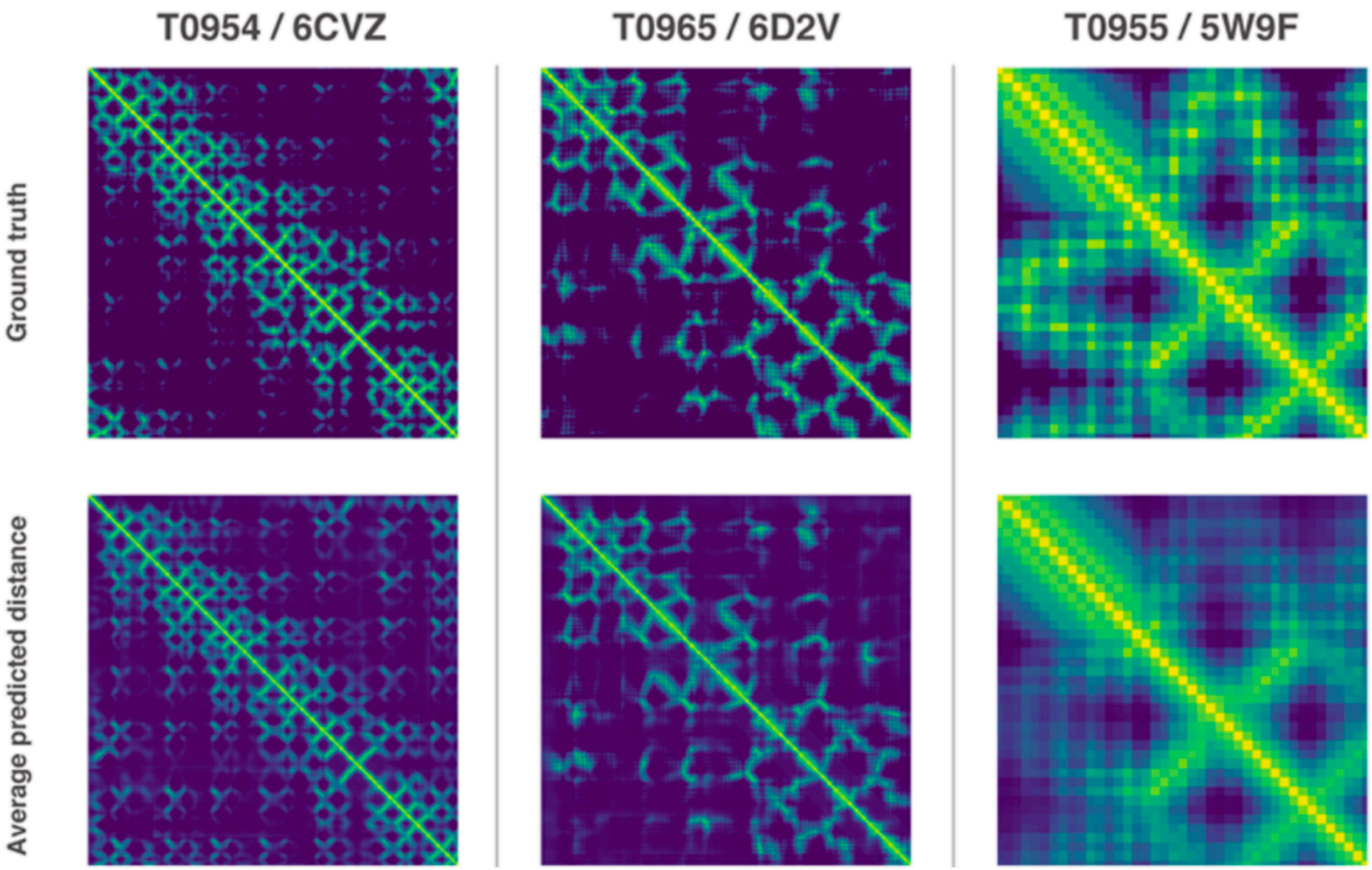
Albert Einstein College of Medicine

Assessor: András Fiser
Department of Systems and Computational Biology
Department of Biochemistry

DeepMind claims early progress in AI-based predictive protein modelling

Natasha Lomas @riptari / 4 hours ago

Comment



specialist, **DeepMind,** has claimed a "significant

Google's DeepMind predicts 3D shapes of proteins

AI program's understanding of proteins could usher in new era of medical progress

most viewed in US

David Attenborough: collapse of civilisation is on the horizon

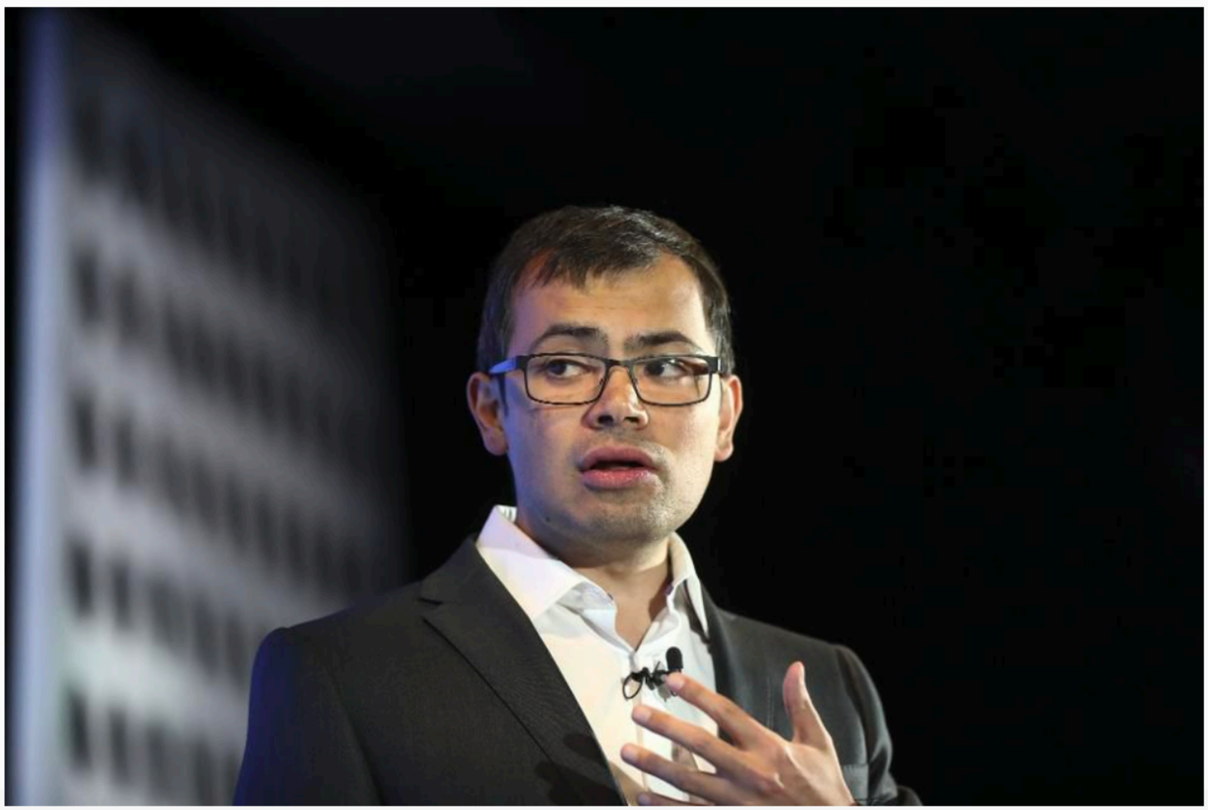
Skin in the game: is live artistic nudity more than titillation?

Trump absent again as Kennedy Center Honors pay tribute to Bush

Trump takes on General Motors (guess who wins?) *Robert Reich*

DeepMind Starts To Show How AI Can Be Used To Solve Scientific Problems

Sam Shead Contributor
AI & Big Data
I cover artificial intelligence and Google DeepMind.



AI

Deepmind's AlphaFold wins CASP13 protein-folding competition

KYLE WIGGERS @KYLE_L_WIGGERS DECEMBER 3, 2018 7:20 AM

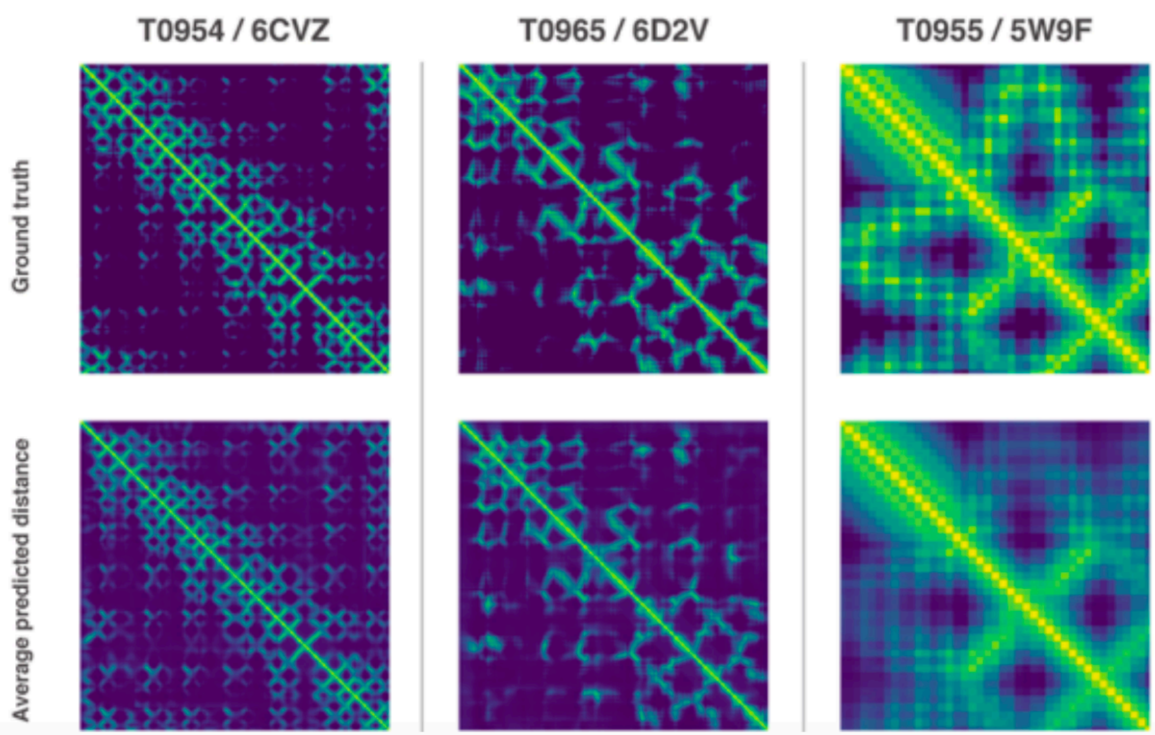


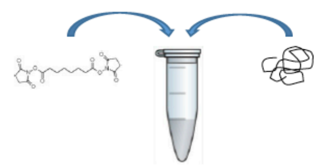
Image Credit: Deepmind

Background 1/4

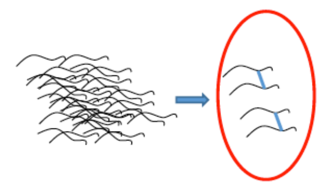
Experimental workflow of a cross-linking experiment

The workflow resembles a conventional proteomics experiment, with some modifications

Sample preparation and cross-linking reaction



Sample work-up: enzymatic digestion, clean-up, enrichment/fractionation (optional)



Cross-linked peptides that need to be identified

LC-MS/MS analysis

Data analysis using specialized software

Different products from a XL experiment

Dead-end link (monolink) *type 0 link*

Loop link (intrapeptide link) *type 1 link*

Intraprotein

Interprotein

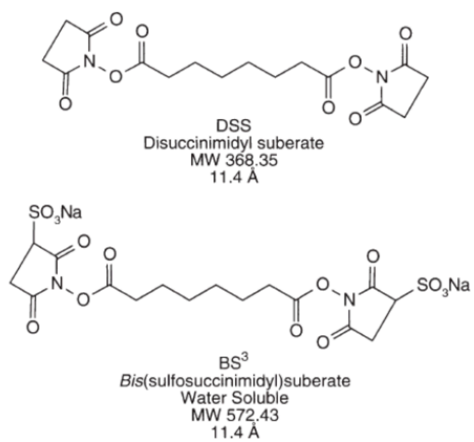
Cross-link *type 2 link*

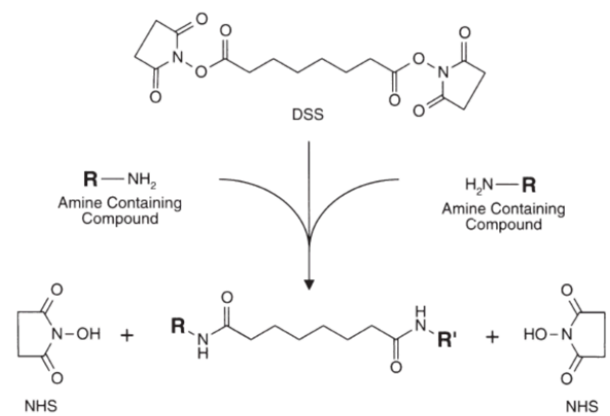
Background 2/4

Cross-linking chemistries

Cross-linking of primary amines (Lys, N-terminus) using succinimide esters, e.g. DSS, BS^3

- Most widely used chemistry in XL-MS
- Side-reactions with Ser/Thr/Tyr possible





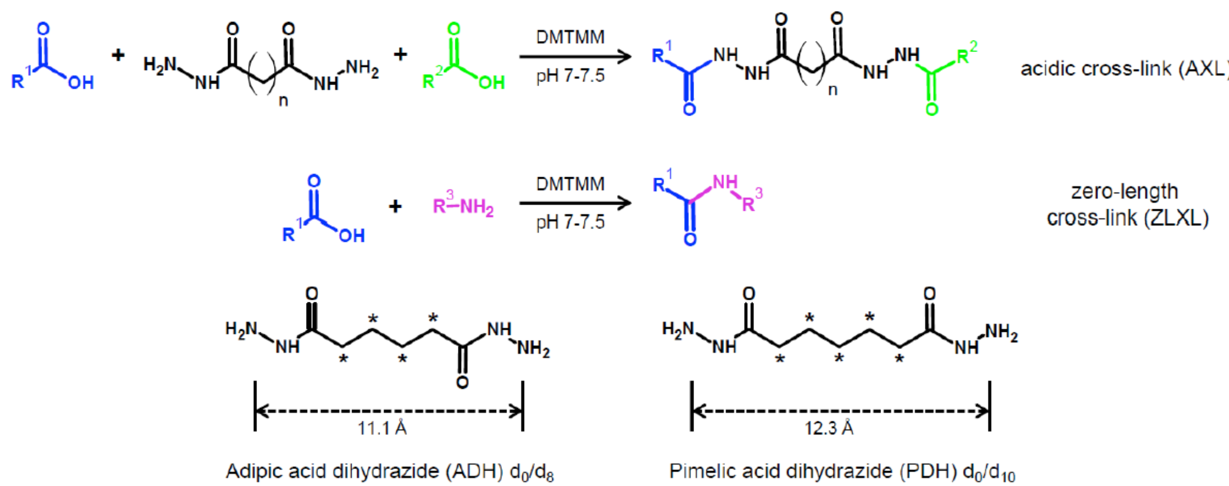
Hermanson, *Bioconjugate Techniques*

Background 3/4

Cross-linking chemistries

Cross-linking of carboxyl groups (Asp, Glu, C-terminus) and of primary amines with carboxyl groups (without spacer)

- Combined reaction will yield two different reaction products
- Lower reaction yields, success depends more on target protein (complex)



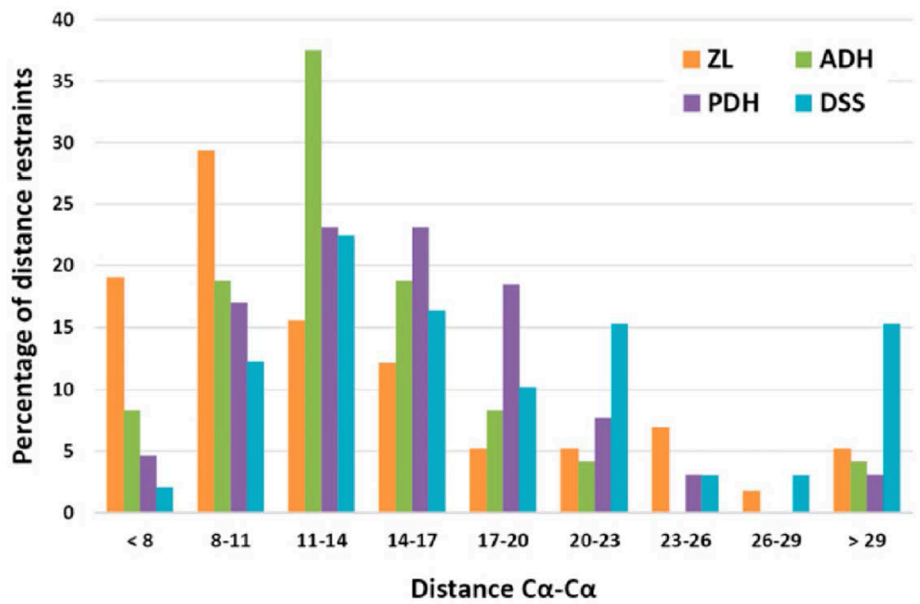
Leitner et al., PNAS, 2014

Background 4/4

How to calculate actual distance restraints?

Practically, larger distances are observed, e.g. up to 30 Å and more (for proteins with known 3D structure)

Note that ZL cross-links bridge shorter distances, but by only approx. 5 Å!



Leitner et al., PNAS, 2014

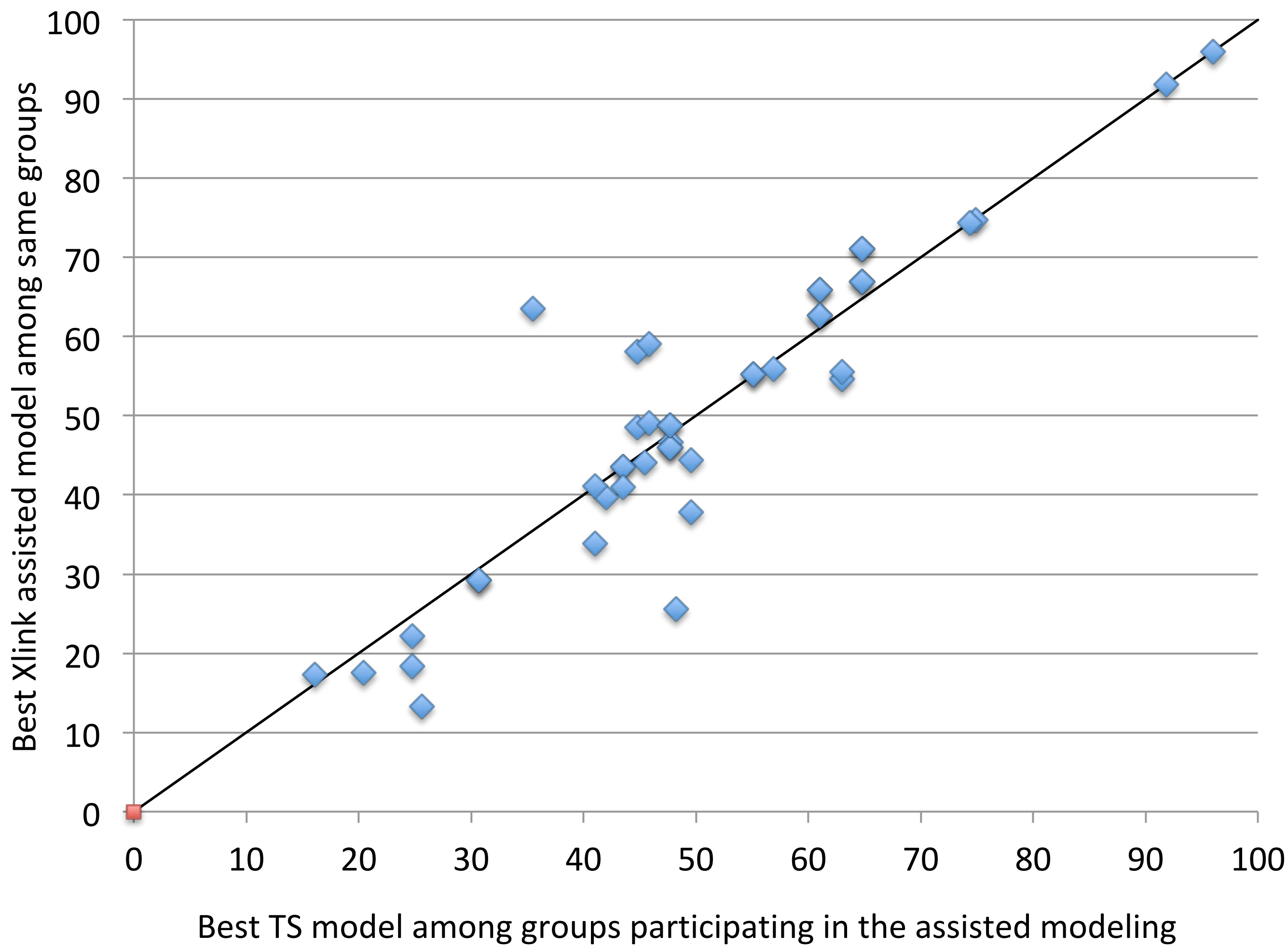
Groups, targets, performance

- Targets: 29 domains/subunits + full complexes in total.
- 14 groups predicted between 3 and 27 targets out of 29
- Number of groups that provided models both with and without Xlink ranges: 3-6 per target

14 groups in total
(but only 6 submitted more than 20 predictions)

#	GR code	GR name	Domains Count	SUM Zscore (>-2.0)	Rank SUM Zscore (>-2.0)	AVG Zscore (>-2.0)	Rank AVG Zscore (>-2.0)	SUM Zscore (>0.0)	Rank SUM Zscore (>0.0)	AVG Zscore (>0.0)	Rank AVG Zscore (>0.0)
1.	105	-	27	10.0907	3	-0.3670	10	7.7168	5	0.2858	9
2.	208	-	26	31.1427	1	0.5055	8	13.5818	3	0.5224	8
3.	492	-	26	3.5963	4	-0.5540	11	0.9136	11	0.0351	11
4.	288	-	26	-1.2892	6	-0.7419	12	0.9080	12	0.0349	12
5.	196	-	23	25.7938	2	0.5997	7	15.2116	1	0.6614	6
6.	122	-	21	-9.6615	9	-0.8410	14	0.5377	13	0.0256	13
7.	000	-	12	1.5362	5	0.9614	3	13.7047	2	1.1421	1
8.	135	-	9	-6.9519	7	1.0053	2	9.0481	4	1.0053	3
9.	359	-	9	-8.5049	8	0.8328	4	7.4951	6	0.8328	4
10.	117	-	8	-16.9087	10	0.1364	9	2.1001	9	0.2625	10
11.	329	-	7	-25.6654	13	-0.8093	13	0.1439	14	0.0206	14
12.	207	-	5	-18.5062	11	1.0988	1	5.4938	7	1.0988	2
13.	364	-	5	-20.5567	12	0.6887	5	3.4433	8	0.6887	5
14.	271	-	3	-26.1149	14	0.6284	6	1.8851	10	0.6284	7

Single best 3D model vs single best assisted model for each single chain target within group of assisted predictors only. (Groups do not need to match!)



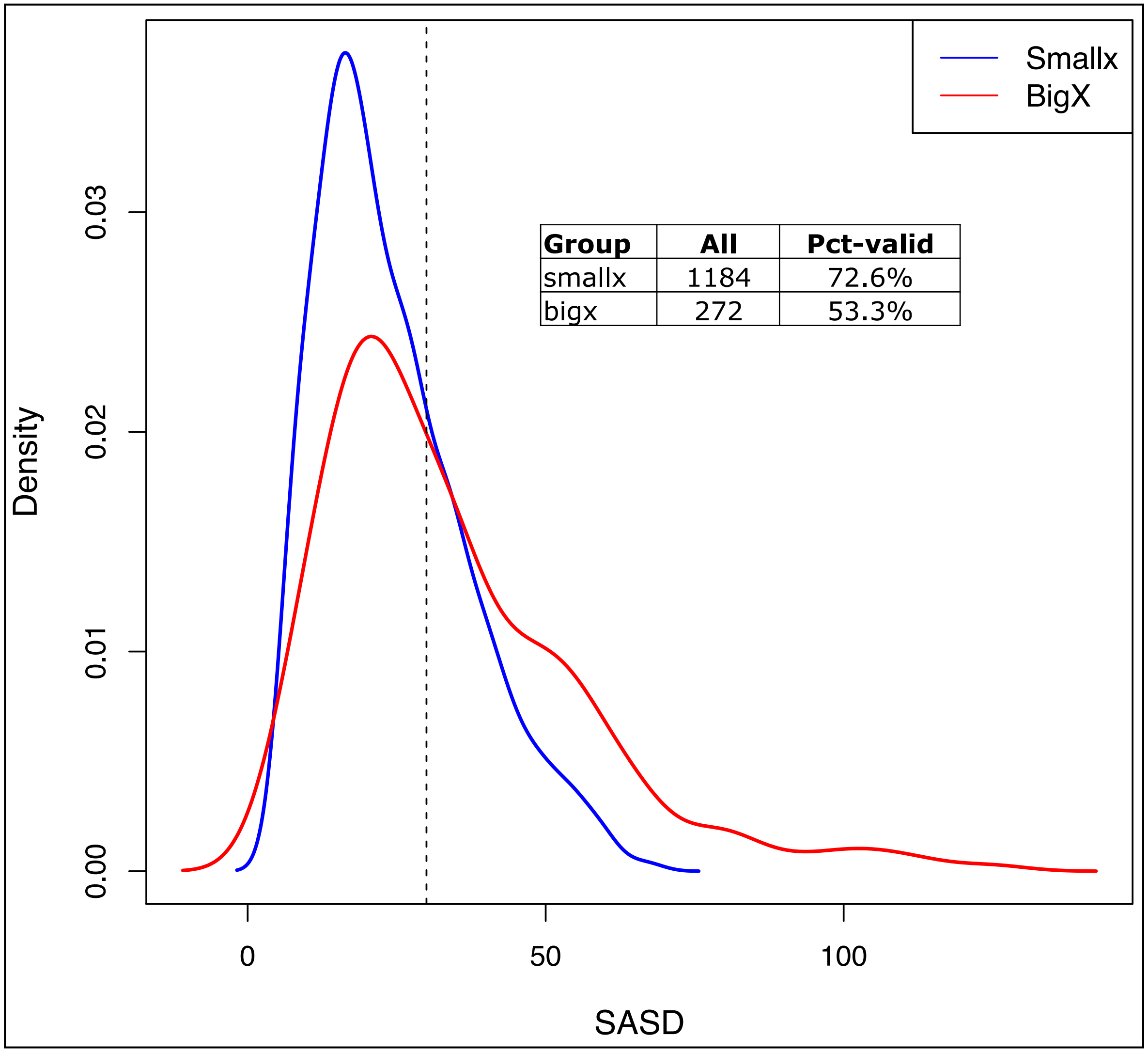
Do the Crosslink data make sense ?

- Check if Xlink data is ***<u>valid</u>*** (within 30 Å on the surface of the solvent accessible area between two linked residue)

- Check if crosslink data is ***<u>informative</u>*** (connecting residues 50 or more positions apart)

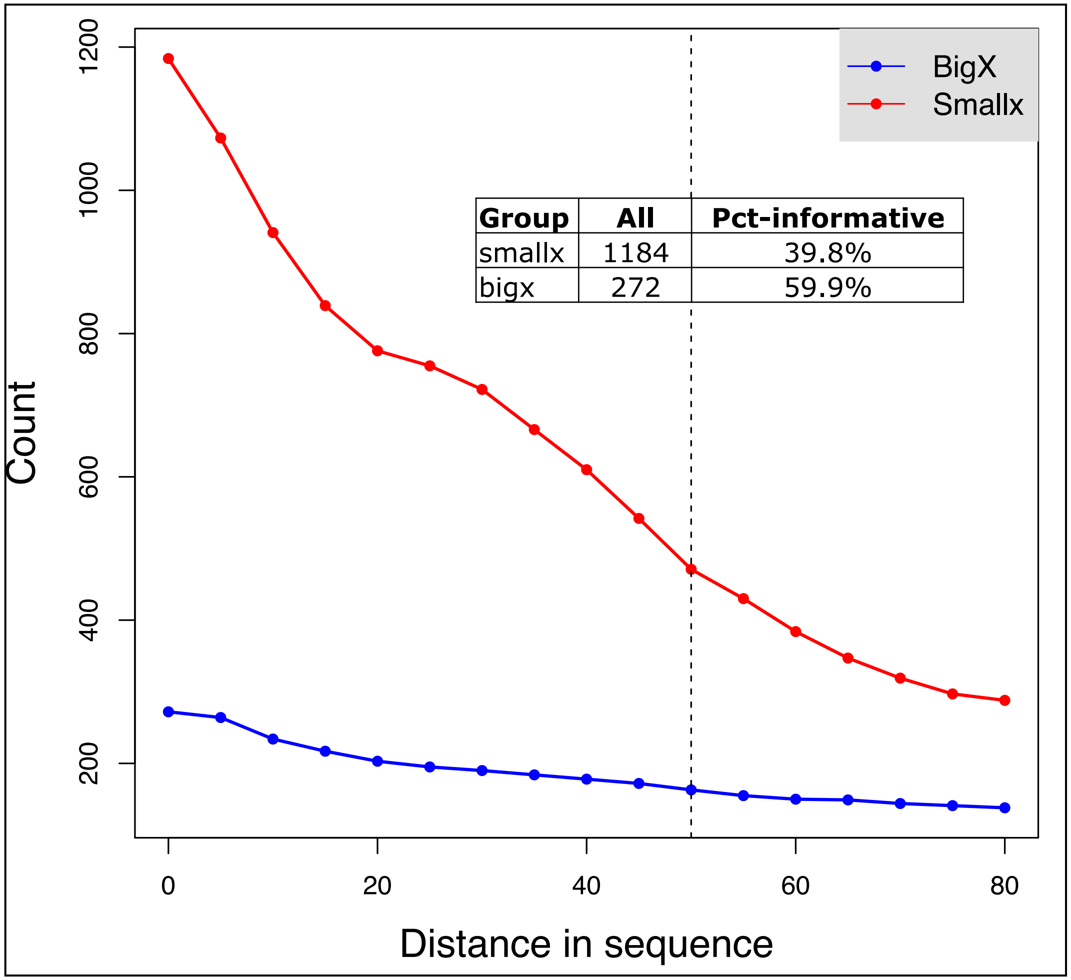
Valid crosslinks (single chains)

Distribution of crosslinks as a function of SASD between residues



Solvent Accessible Surface Distance (SASD)

Number of *informative* crosslinks in each group



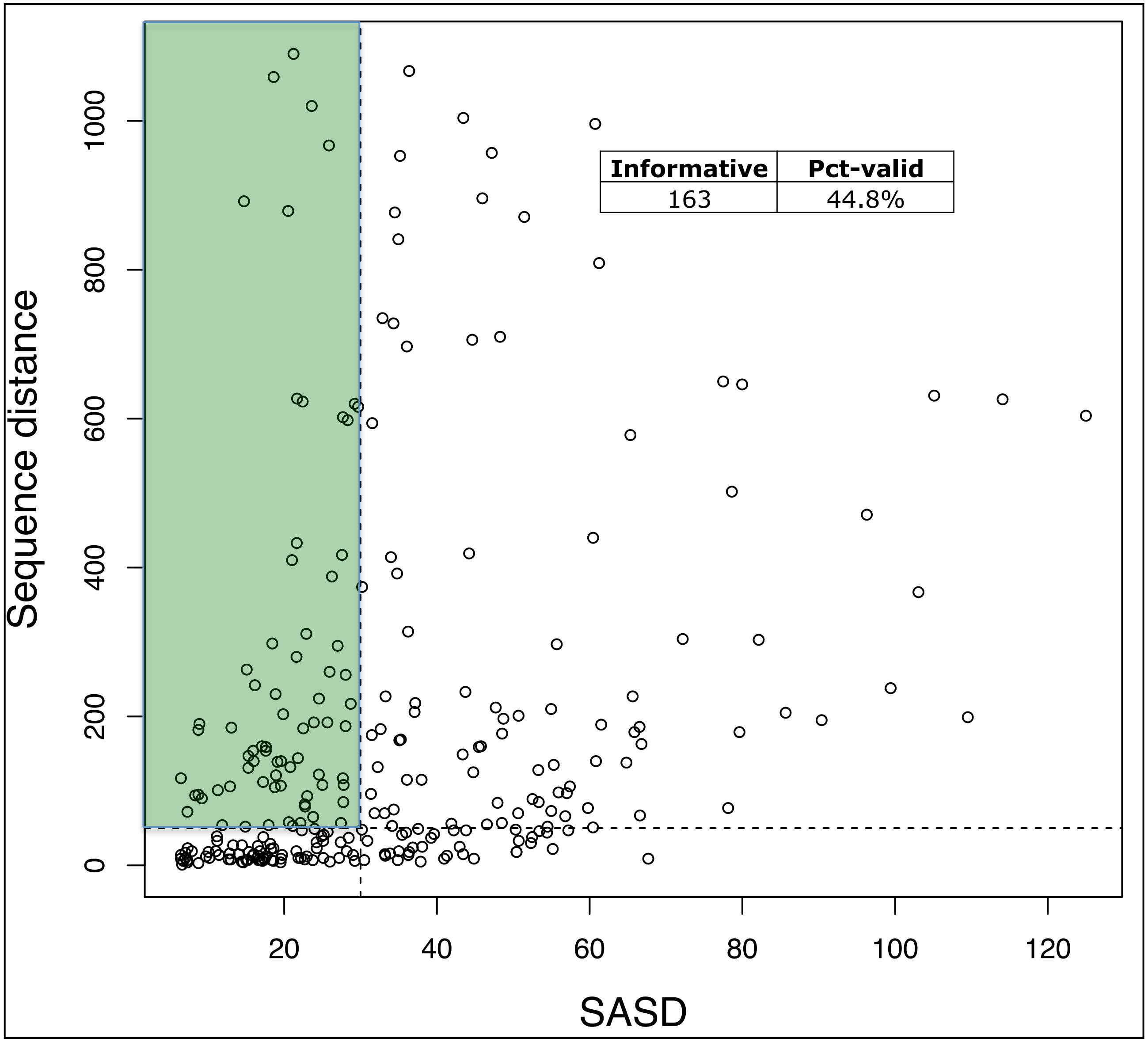
Group	All	Pct-informative
smallx	1184	39.8%
bigx	272	59.9%

Overall numbers of *valid* and *informative* crosslinks

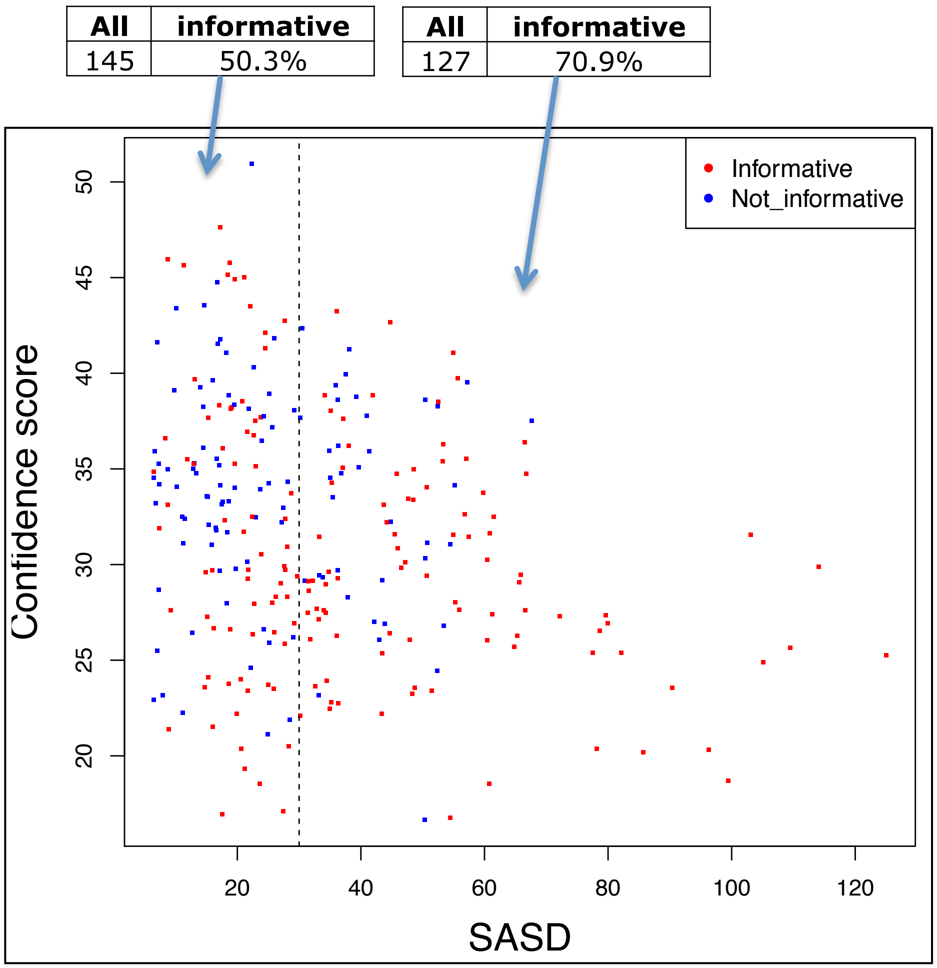
Set	**All crosslinks**	**Valid**	**Informative**	**Valid+Info**	**Valid+info/ informative**
smallX	1184	859	471	277	58.8%
bigX	272	145	163	73	44.8%

Can confidence scores help to enrich valid crosslinks?

Crosslinks in BigX group



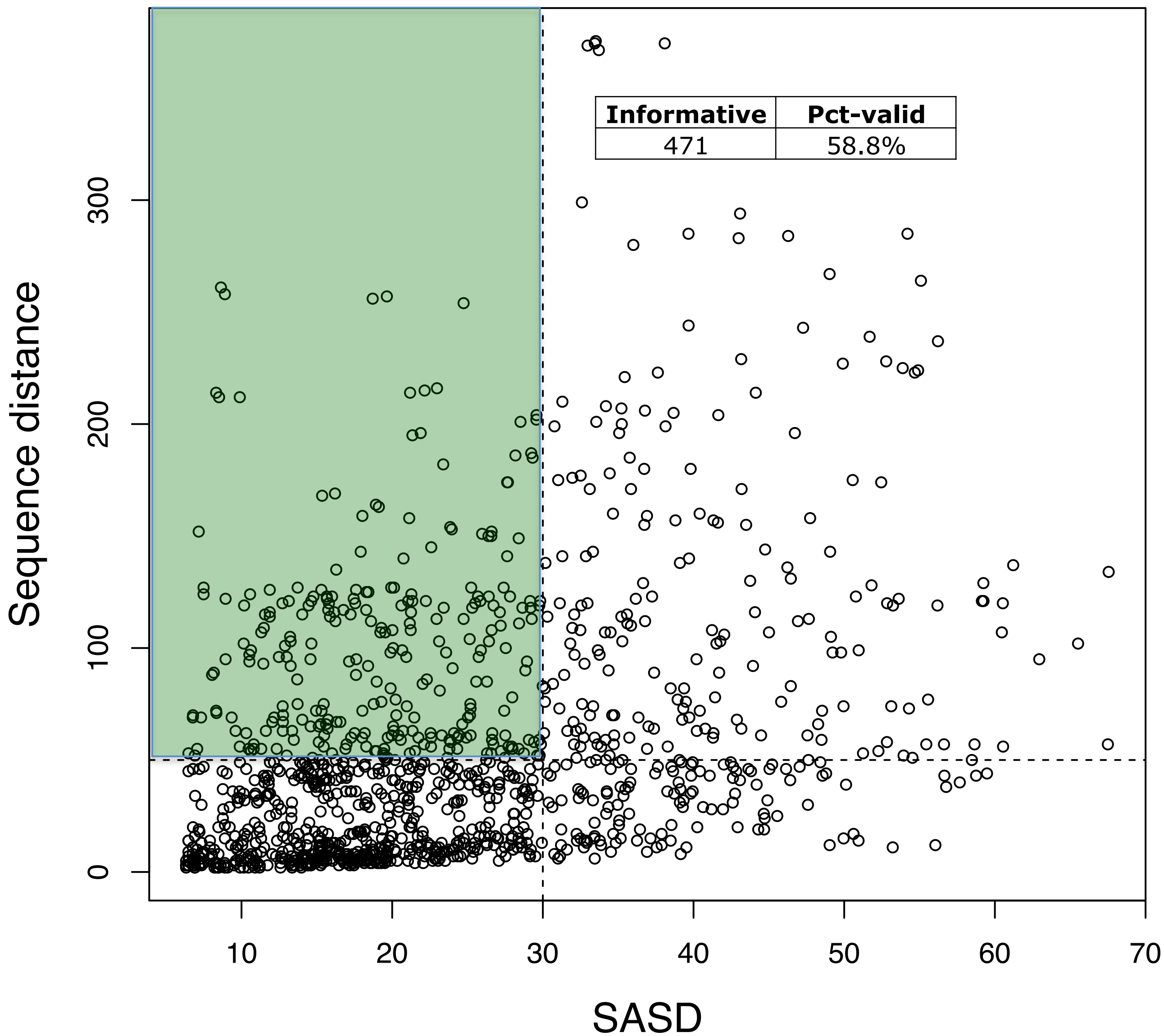
Usefulness of confidence scores for crosslinks in BigX group



Confidence cutoff vs Informative Xlinks

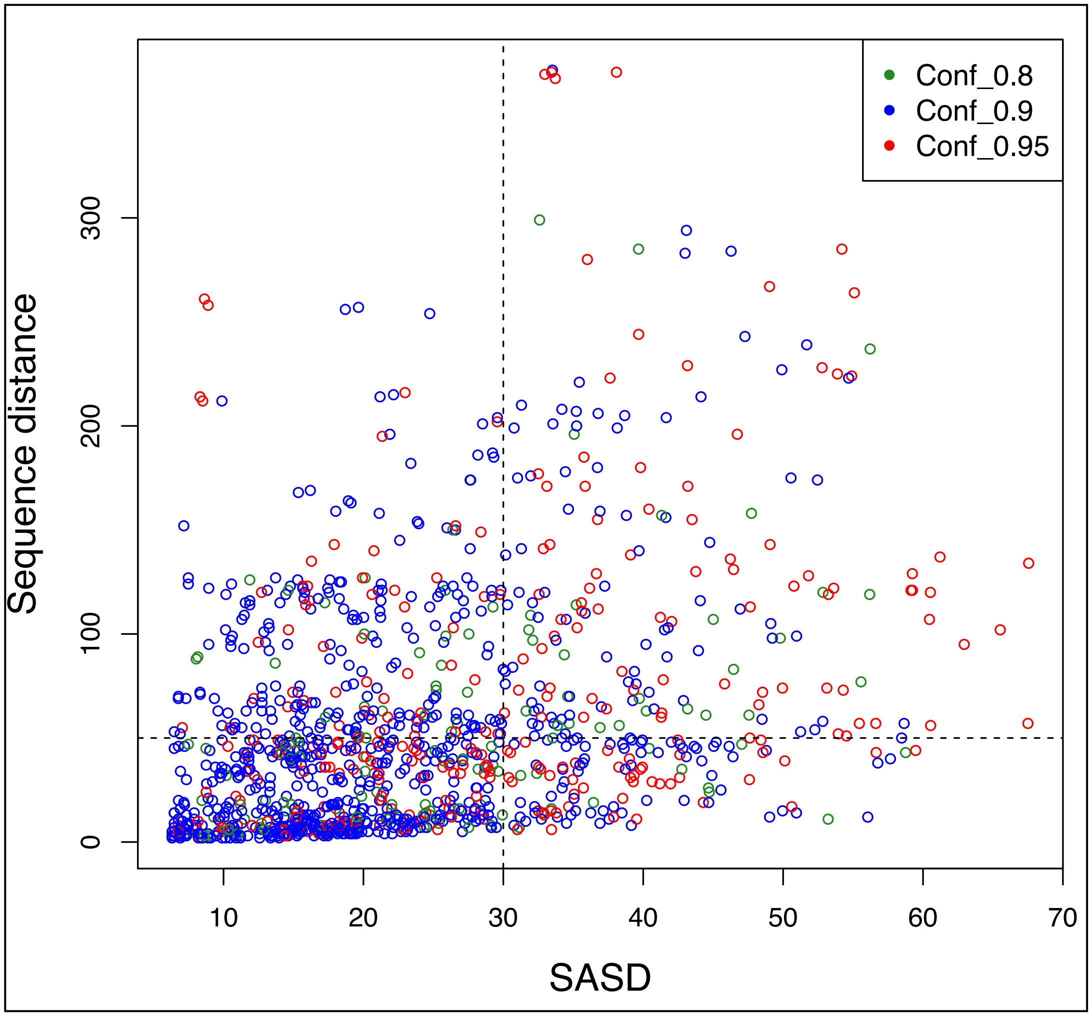
Confidence cutoff	# Xlinks left	% of total	% valid
15	163	100%	45%
20	156	96%	44%
25	129	79%	44%
30	73	45%	51%
35	42	26%	64%
40	14	9%	79%
45	6	4%	100%
50	0	0%	0%

Crosslinks in smallX group



Informative	Pct-valid
471	58.8%

Usefulness of Confidence scores for crosslinks in smallX group



Distribution of crosslinks in smallX group by confidence score

All

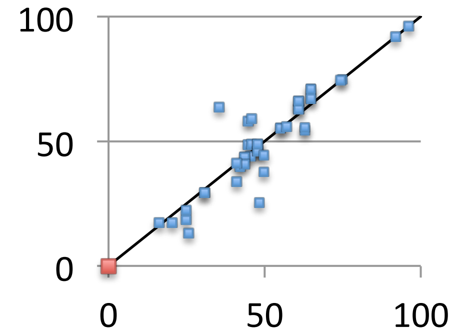
	All	Percent valid
All (80% and up)	1184	39.00%
>= 90% confidence	901	42.81%
95% confidence	768	43.4%

Informative

	All	Percent valid
All (80% and up)	471	58.8%
>= 90% confidence	336	67.0%
95% confidence	282	70.6%

All targets (single chains)

Target	All	Valid	Informative	Valid+Inf	Valid+inf/ informative
x0957S2	85	70	13	0	0.0%
X0953S2	5	3	0	0	0.0%
X0957S2	2	2	0	0	0.0%
X0968S2	5	5	0	0	0.0%
x0957S1D2	12	9	0	0	0.0%
x0957S2D1	83	68	13	0	0.0%
X0953S1	0	0	0	0	0.0%
X0953S2D2	2	2	0	0	0.0%
X0953S2D3	1	1	0	0	0.0%
X0957S1D2	2	2	0	0	0.0%
X0968S2D1	5	5	0	0	0.0%
X0999D5	0	0	0	0	0.0%
X0957S2	2	2	0	0	0.0%
X0968S1	9	8	1	1	100.0%
X0968S1D1	9	8	1	1	100.0%
X0957S1	7	7	2	2	100.0%
x0957S1D1	73	66	6	2	33.3%
X0957S1D1	2	2	2	2	100.0%
X0999D3	8	3	5	2	40.0%
X0999D4	5	4	2	2	100.0%
X0987D1	15	9	4	3	75.0%
X0999D1	12	10	5	3	60.0%
X0999D2	10	5	7	3	42.9%
x0968S2	76	69	5	5	100.0%
x0968S2D1	76	69	5	5	100.0%
X0987D2	20	12	6	6	100.0%
X0975	19	14	10	7	70.0%
X0975D1	19	14	10	7	70.0%
X0981	25	15	14	8	57.1%
X0985	37	21	19	9	47.4%
X0985D1	38	22	19	9	47.4%
x0968S1	68	50	20	16	80.0%
X0987	66	28	37	16	43.2%
x0968S1D1	68	50	20	16	80.0%
x0987D1	147	108	29	20	69.0%
x0957S1	144	116	41	26	63.4%
X0999	97	42	80	30	37.5%
x0987D2	246	193	96	77	80.2%
x0975	272	192	144	90	62.5%
x0975D1	272	192	144	90	62.5%
x0987	539	362	248	140	56.5%

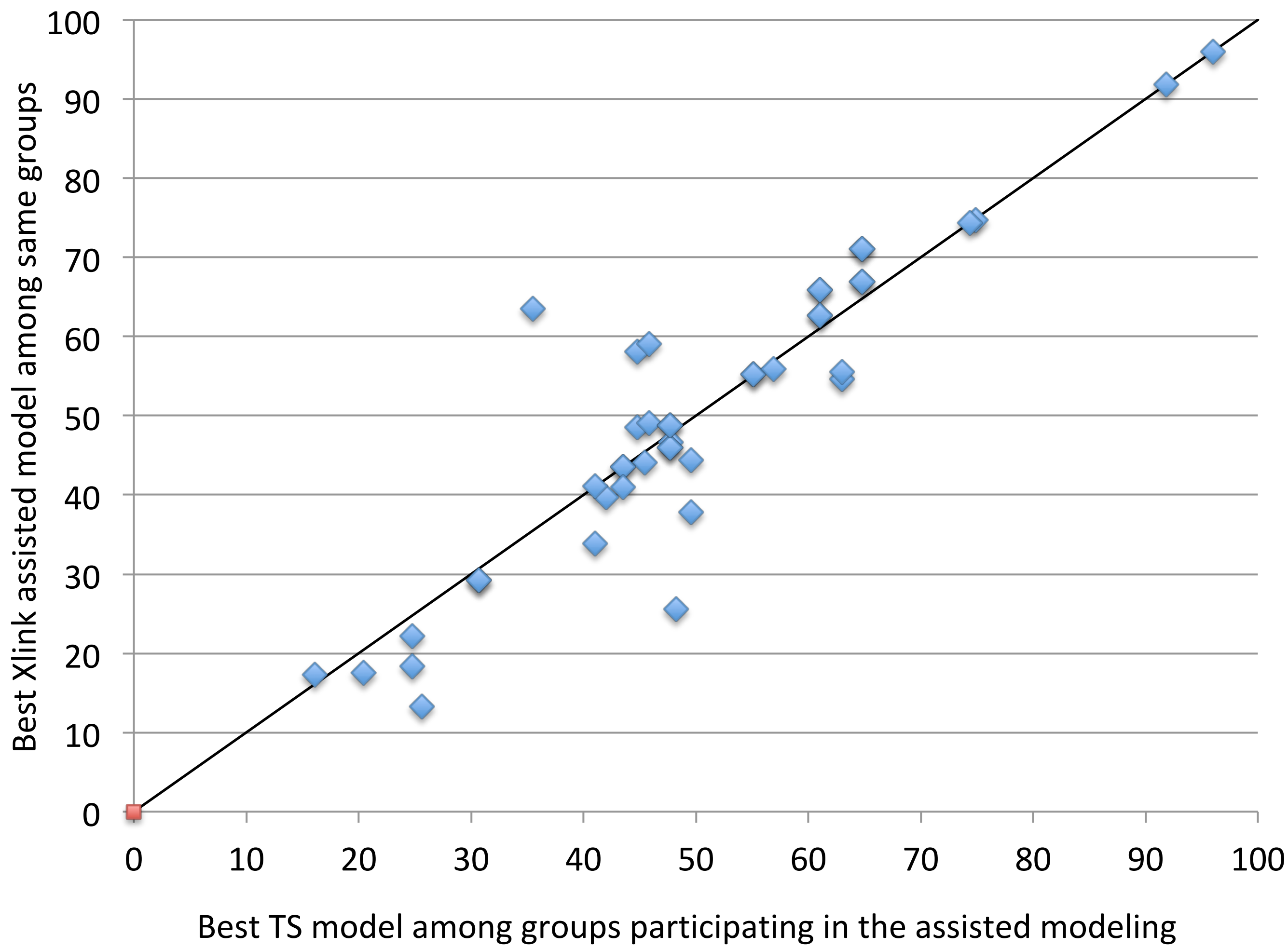


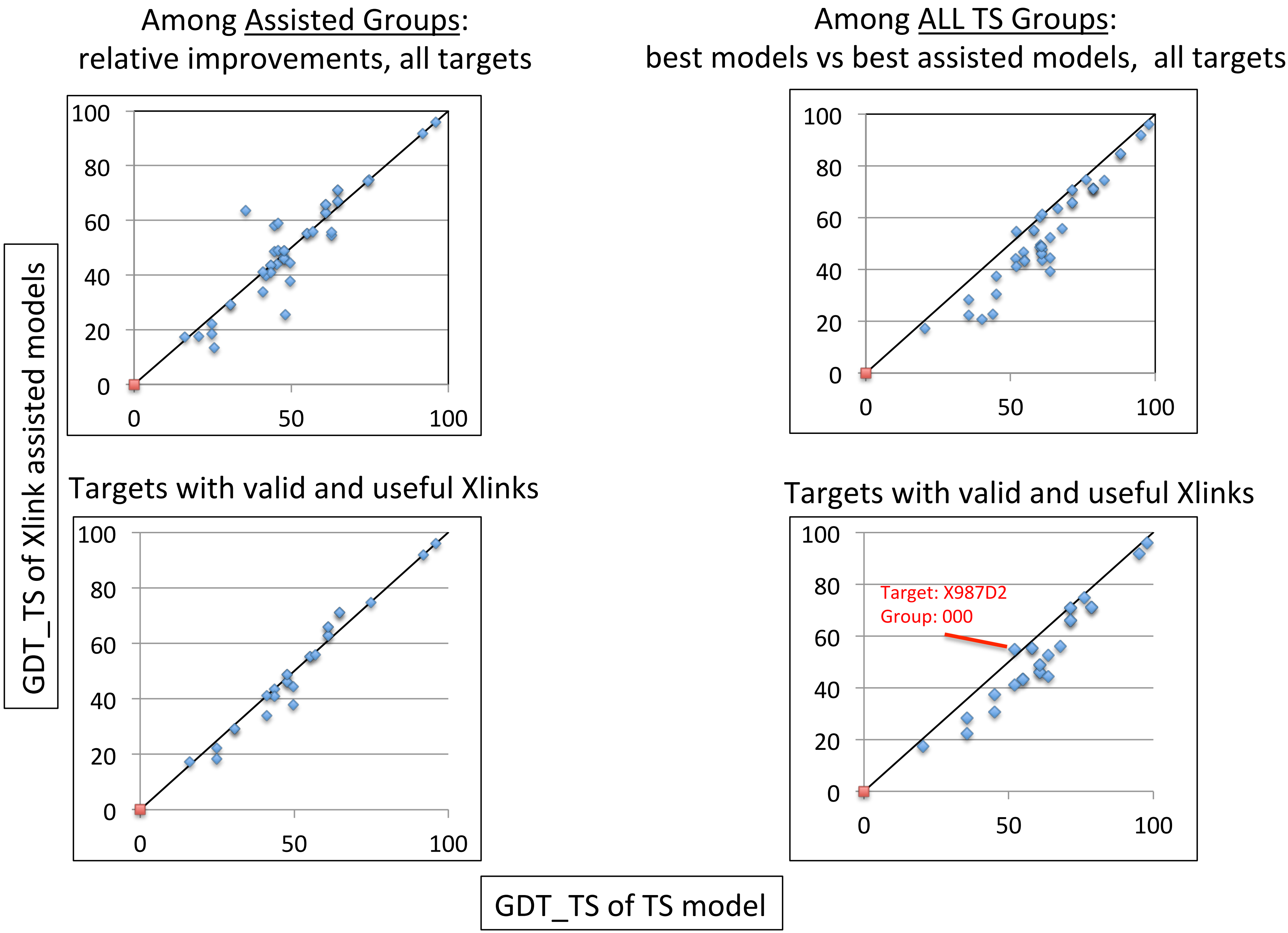
Subset of targets with at least 1 valid and informative Xlink

Targets with at least one valid+informative crosslink, (sorted by valid+inf/informative)

Target	All	Valid	Informative	Valid-Inf	valid-inf/ inform
x0957S1D1	73	66	6	2	33.3%
X0999	97	42	80	30	37.5%
X0999D3	8	3	5	2	40.0%
X0999D2	10	5	7	3	42.9%
X0987	66	28	37	16	43.2%
X0985	37	21	19	9	47.4%
X0985D1	38	22	19	9	47.4%
x0987	539	362	248	140	56.5%
X0981	25	15	14	8	57.1%
X0999D1	12	10	5	3	60.0%
x0975	272	192	144	90	62.5%
x0975D1	272	192	144	90	62.5%
x0957S1	144	116	41	26	63.4%
x0987D1	147	108	29	20	69.0%
X0975	19	14	10	7	70.0%
X0975D1	19	14	10	7	70.0%
X0987D1	15	9	4	3	75.0%
x0968S1	68	50	20	16	80.0%
x0968S1D1	68	50	20	16	80.0%
x0987D2	246	193	96	77	80.2%
X0987D2	20	12	6	6	100.0%
X0999D4	5	4	2	2	100.0%
X0968S1	9	8	1	1	100.0%
X0968S1D1	9	8	1	1	100.0%
x0968S2	76	69	5	5	100.0%
x0968S2D1	76	69	5	5	100.0%
X0957S1	7	7	2	2	100.0%
X0957S1D1	2	2	2	2	100.0%

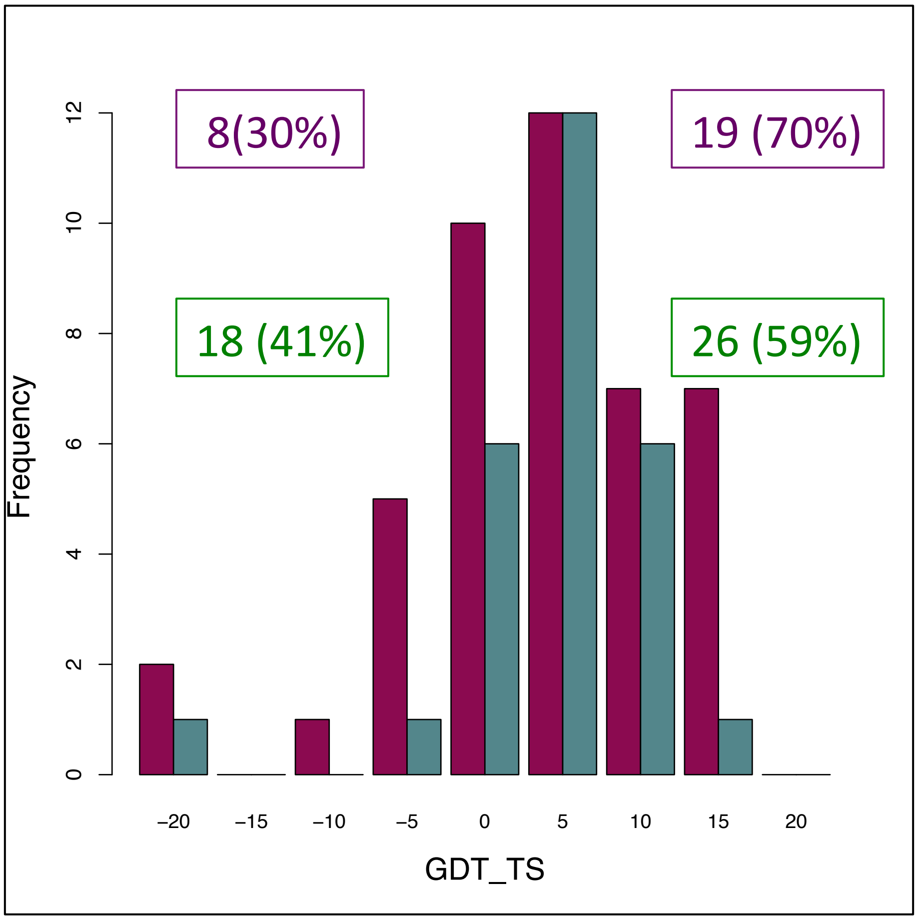
Single best 3D model vs single best assisted model for each target within group of assisted predictors only
(Groups do not need to match!)




Among Assisted Groups:
relative improvements, all targets
Among ALL TS Groups:
best models vs best assisted models, all targets
Targets with valid and useful Xlinks
Targets with valid and useful Xlinks
Target: X987D2
Group: 000
GDT_TS of Xlink assisted models
GDT_TS of TS model
0
20
40
60
80
100
0
50
100

CASP:
Page redacted

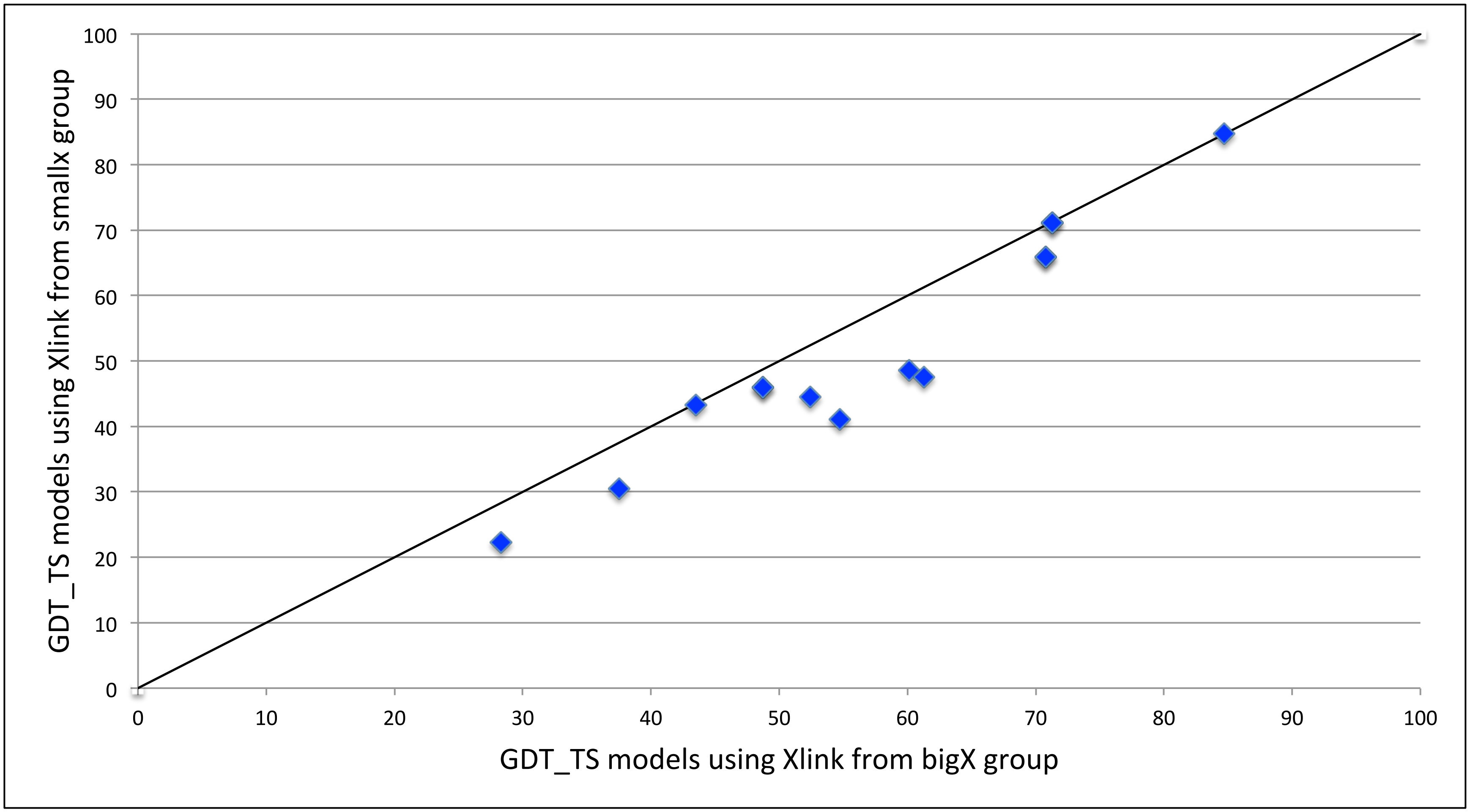
Head-to-head comparison of GDT_TS changes for each group and each model (when available)



Red: all targets (44 data points)
Green: subset of targets with valid and informative Xlinks (27 data points)

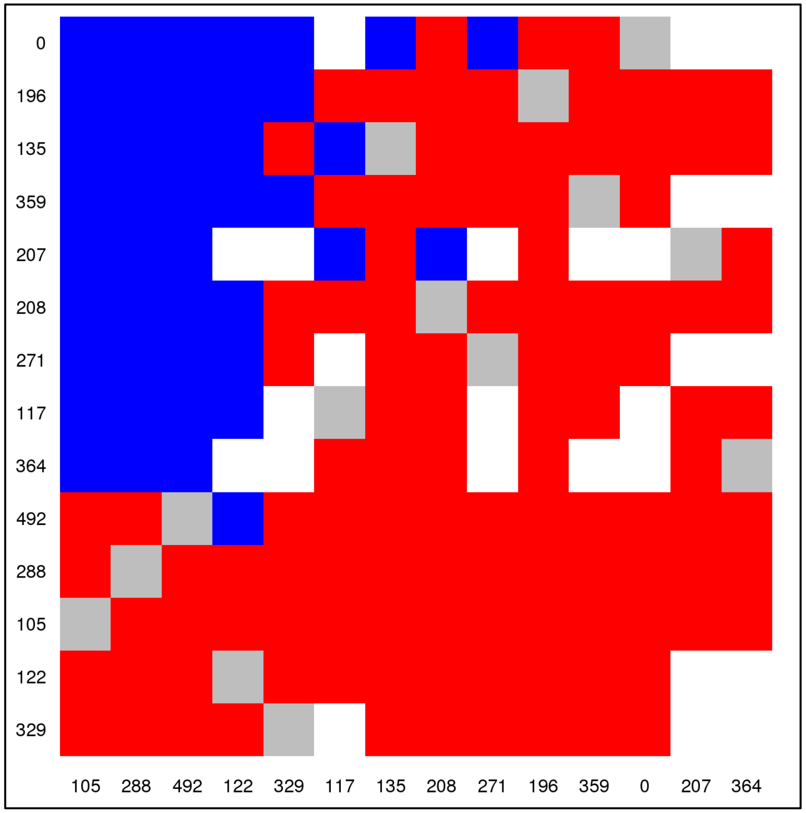
Averages: **1.75** and **2.12** GDT_TS

Best Xlink assisted targets using Xlink from smallx or bigX source

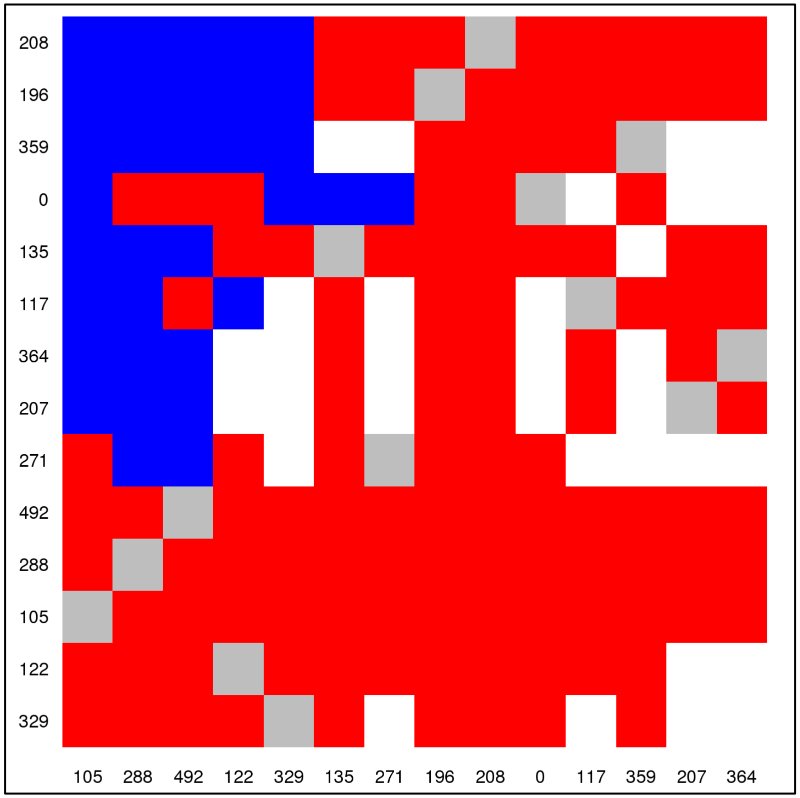


Group performance

All targets



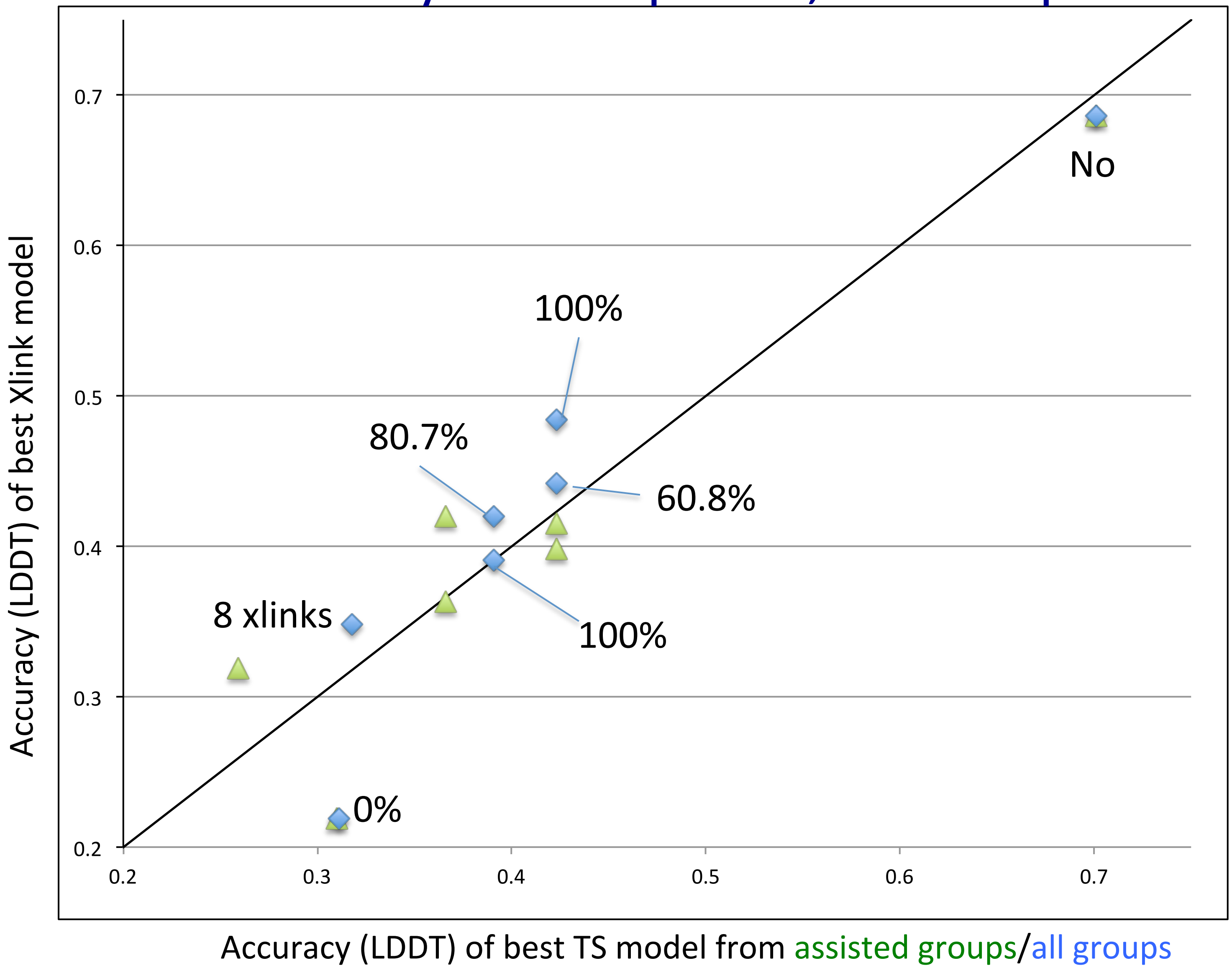
Targets with *valid* and *informative* Xlinks



At each square, GDT-TS score distribution of vertical and horizontal axis groups was compared with one-tailed Wilcoxon test at a significance level of 0.05

Blue: Vertical performed better than horizontal
Red: Vertical not significantly better than horizontal
White: Not enough shared targets between groups
Gray: Vertical and horizontal are the same group

Accuracy on complexes, a blowup



Information about Xlinks added: % valid or NO suitable crosslinks

Conclusions

- The largest exercise to date about assisted modeling
- Algorithmic challenge how to select and incorporate xlinks in modeling
- Confidence scores can help in enrichment but at the expense of losing a large number of correct xlinks
- Single chain Xlink assisted modeling shows anecdotal promise over all models
- Single chain Xlink assisted models have a trend to improve over unassisted models of the same group
- Modeling heterocomplexes is promising but very few data

Acknowledgement

CASP and Predictioncenter at UC Davies, Davies, USA:

Andriy Kryshtafovych
Bohdan Monastyrskyy
Krzysztof Fidelis

CASP organizers

Albert Einstein College of Medicine, New York, USA:

Rojan Shrestha
Eduardo Fajardo
Nelson Gil

Questions to presenters

- the generation of initial decoys
- the actual formalism to incorporate crosslinks
- the function to weigh in confidence values (if any)
- if you discriminated between short and long range crosslinks (in terms of sequence separation)
- if you were combining information from various types of assisted modeling data, e.g. SAXS and Xlinks, together.