

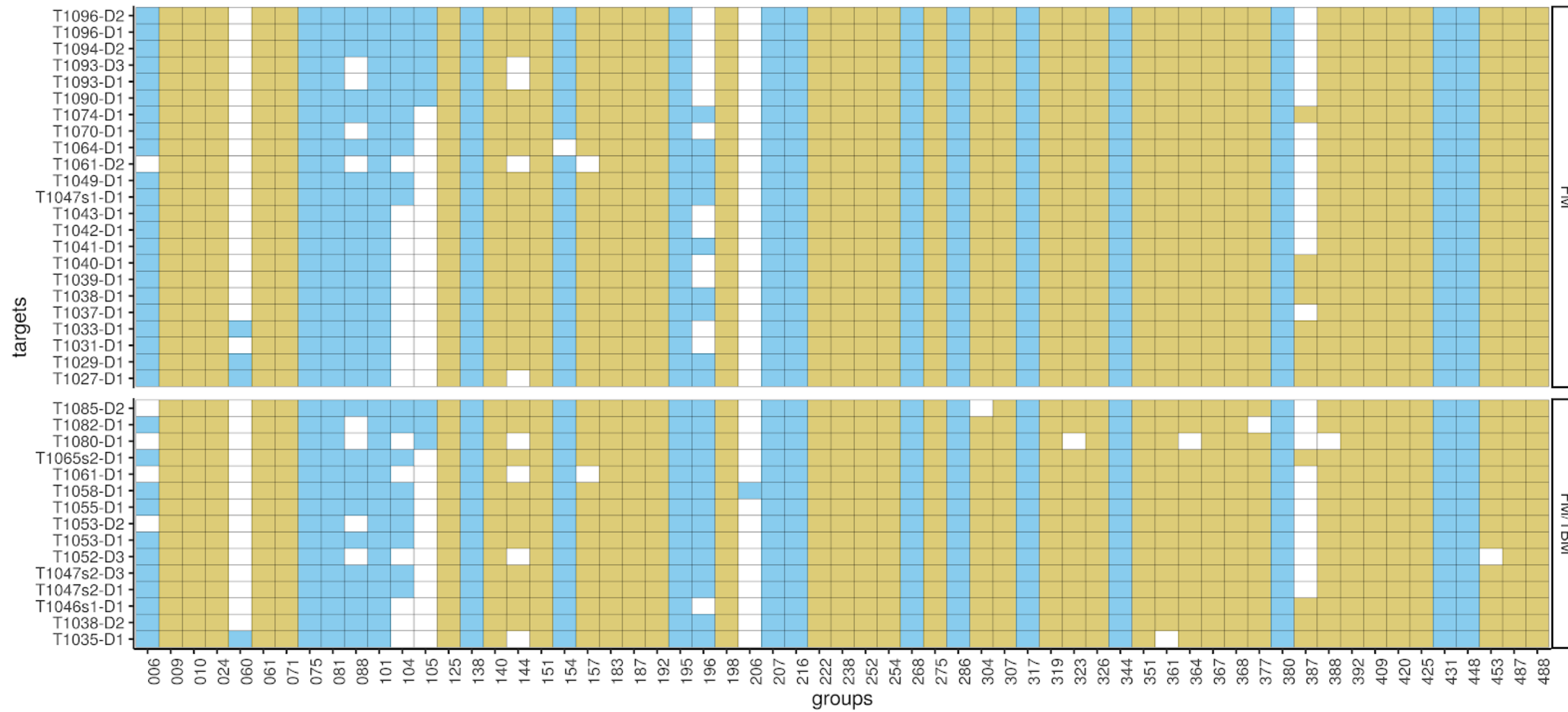
Evaluation of contact and distance prediction in CASP14

Rosalba Lepore
BSC Barcelona Supercomputing Center

3 Dec 2020

Overview of targets and participants

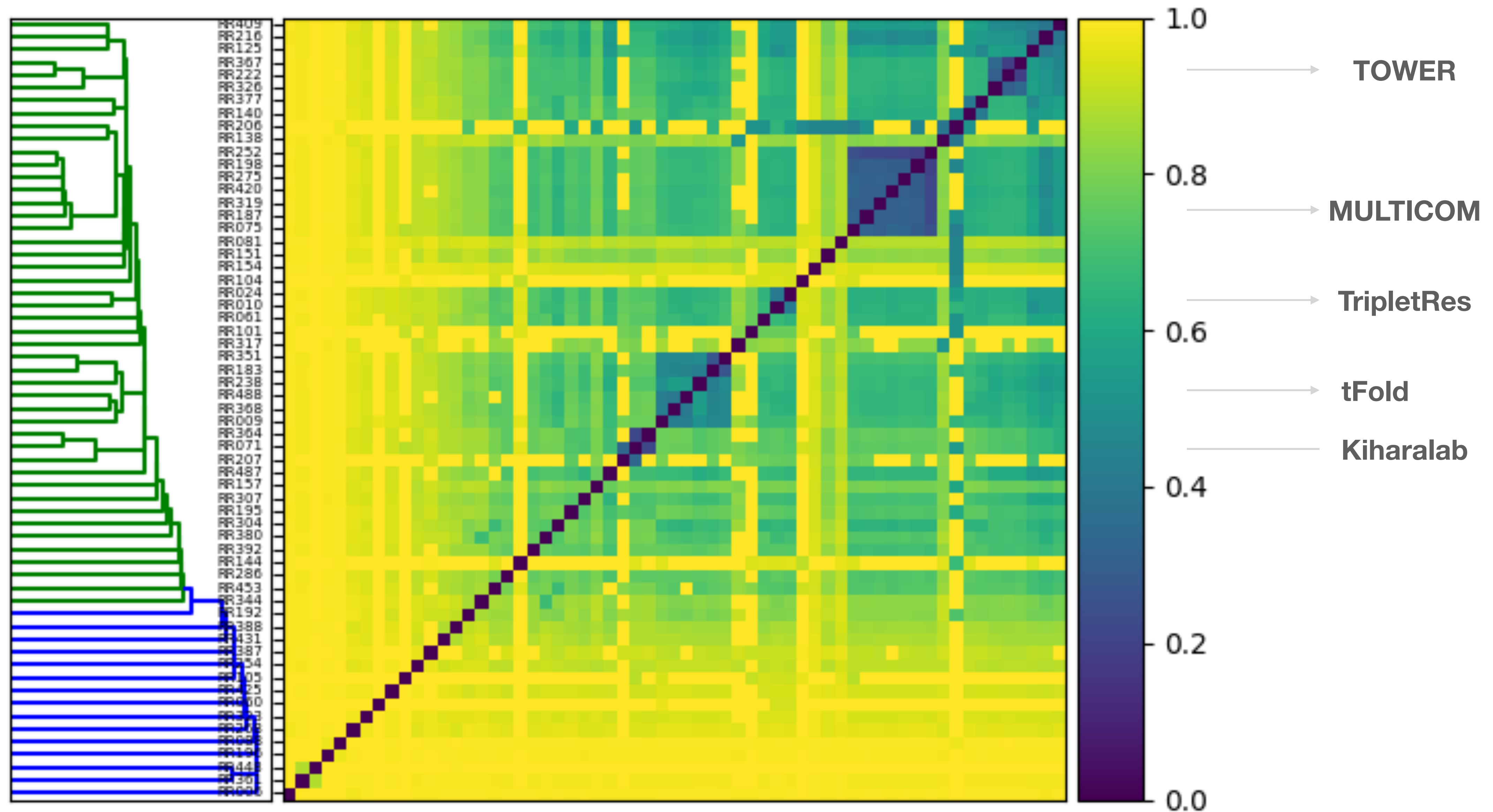
Submission Dist + Contacts Contacts Not submitted



38 evaluation units:
 • 23 FM
 • 15 FM/TBM

61 predictors:
39 contacts & distances

overview of submitted predictions (Jaccard distance = 1- Intersection / Union)



Contact assessment

Prediction format

List of contacts in 3 columns format:

- **i j p0**
- **i** and **j** are the indices of two the amino acids
- **p0** probability of the two residues being within 8Å distance

Evaluation:

Prediction are trimmed to domains

i-j pairs excluded if sequence space separation is < 6 aa

Non-listed aa pairs are assigned p0 = 0

Metrics

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

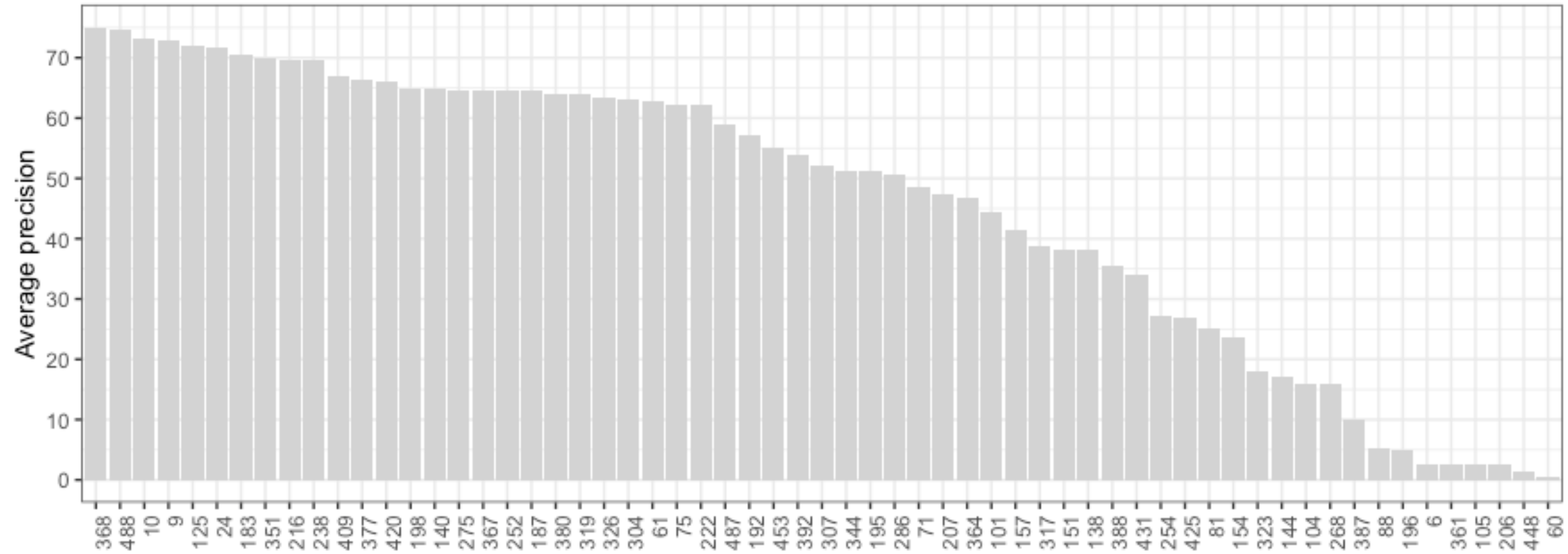
$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{Entropy score} = 100 * \frac{H_{\text{struct}|_{\text{ext}}} - H_{\text{struct}|_{\text{contact}}}}{H_{\text{struct}|_{\text{ext}}}}$$

Results in contact prediction

Average precision - all groups all targets

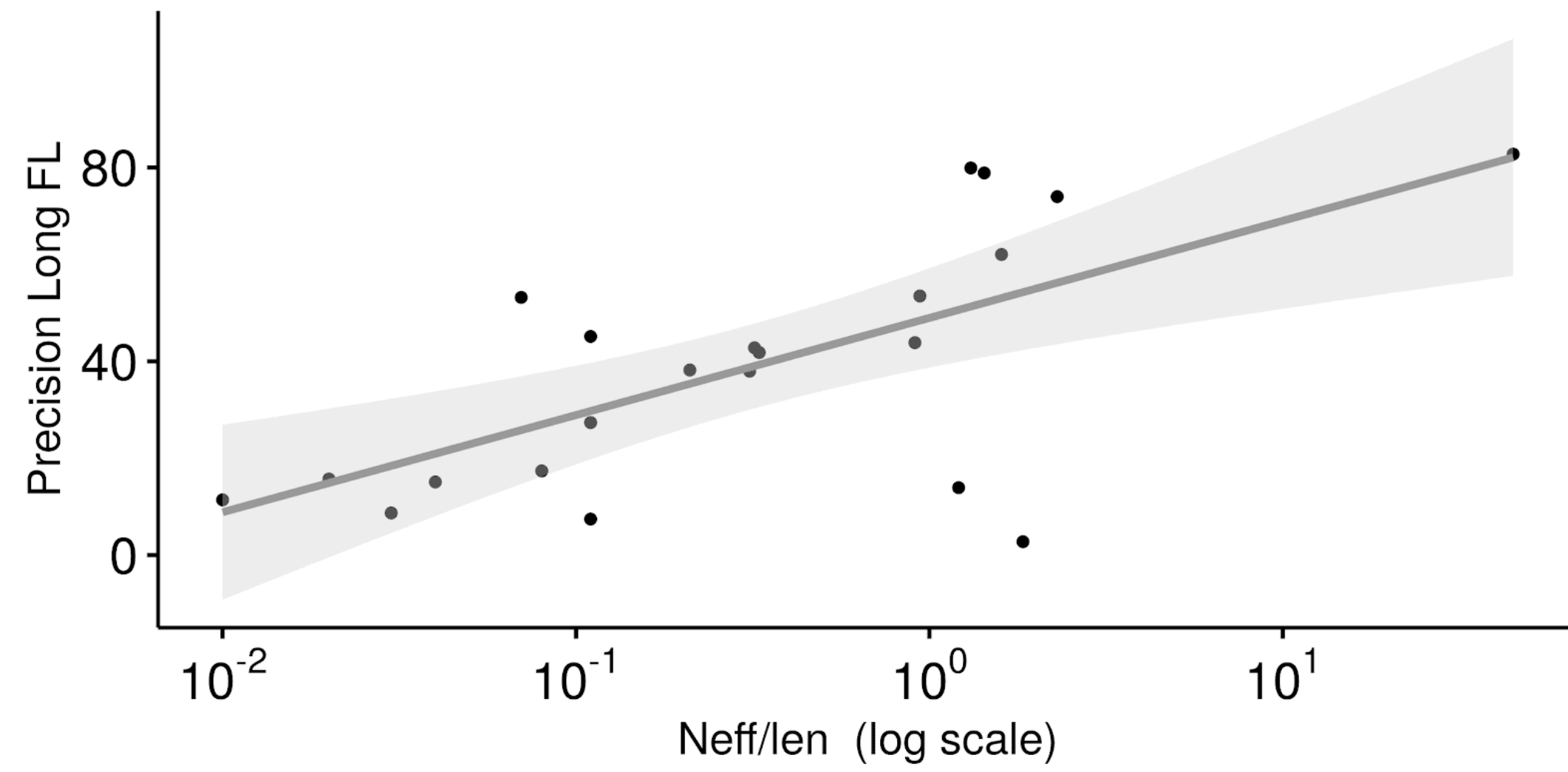
Average precision FM+FM/TBM Long L5 p>0



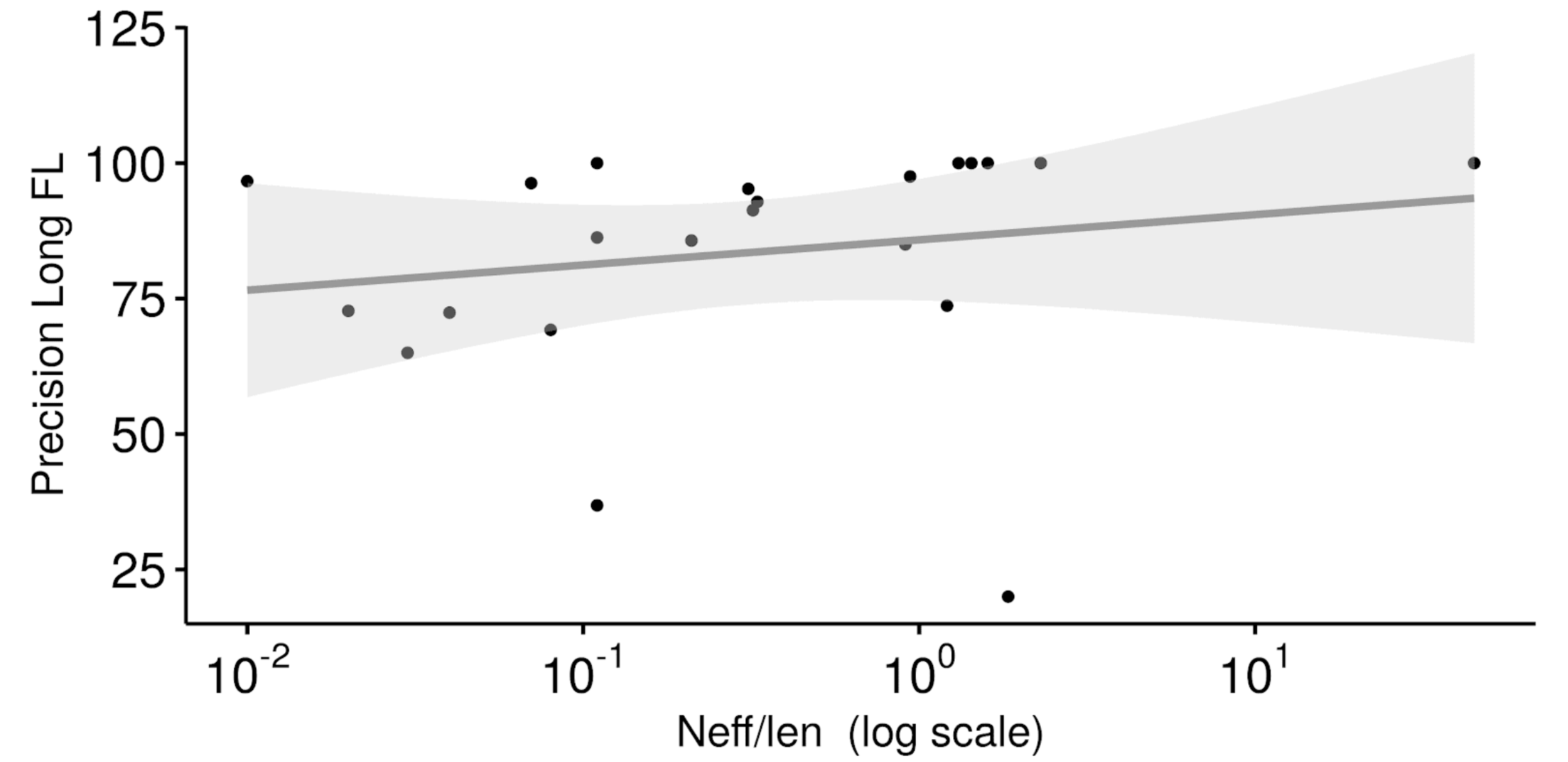
Top 10 groups reached ~ 70% average precision (an excellent result)

Dependence on alignment depth

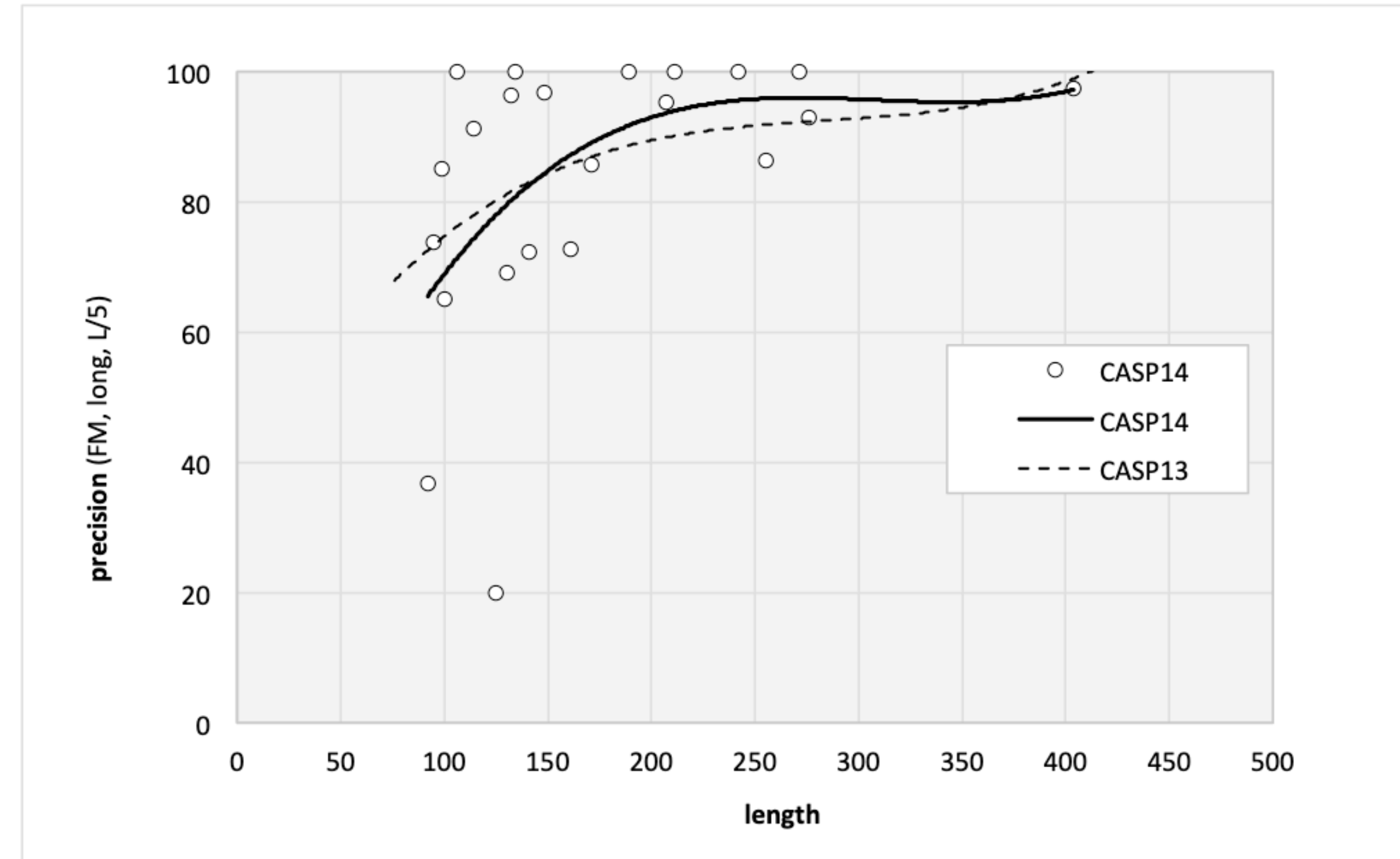
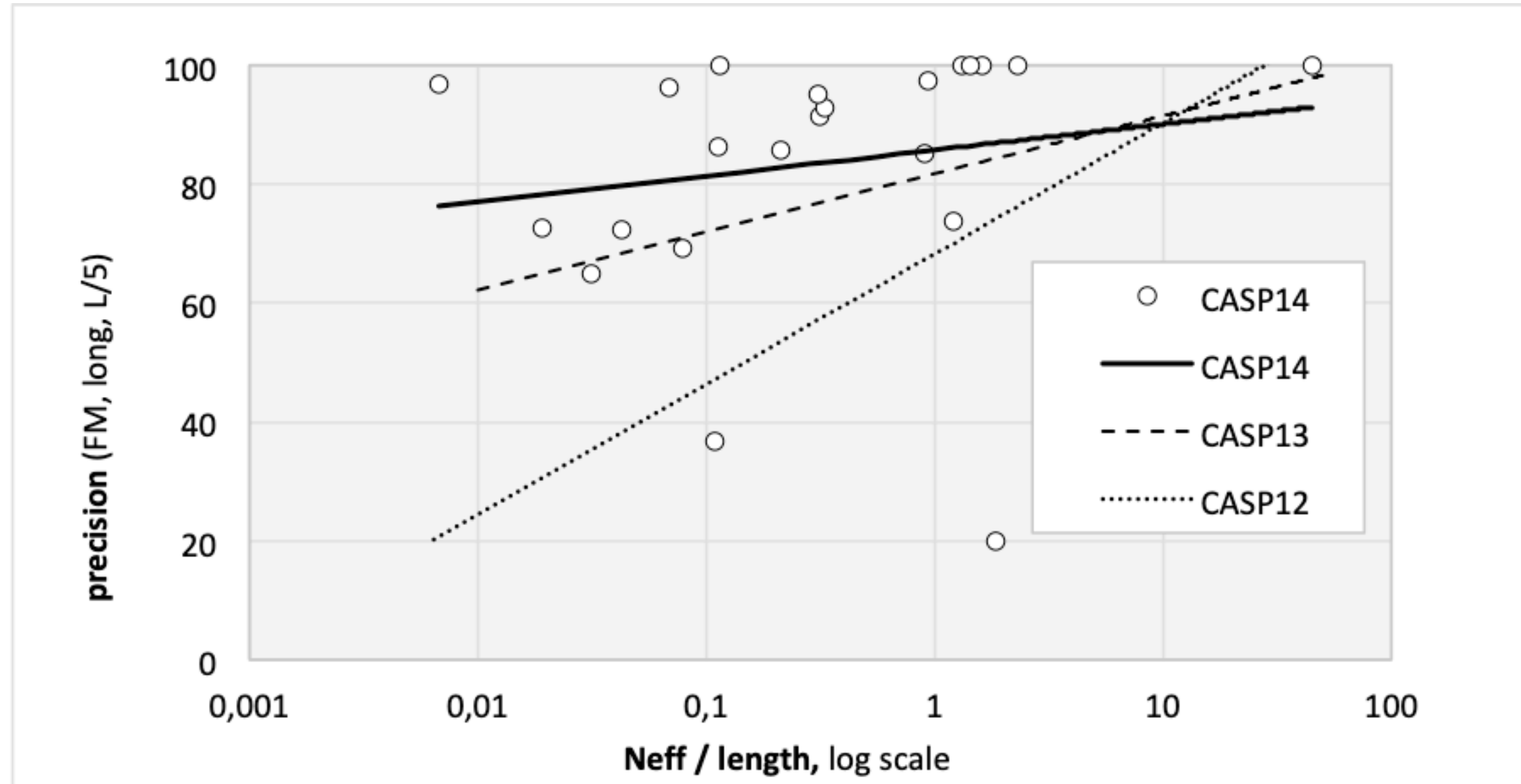
average precision



best precision

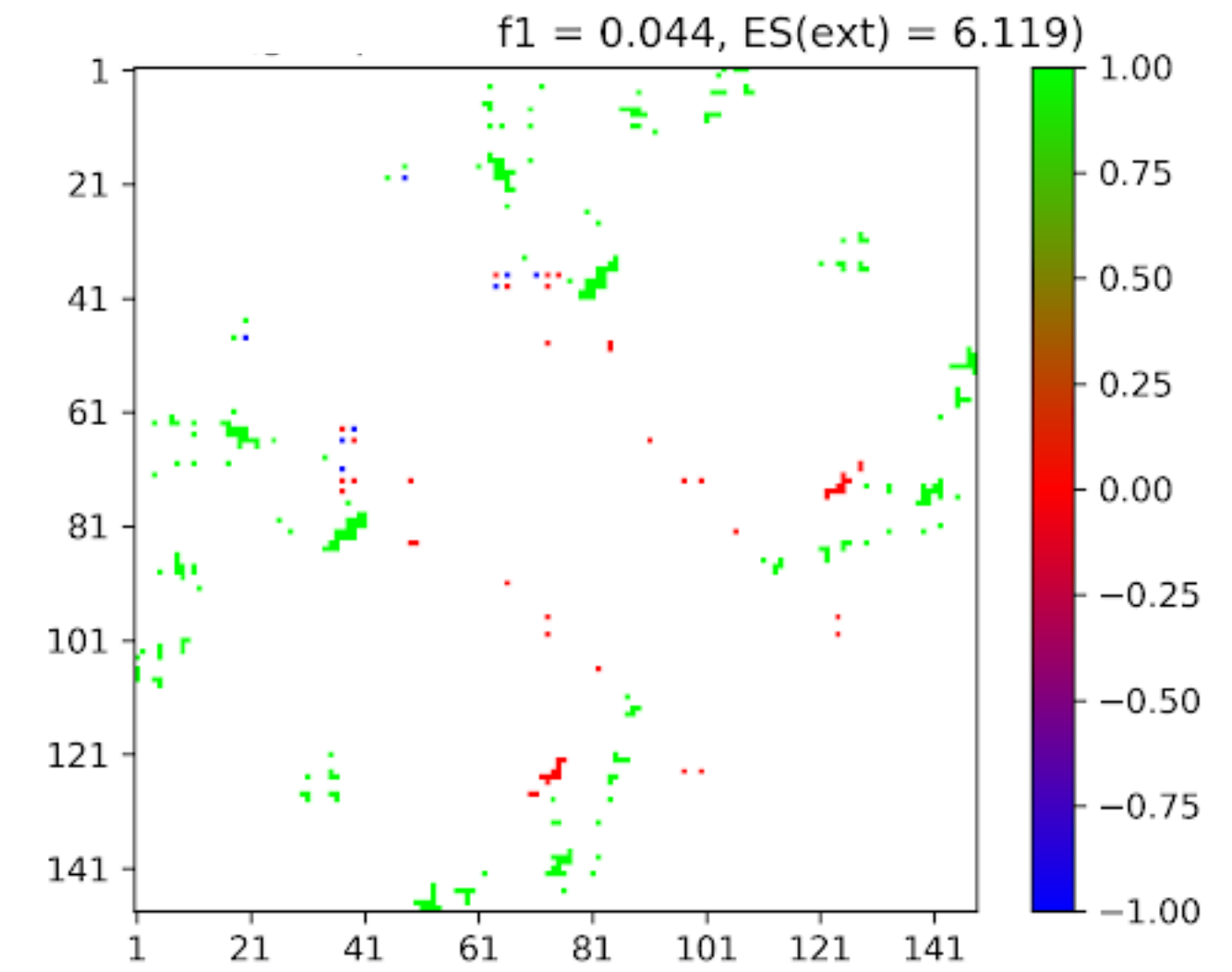
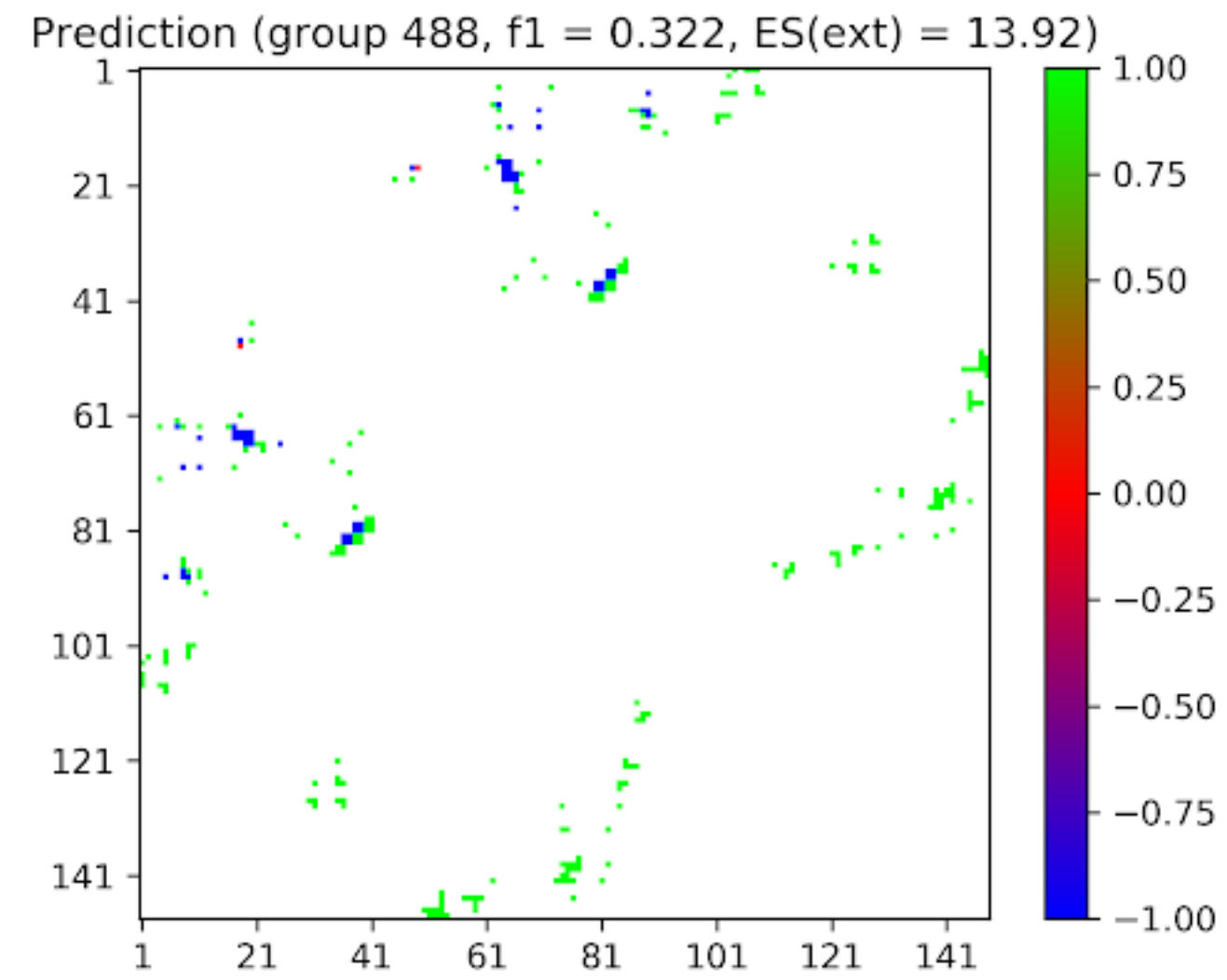
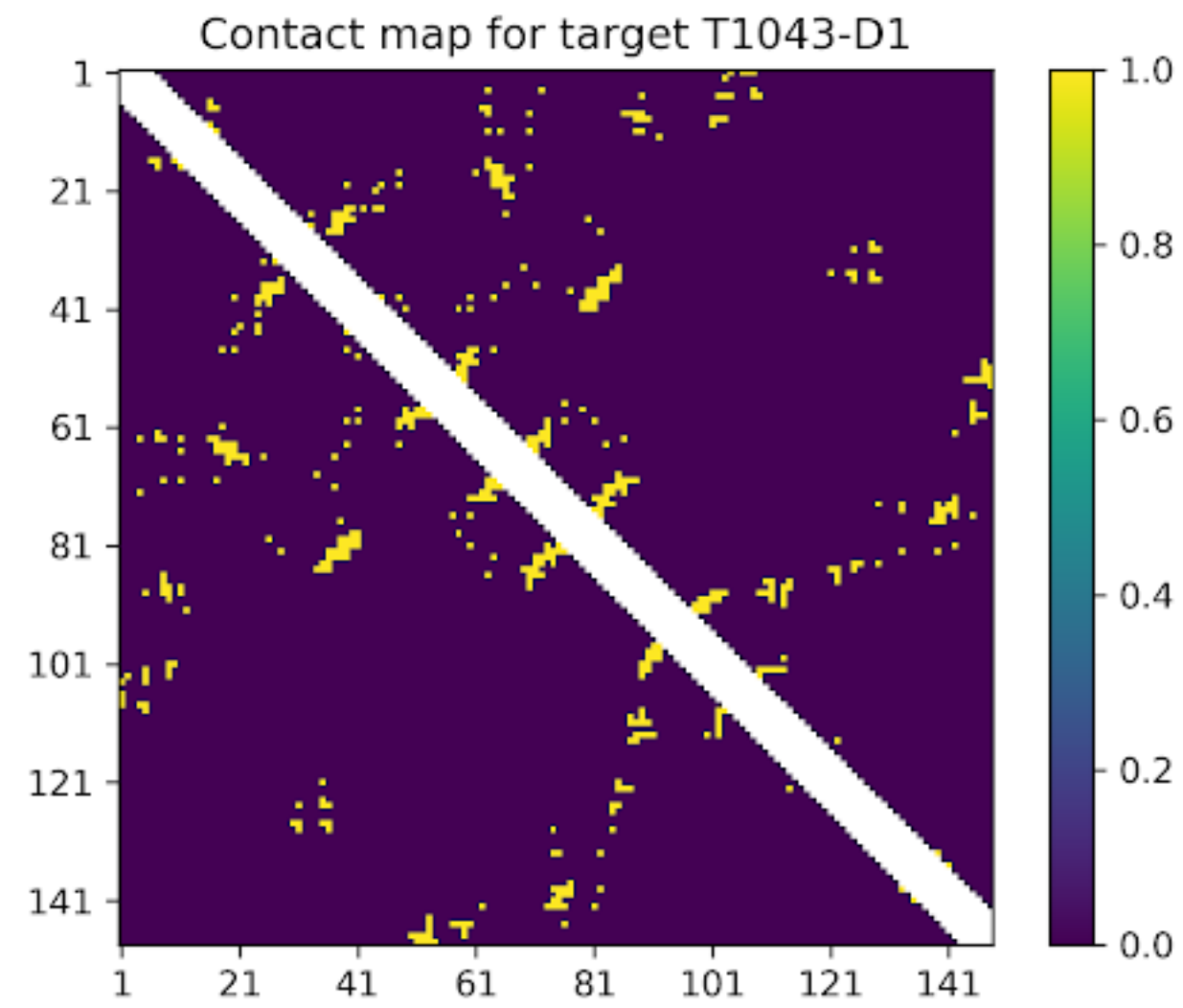


Best precision in CASP14 vs previous CASPs



Less reliant on deep alignments / better at extracting signal from small alignments

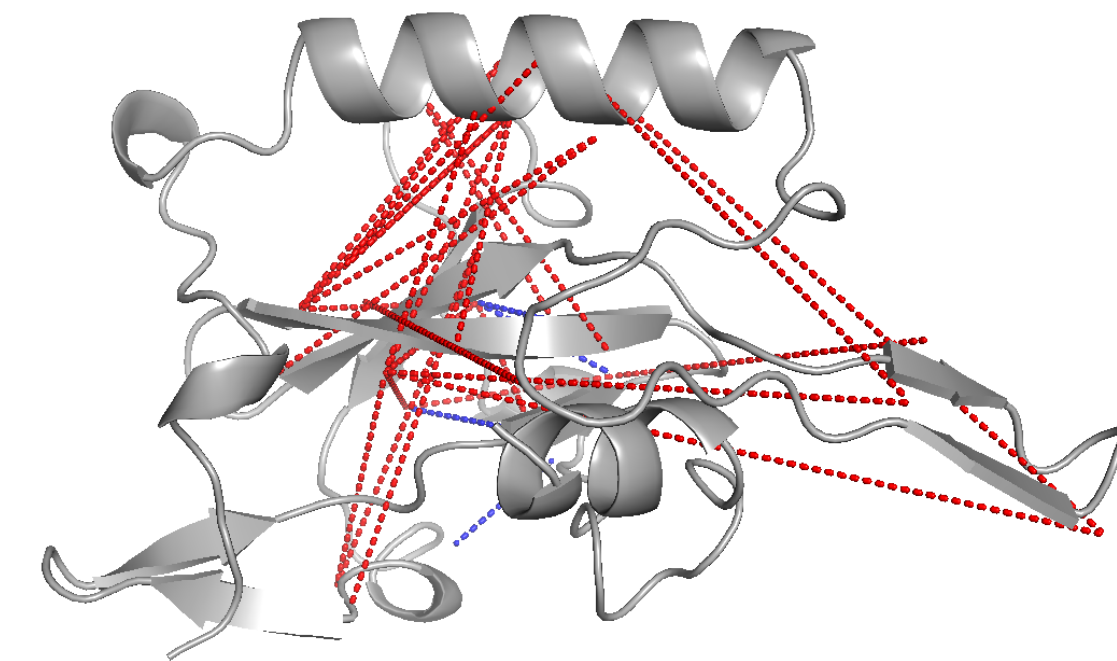
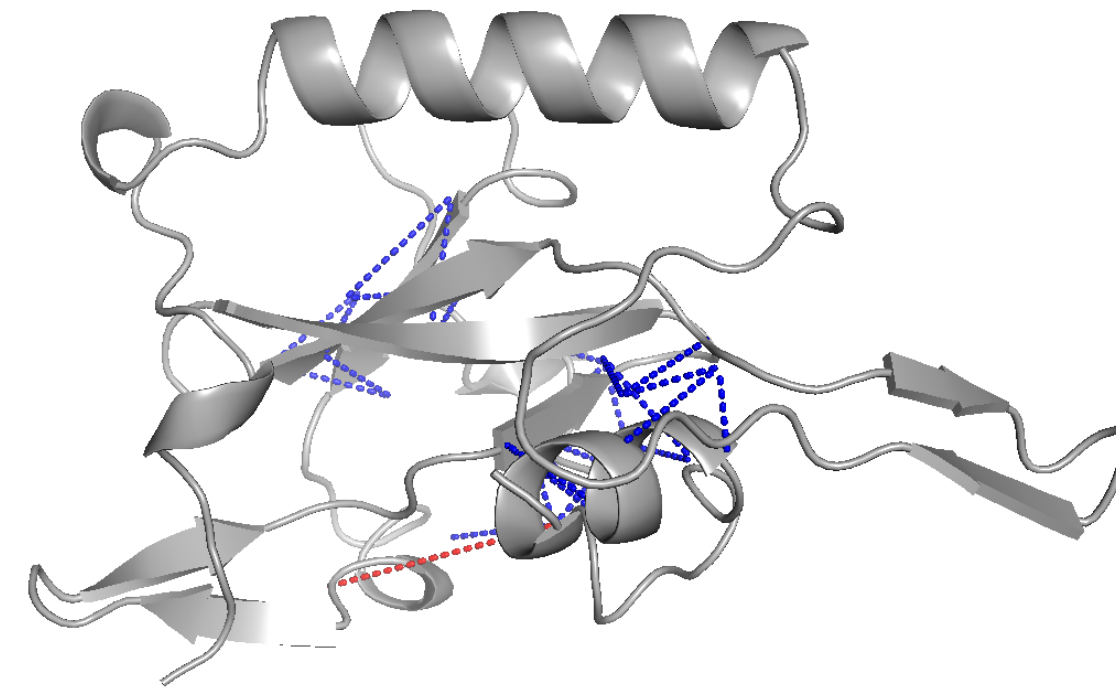
T1043-D1 - FM $\log(\text{Neff} / \text{len}) = 0.01$



Long L5 contacts

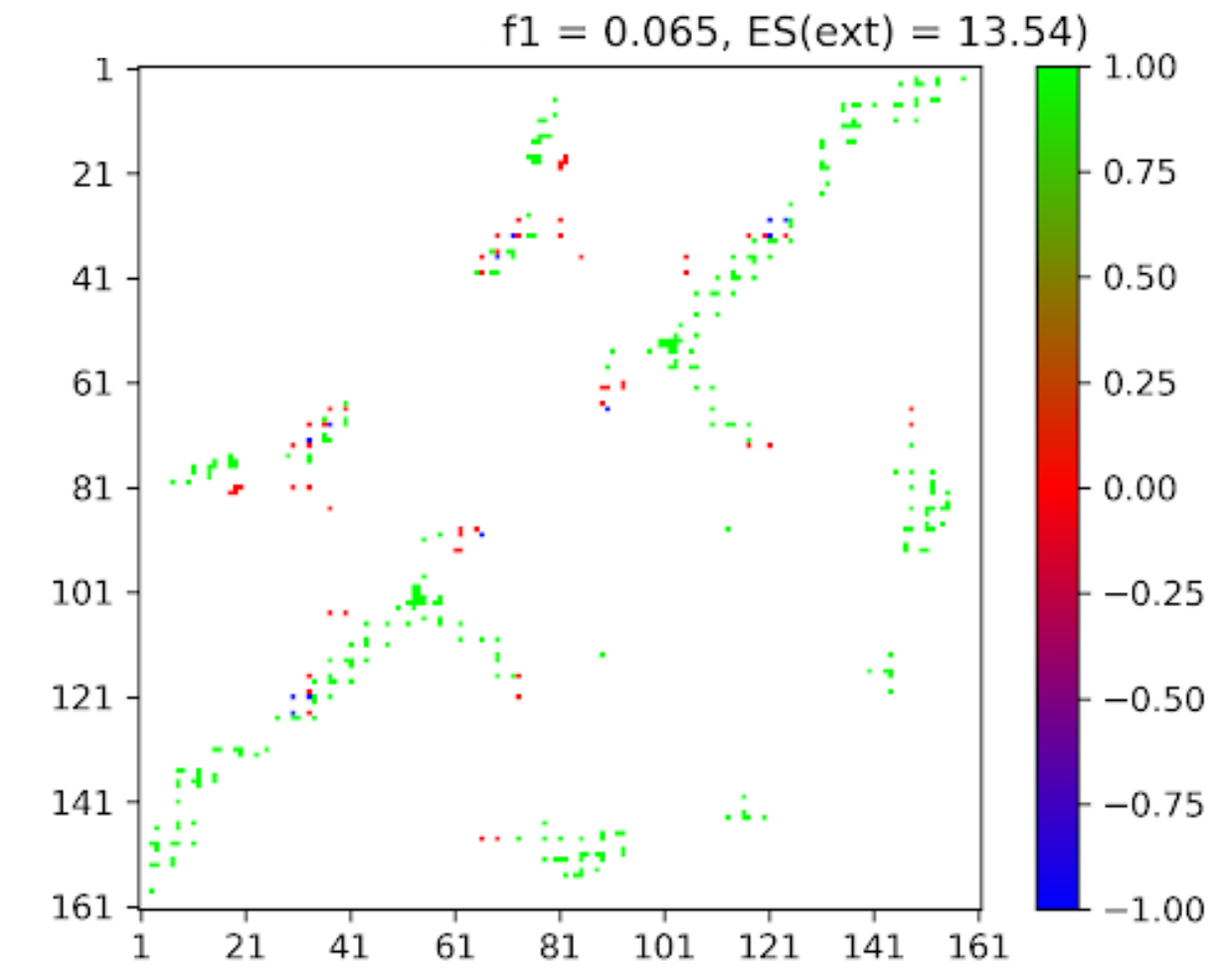
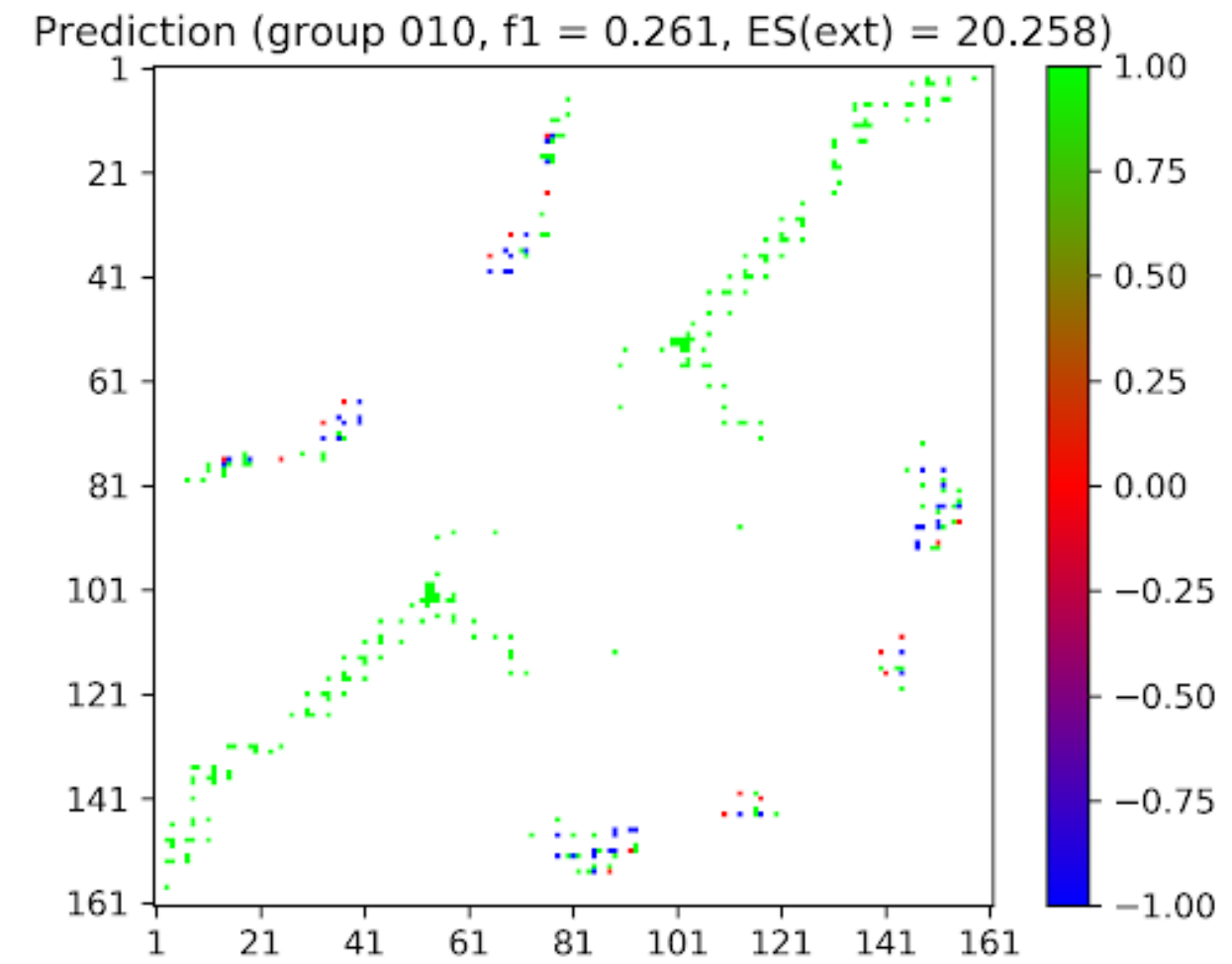
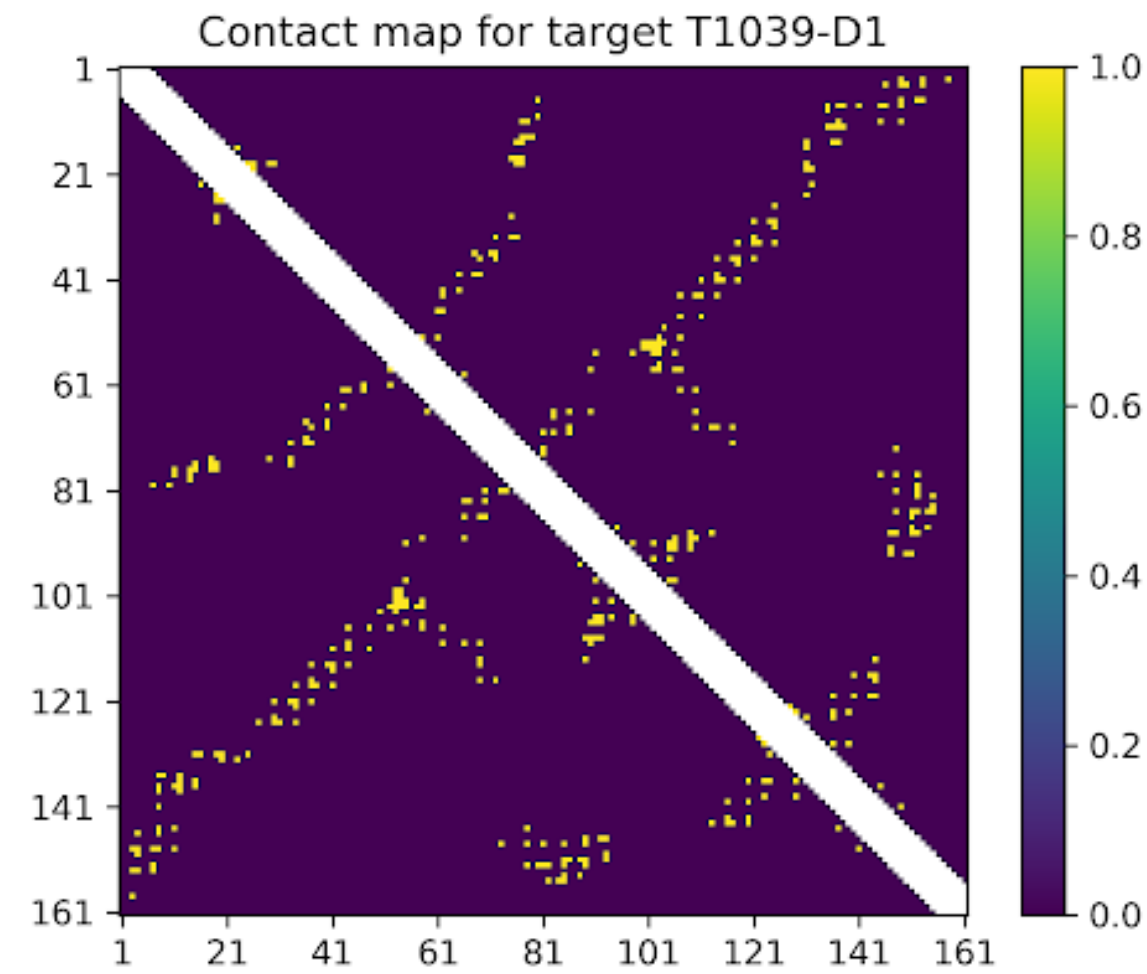
Average precision $\approx 10\%$

Top performing precision = 96.7%



Long L5 predictions Red = FP, Blue = TP

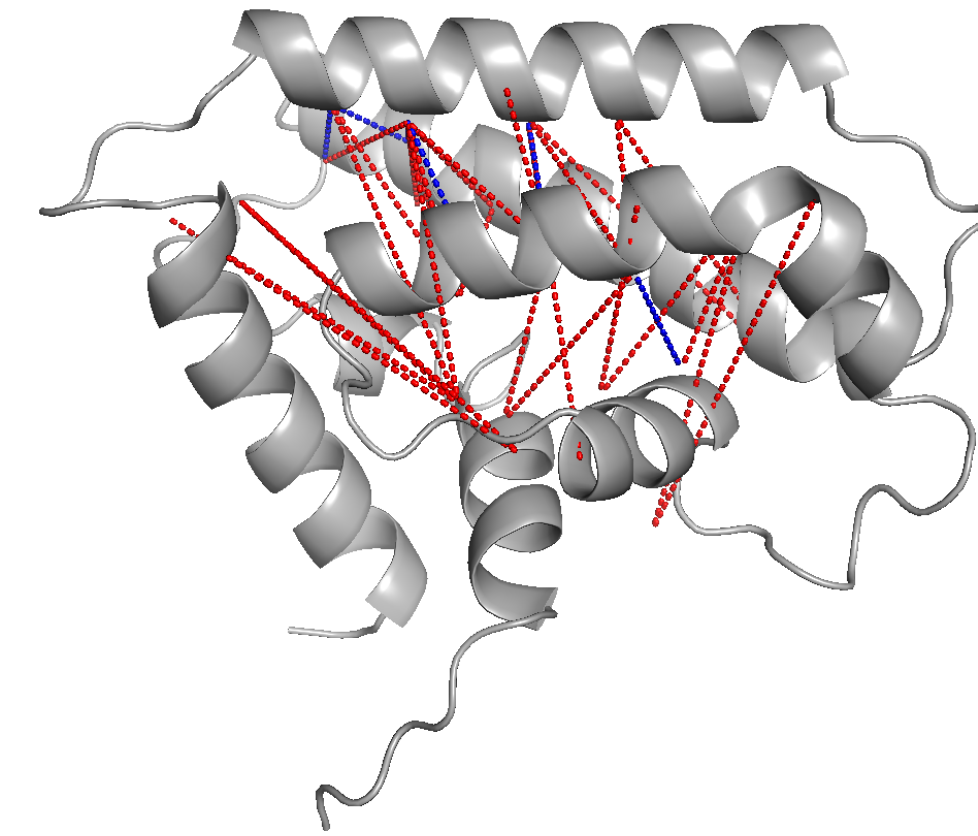
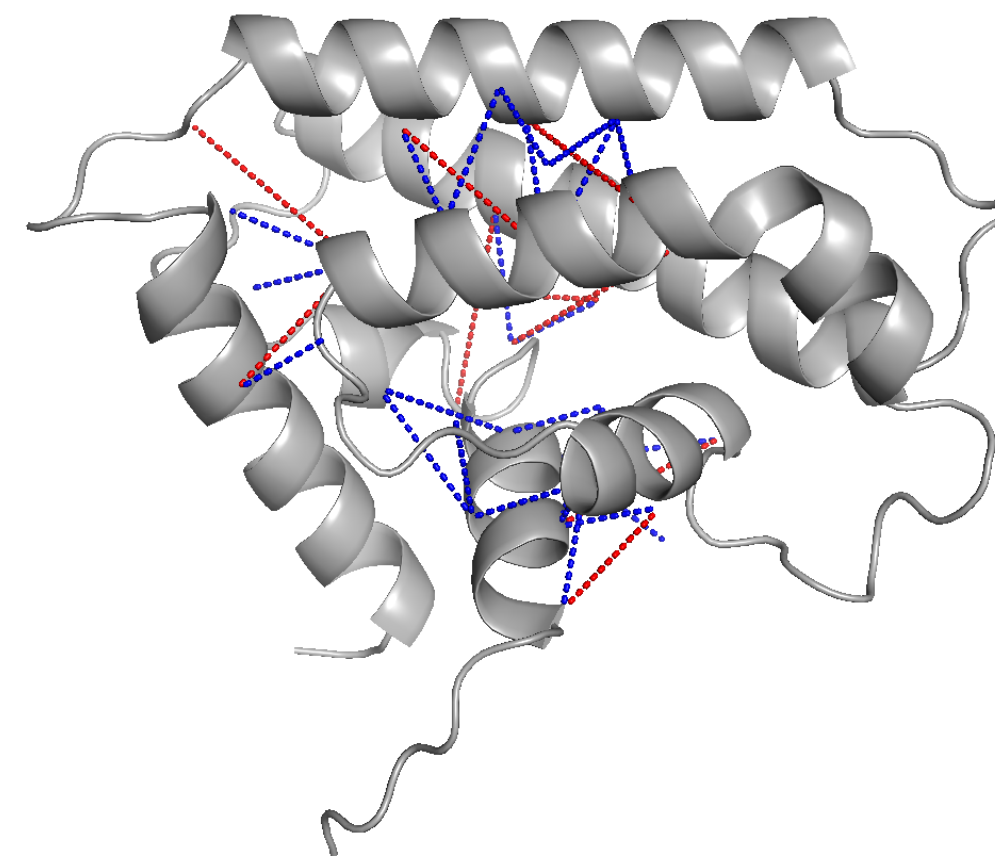
T1039-D1 - FM $\log(\text{Neff} / \text{len}) = 0.02$



Long L5 contacts

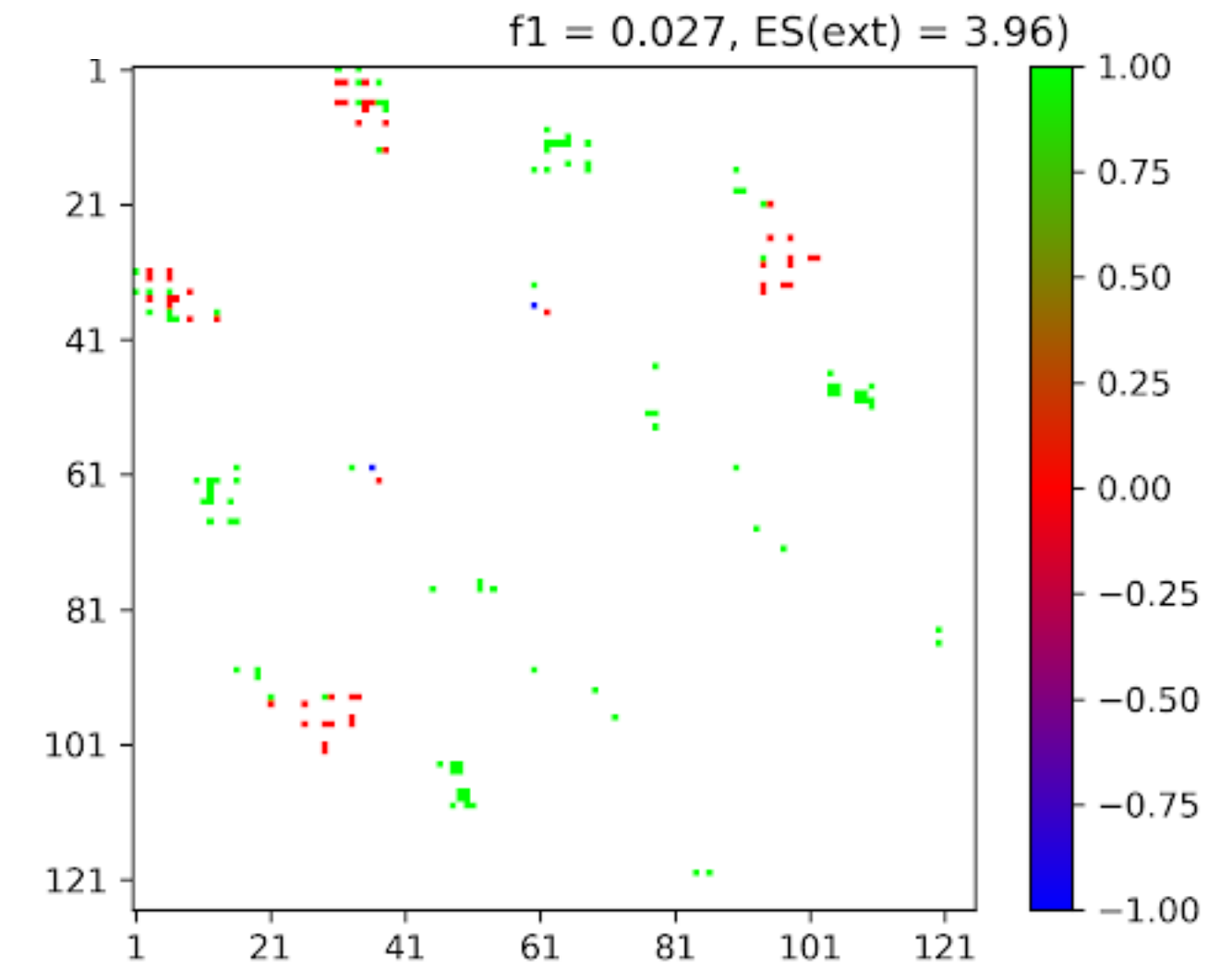
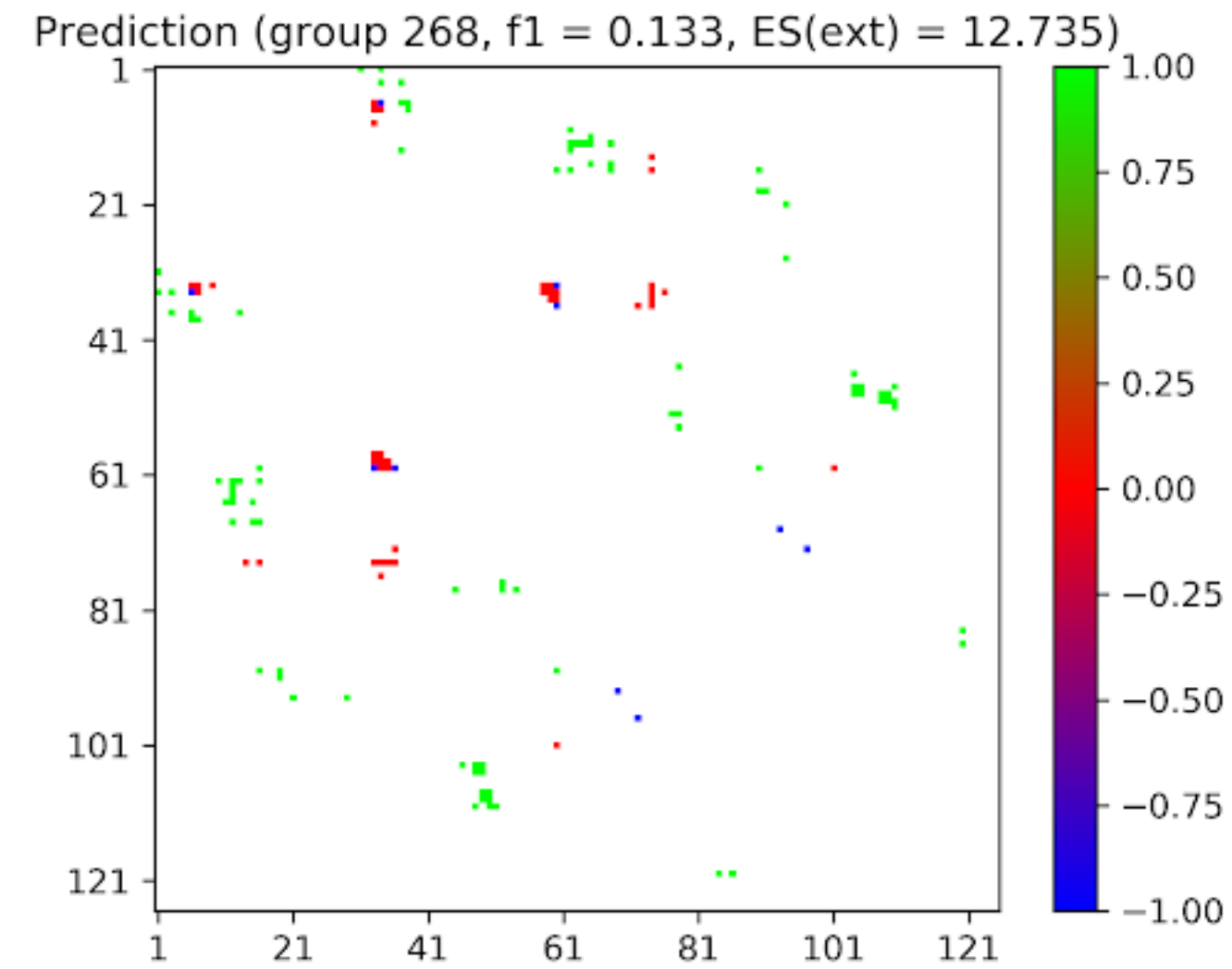
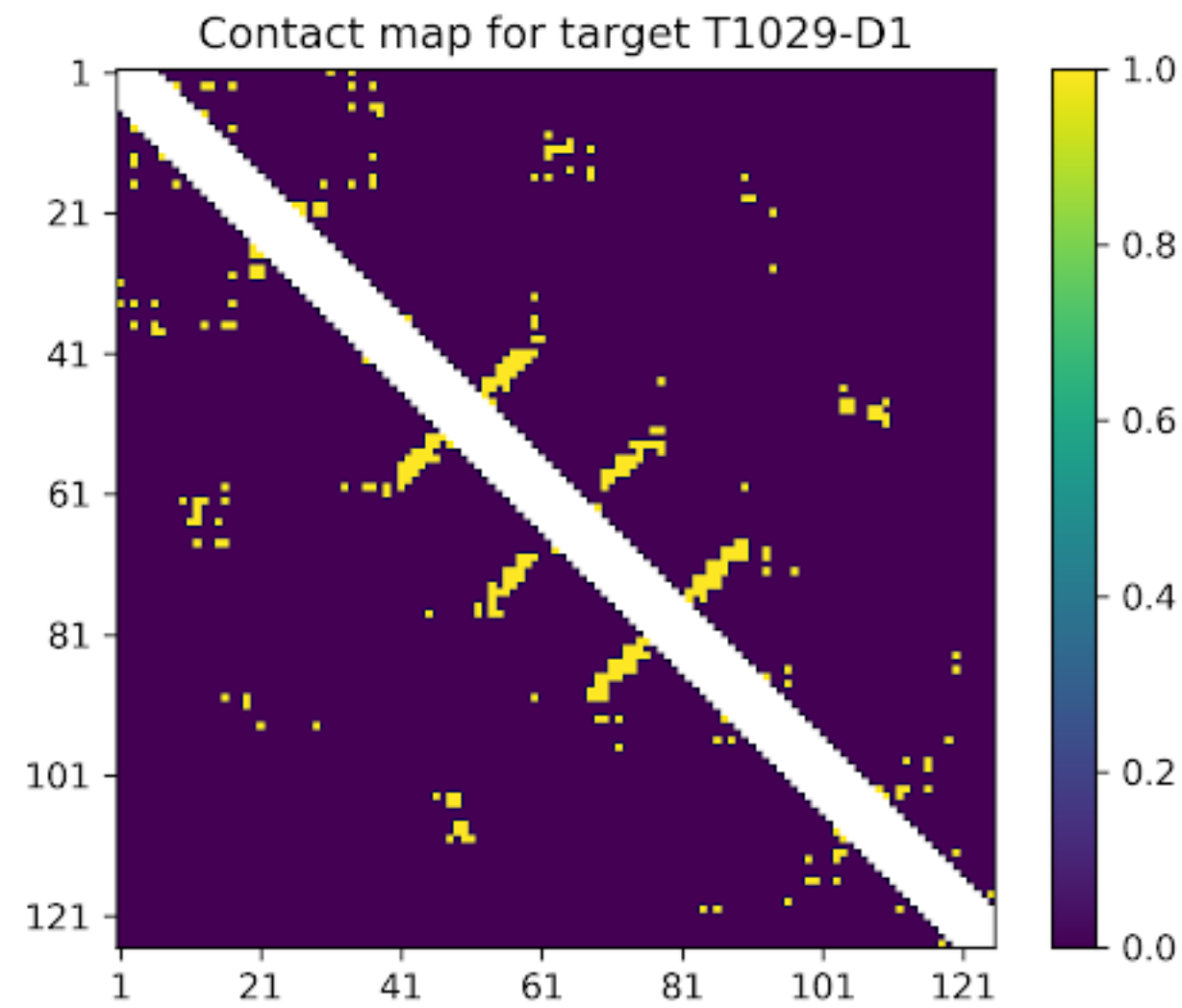
Average precision < 20%

Top performing precision = 72.7 %



Long L5 predictions Red = FP, Blue = TP

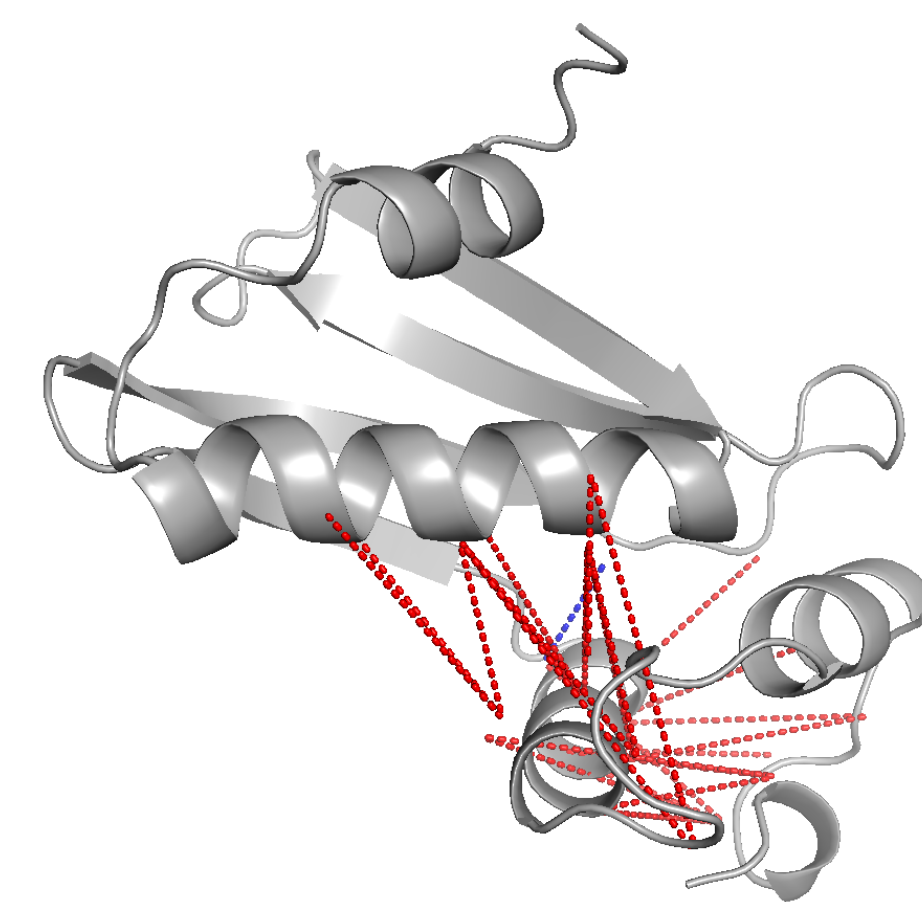
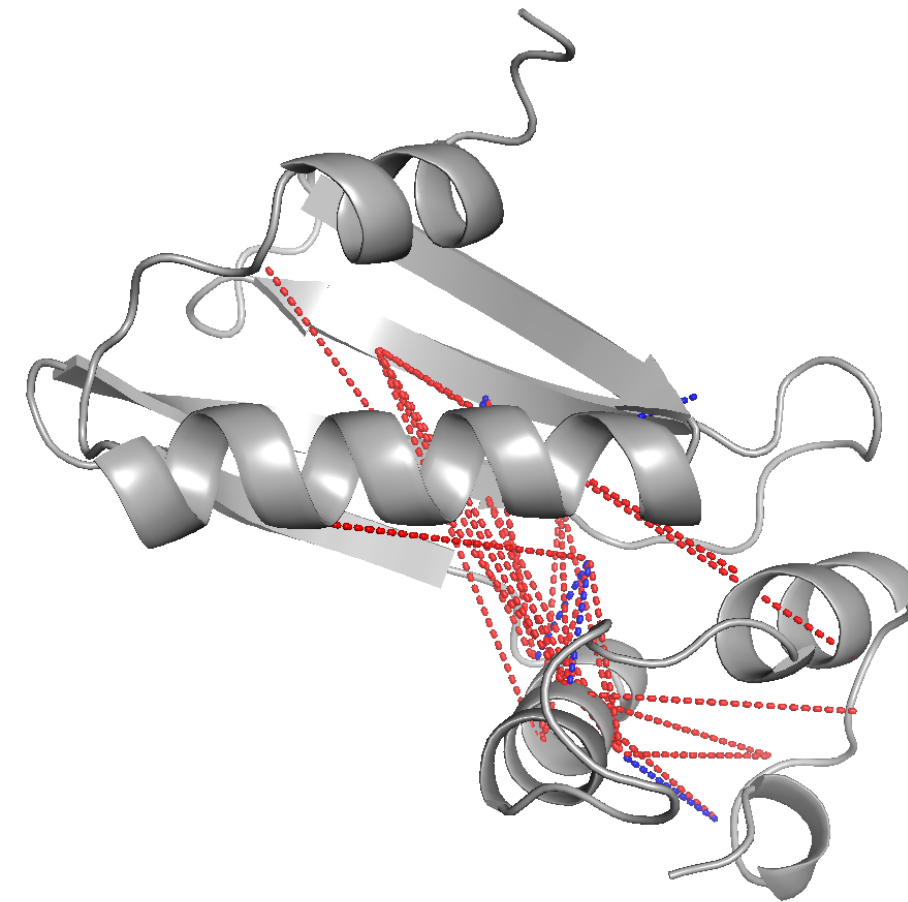
T1029-D1 - FM $\log(\text{Neff} / \text{len}) = 1.84$



Long L5 contacts

Average precision 2.8% (sd ≈ 4.2)

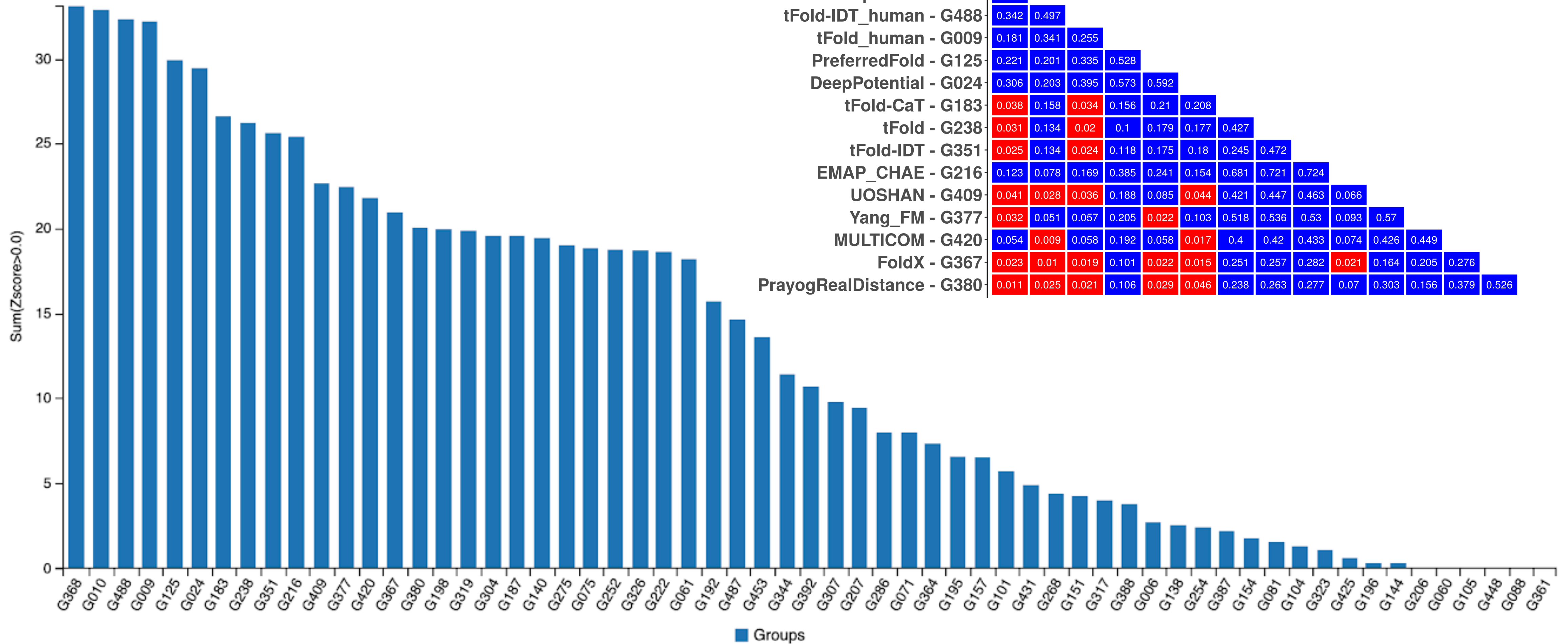
Top performing precision = 20%



Long L5 predictions Red = FP, Blue = TP

Ranking - sum(z-score > 0)

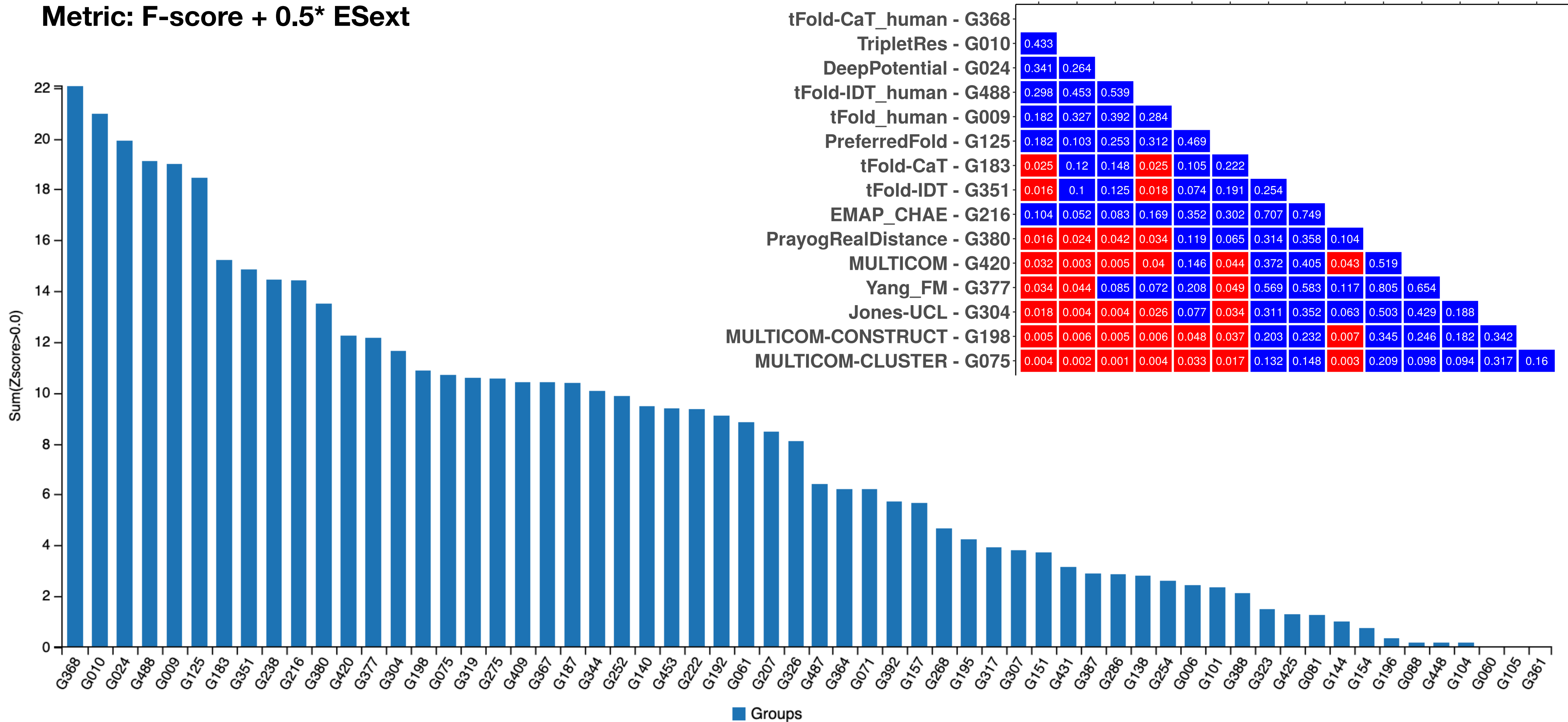
Metric: F-score



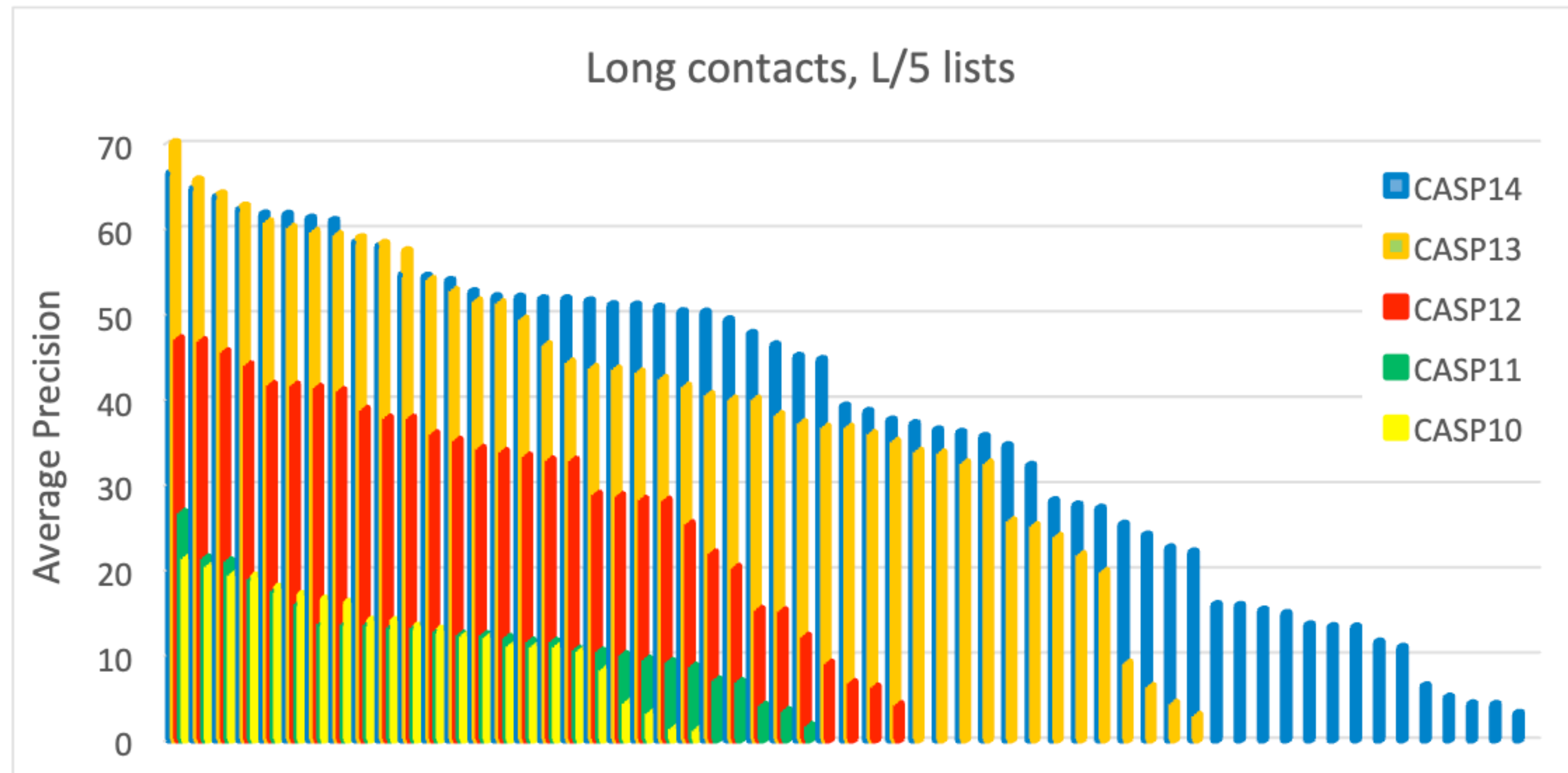
Head to head & paired t-test

	G368	G010	G488	G009	G125	G024	G183	G238	G351	G216	G409	G377	G420	G367	G380
tFold-CaT_human - G368															
TripletRes - G010	0.413														
tFold-IDT_human - G488	0.342	0.497													
tFold_human - G009	0.181	0.341	0.255												
PreferredFold - G125	0.221	0.201	0.335	0.528											
DeepPotential - G024	0.306	0.203	0.395	0.573	0.592										
tFold-CaT - G183	0.038	0.158	0.034	0.156	0.21	0.208									
tFold - G238	0.031	0.134	0.02	0.1	0.179	0.177	0.427								
tFold-IDT - G351	0.025	0.134	0.024	0.118	0.175	0.18	0.245	0.472							
EMAP_CHAE - G216	0.123	0.078	0.169	0.385	0.241	0.154	0.681	0.721	0.724						
UOSHAN - G409	0.041	0.028	0.036	0.188	0.085	0.044	0.421	0.447	0.463	0.066					
Yang_FM - G377	0.032	0.051	0.057	0.205	0.022	0.103	0.518	0.536	0.53	0.093	0.57				
MULTICOM - G420	0.054	0.009	0.058	0.192	0.058	0.017	0.4	0.42	0.433	0.074	0.426	0.449			
FoldX - G367	0.023	0.01	0.019	0.101	0.022	0.015	0.251	0.257	0.282	0.021	0.164	0.205	0.276		
PrayogRealDistance - G380	0.011	0.025	0.021	0.106	0.029	0.046	0.238	0.263	0.277	0.07	0.303	0.156	0.379	0.526	

Ranking - sum(z-score > 0)

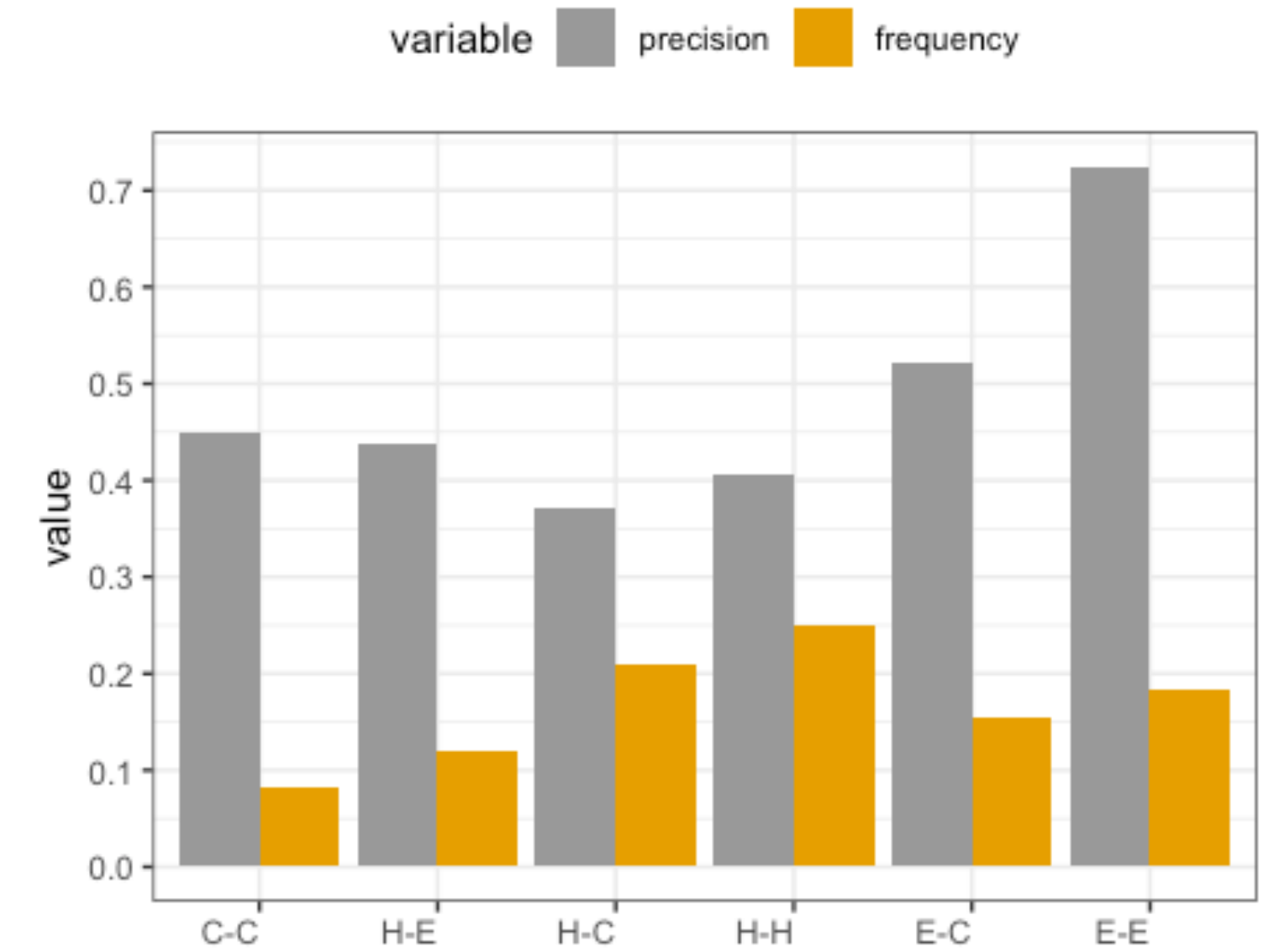
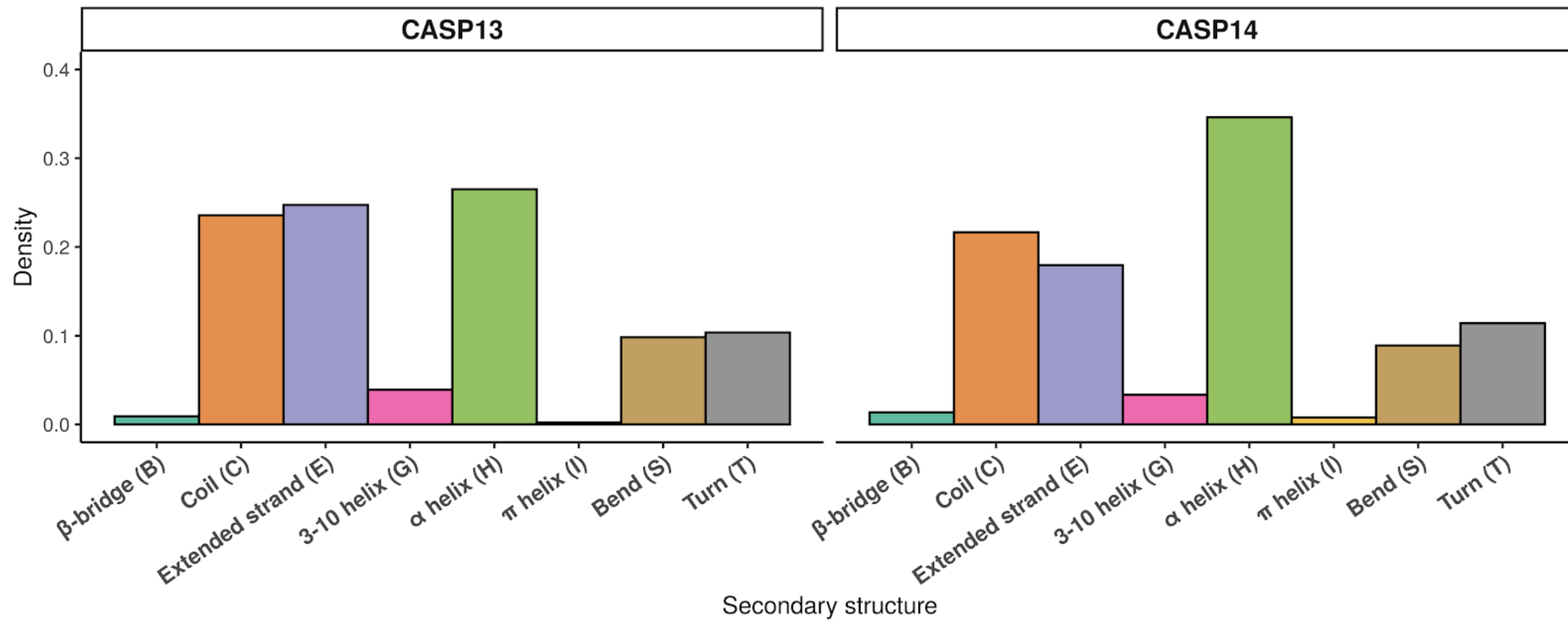


Progress with respect to CASP10-13



- No difference in performance over top 10 groups
- Increased # of predictions at 50% precision
- Increased # of participants

CASP14 vs CASP13: Secondary structure content



Precision vs SS type - FL p > 0.5

Results in distance prediction

Distance assessment

Prediction format

List of contacts in 13 columns format:

- **i j pN**
- **i** and **j** are the indices of two the amino acids
- **pN**: C-beta distance within boundaries of the N-th bin: p1 ($d \leq 4$), p2 ($4 \leq d \leq 6$), p3 ($6 \leq d \leq 8$) p10 ($d > 20$)

Evaluation:

Prediction are trimmed to domains

i-j pairs excluded if sequence space separation is < 6 aa

Non-listed aa pairs are assigned $p_{10} = 1$

Metrics

Precision, Recall and F over each bin:

- TP, FP and FN computed over binarised vector and $\text{Max}(pN)$

$$\text{Mean Distance Difference} = 1 - \left(\frac{1}{10} \sum_{k=1}^{10} \frac{1}{N_k} \sum_{a=1}^{N_k} \sum_{b=1}^{10} \frac{p_a(d_b) |D_k - d_b|}{D_{max}} \right)$$

$$\text{Mean Bin Neighbours} = \frac{1}{10} \sum_{k=1}^{10} \frac{1}{N_k} \sum_{b=1}^{N_k} \left(p_b(D_k) + \frac{p_b(d_{k-1}) + p_b(d_{k+1})}{2} \right)$$

Graph-based metrics

Strength

$$s_i = \sum_{j=1}^N a_{ij} w_{ij}$$

Clustering Coefficient

$$c_i^w = \frac{1}{s_i(k_i - 1)} \sum_{j,h} \frac{(w_{ij} + w_{ih})}{2} a_{ij} a_{ih} a_{jh}$$

Average Shortest Path

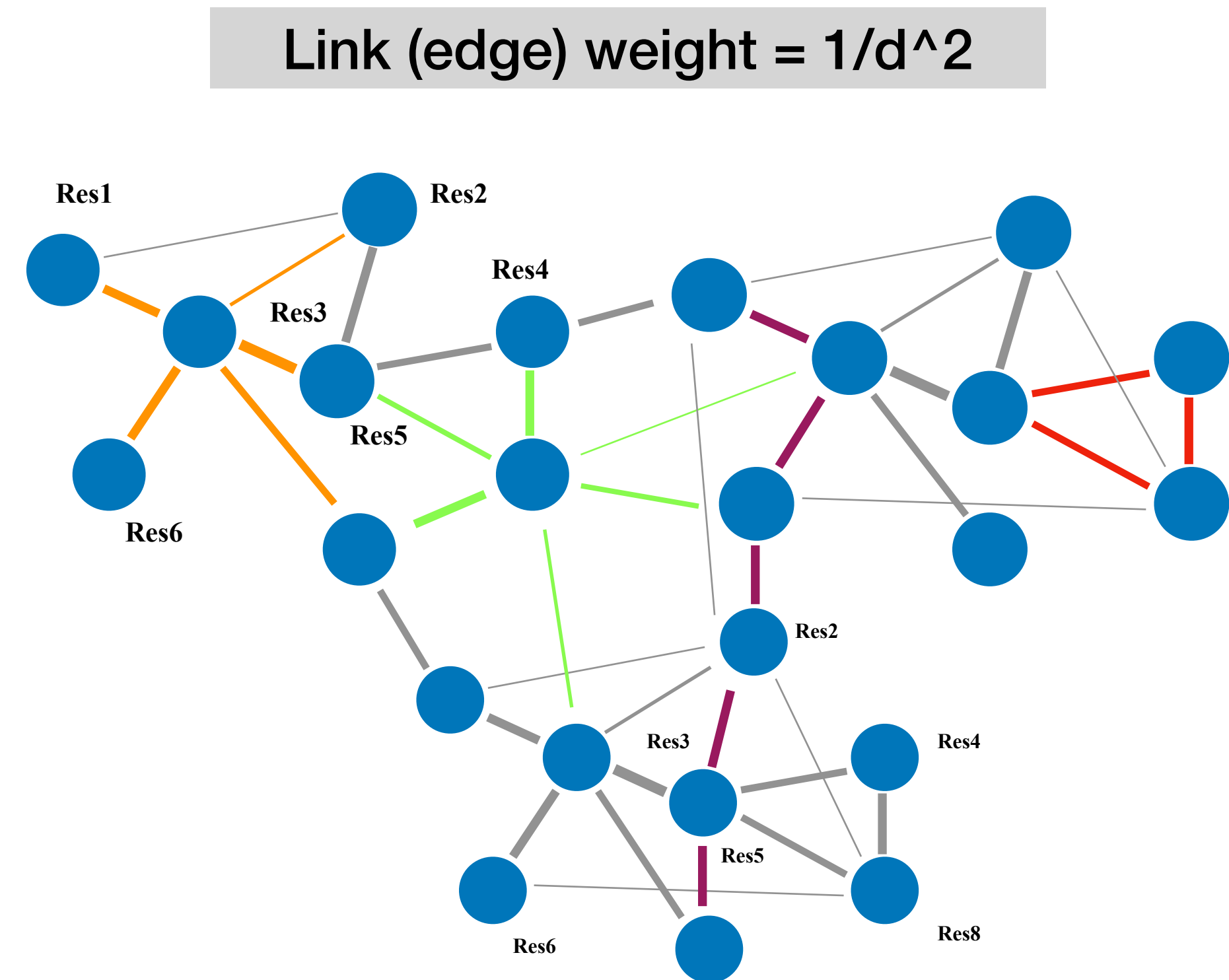
$$\overline{sp}_i = \frac{1}{N} \sum_{j=1}^N sp_j$$

$$sp_{i \rightarrow j} = P(v_1, \dots, v_i, \dots, v_n) \mid P = \min \left(\sum_{i=1}^{n-1} f(w_{i,i+1}) \right)$$

Diversity

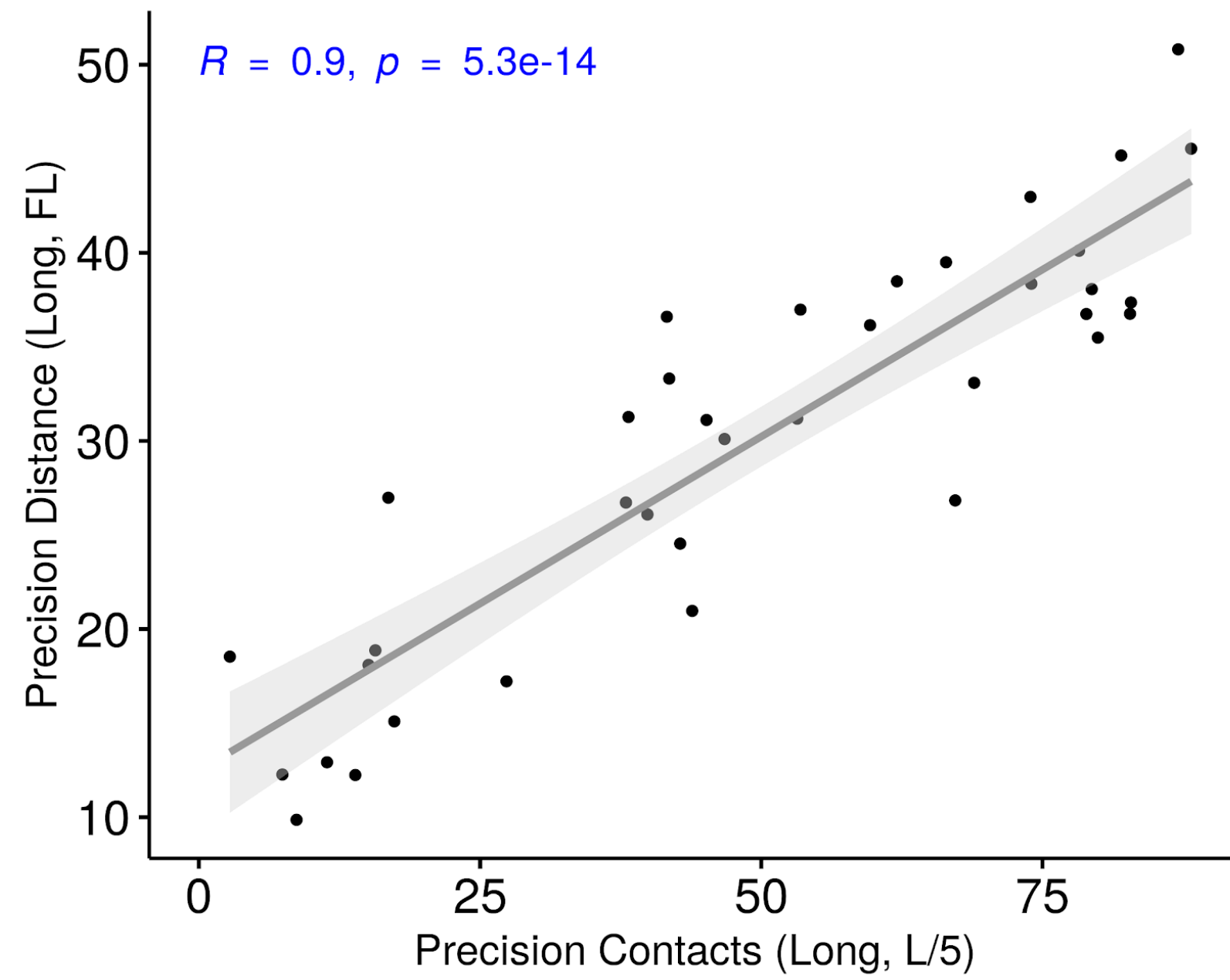
$$D(i) = \frac{H(i)}{\log(k_i)}$$

$H(i)$ is the Shannon entropy

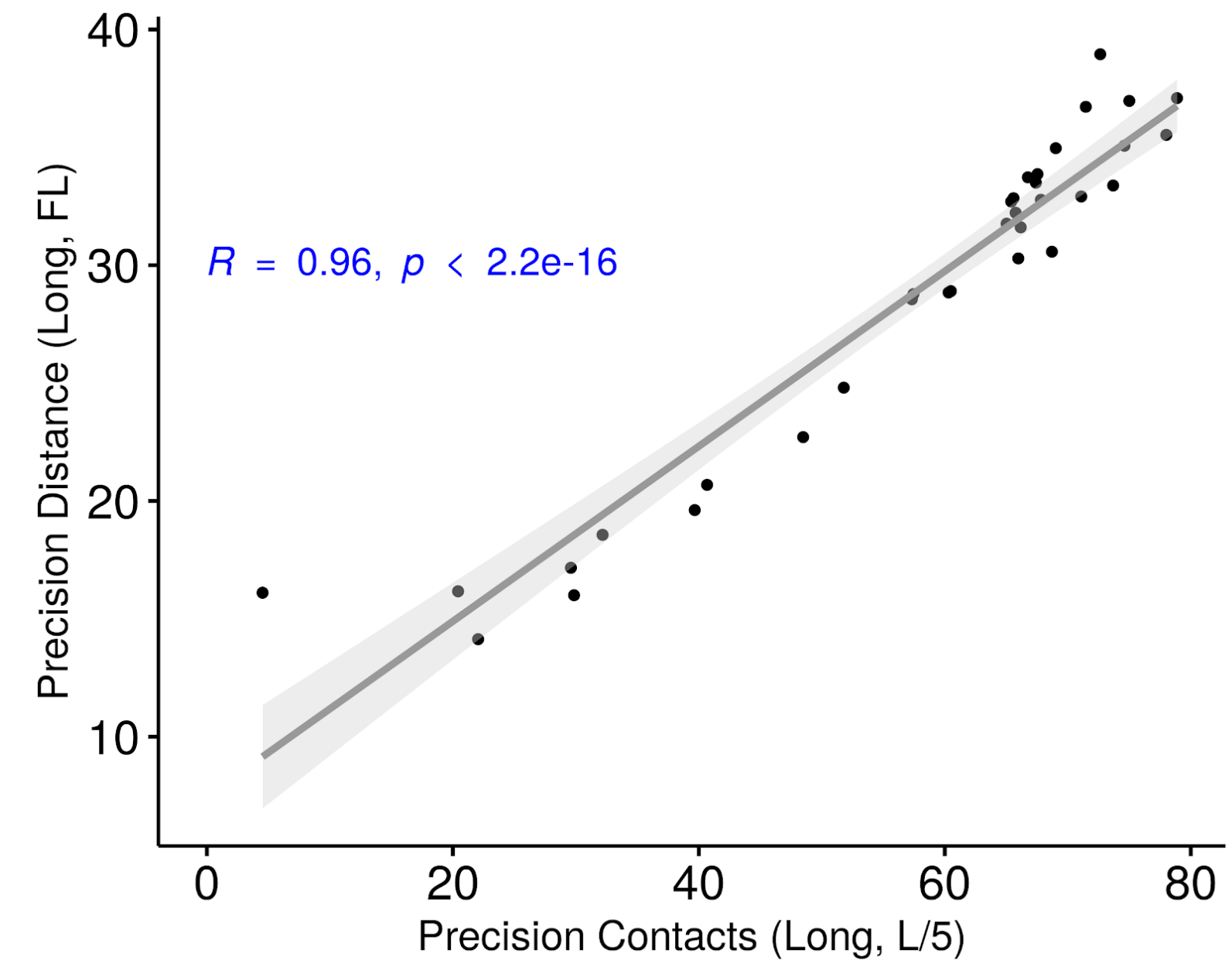


Distance vs Contacts (average performance)

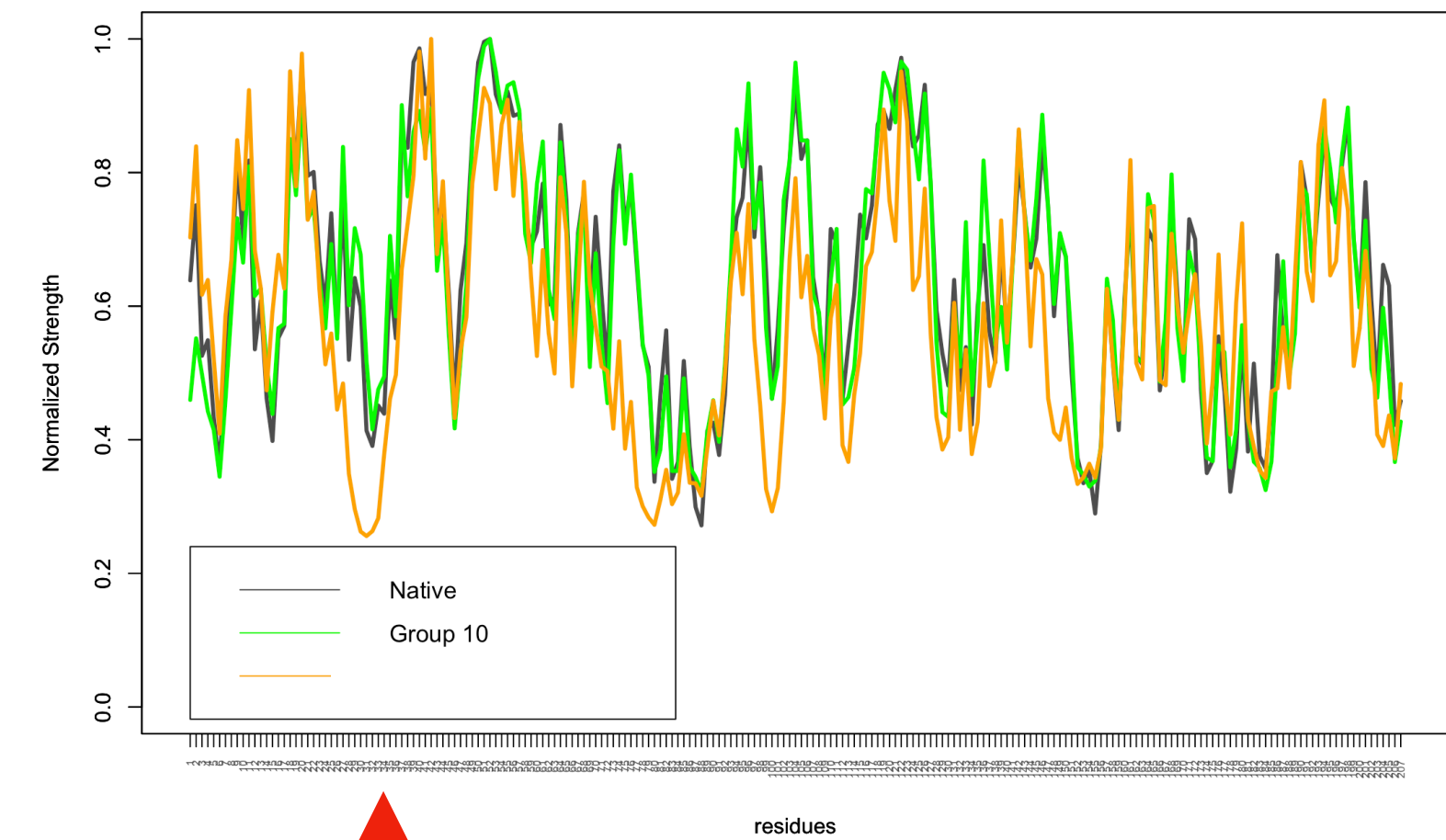
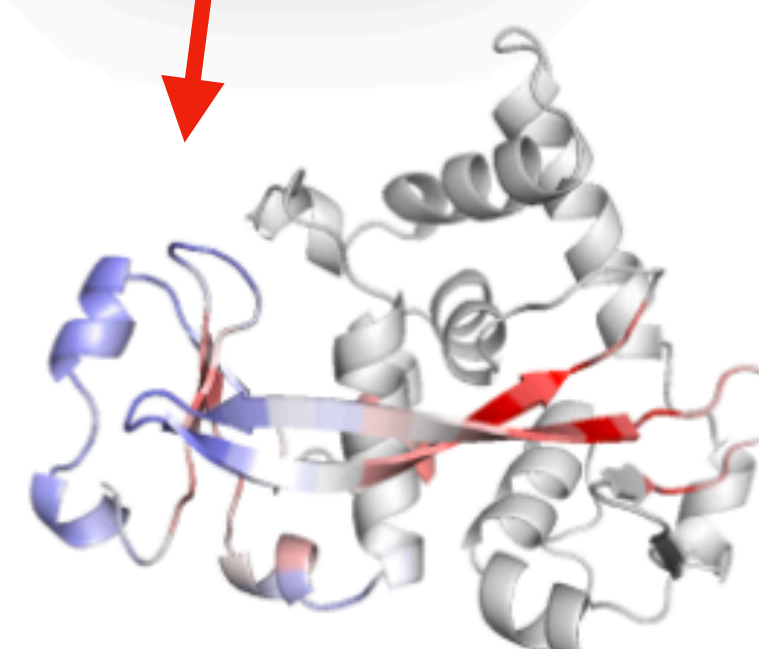
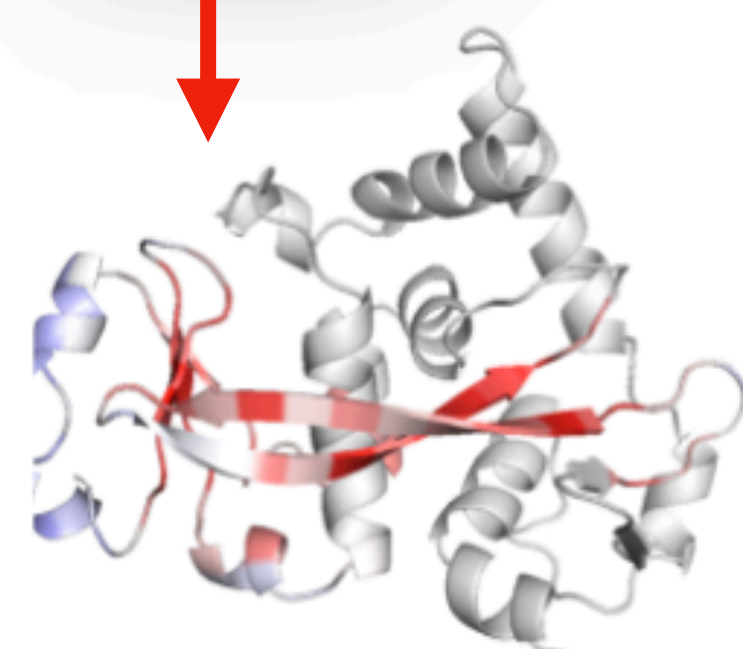
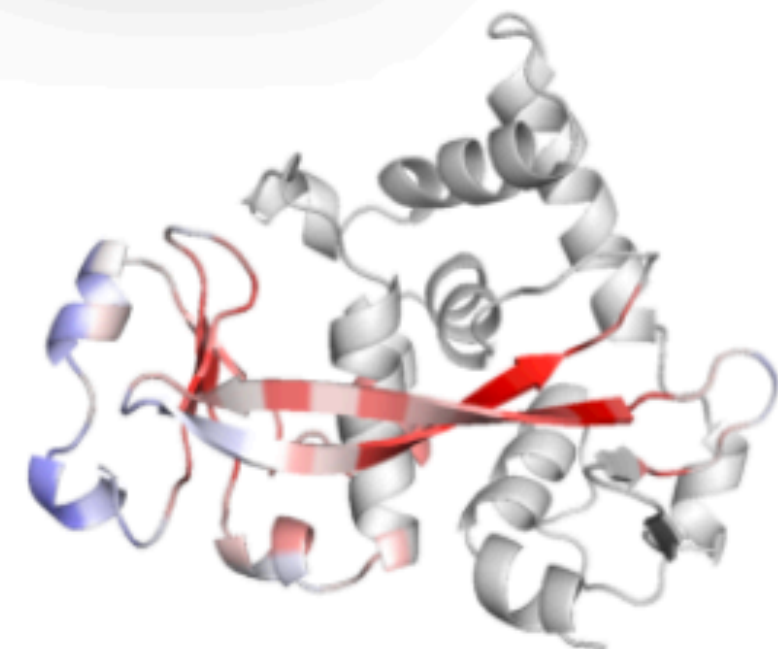
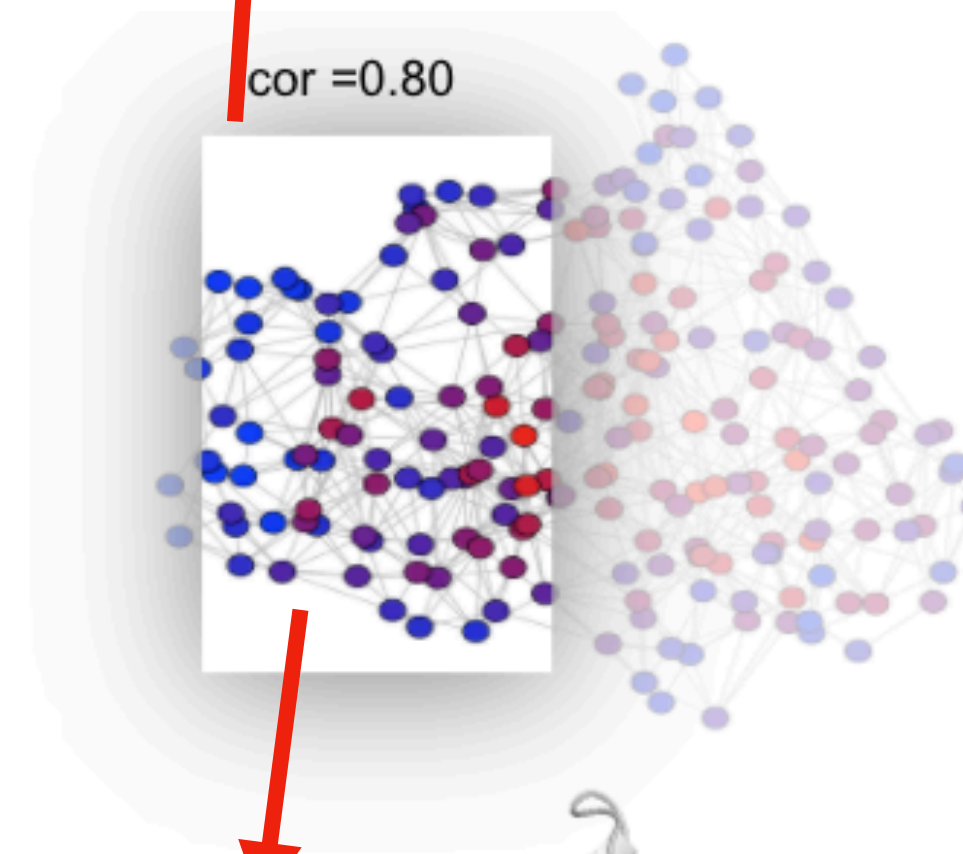
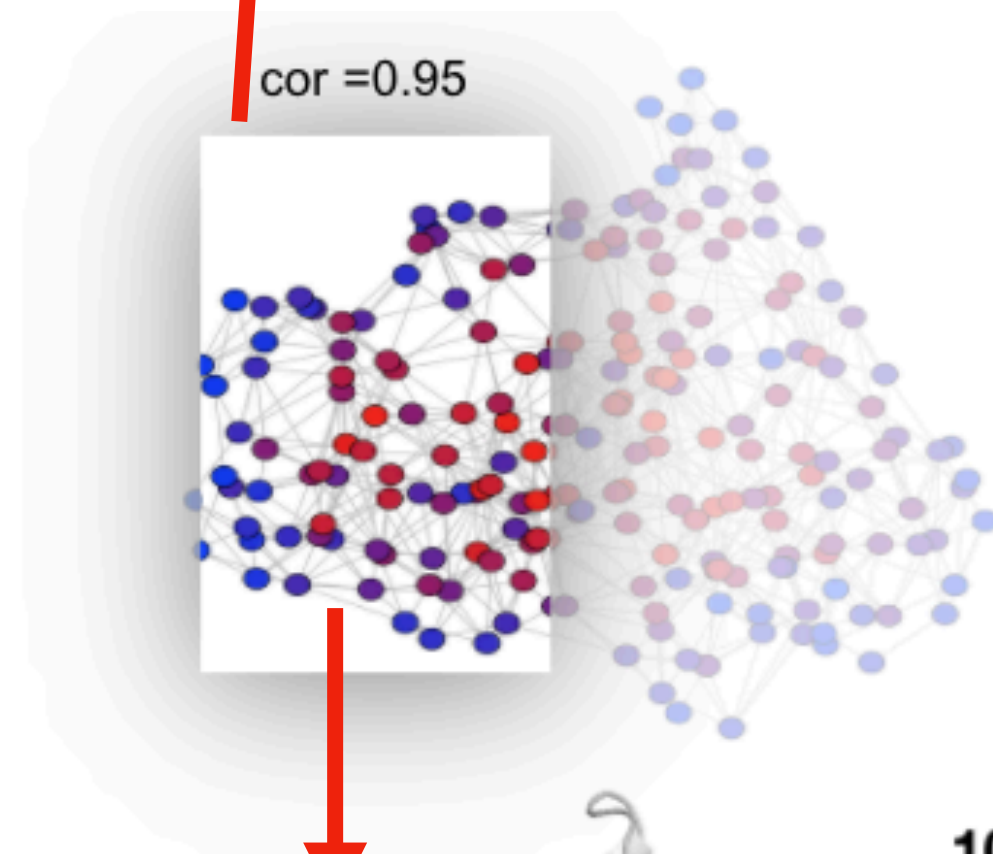
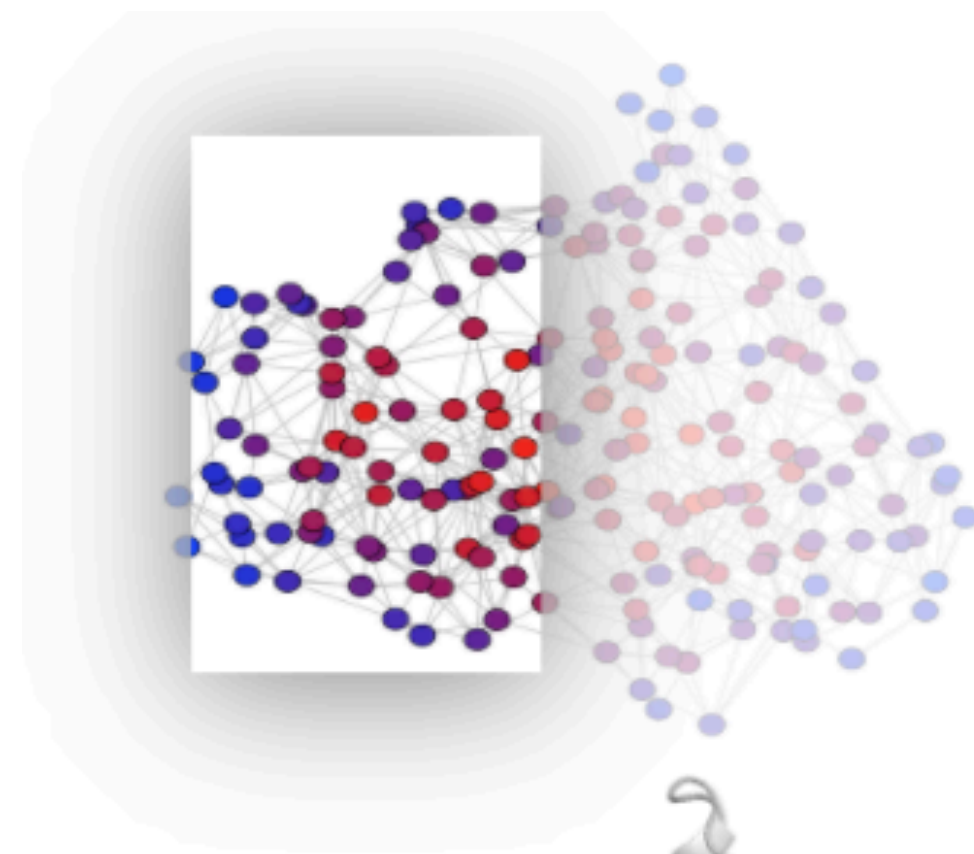
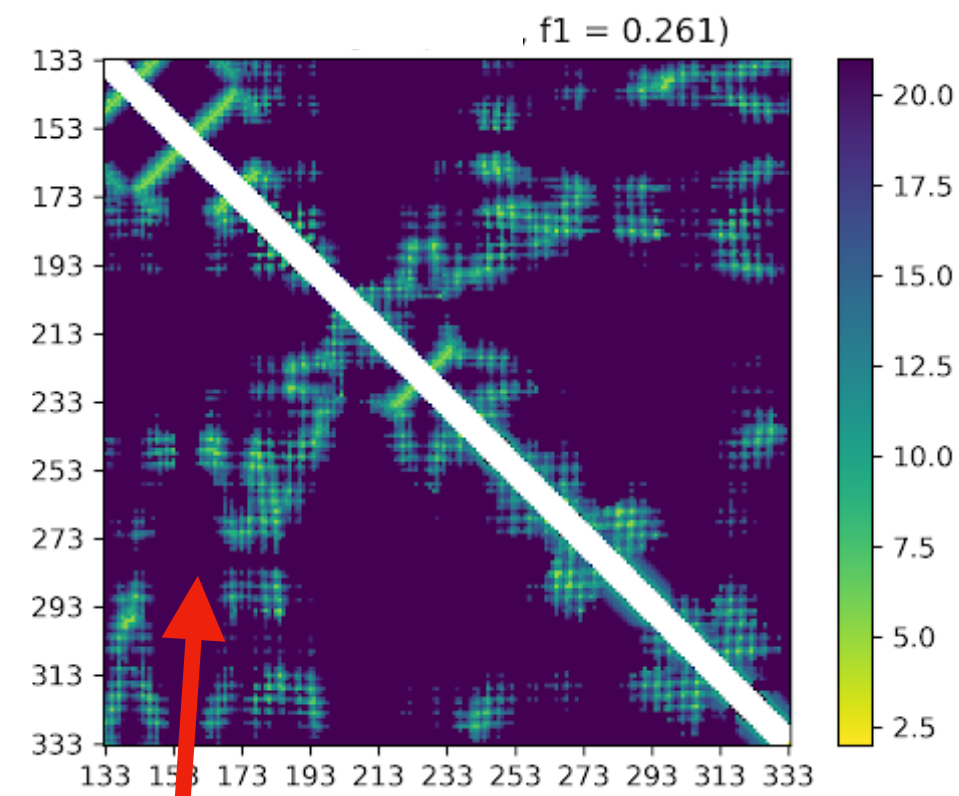
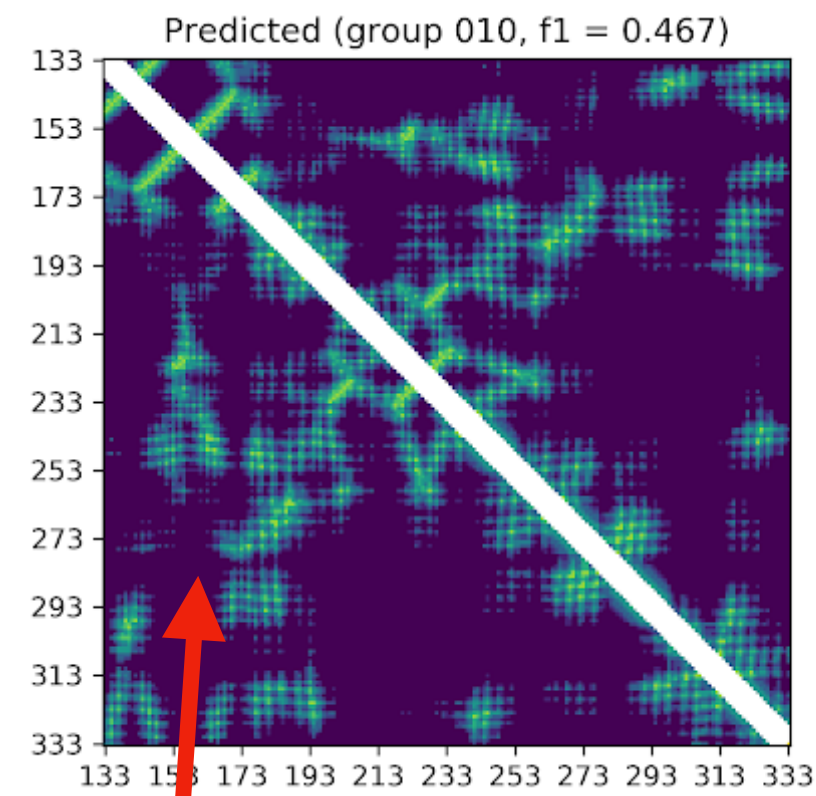
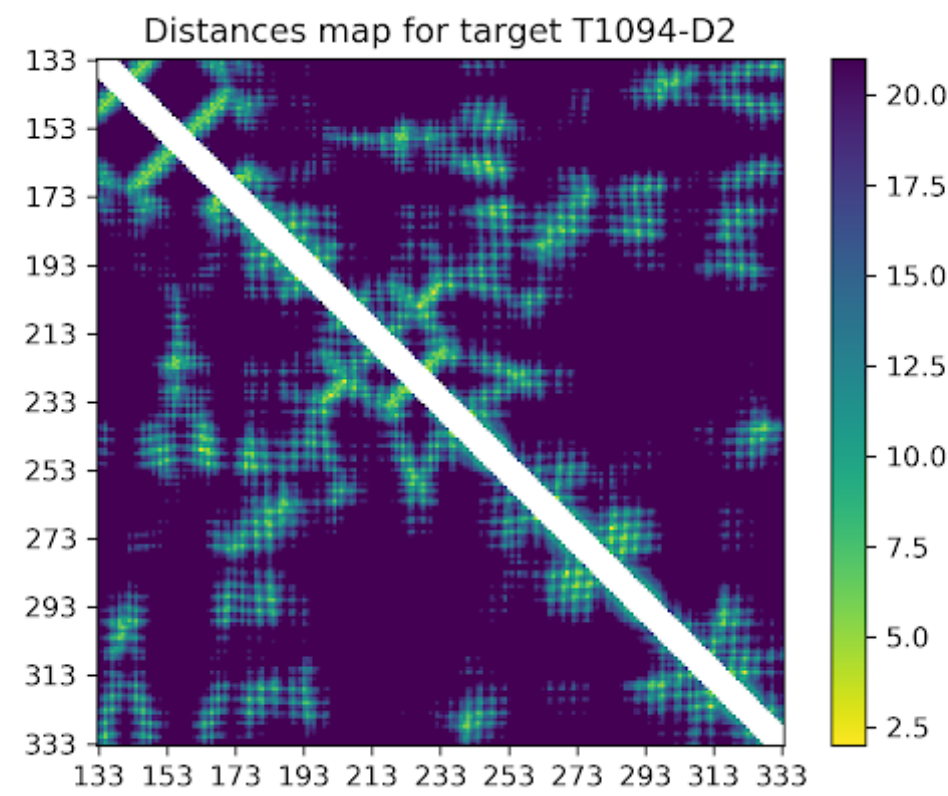
Targets



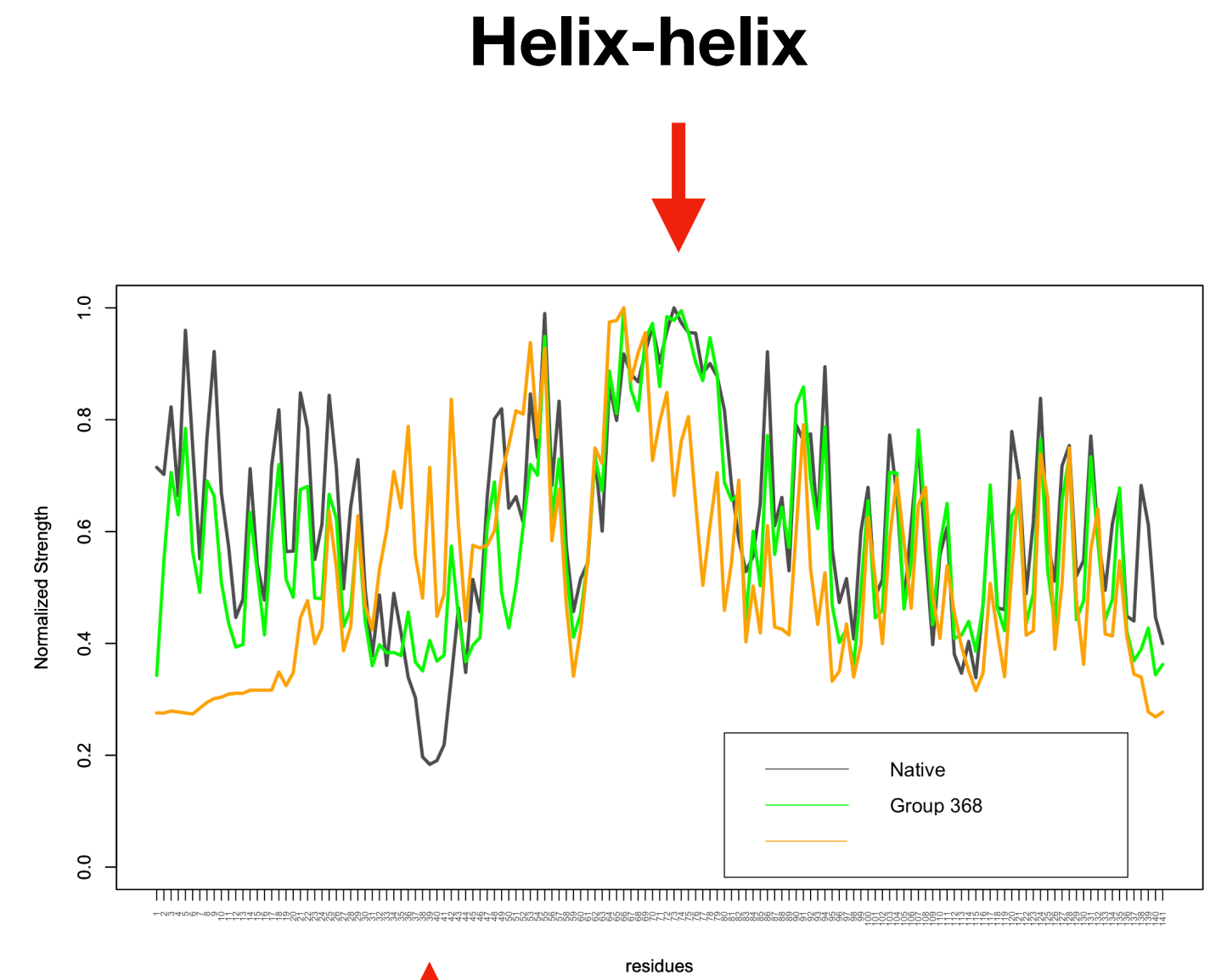
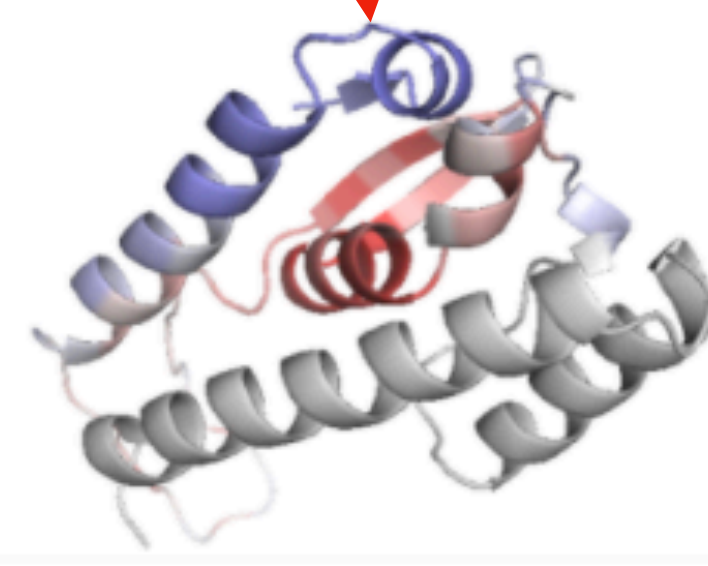
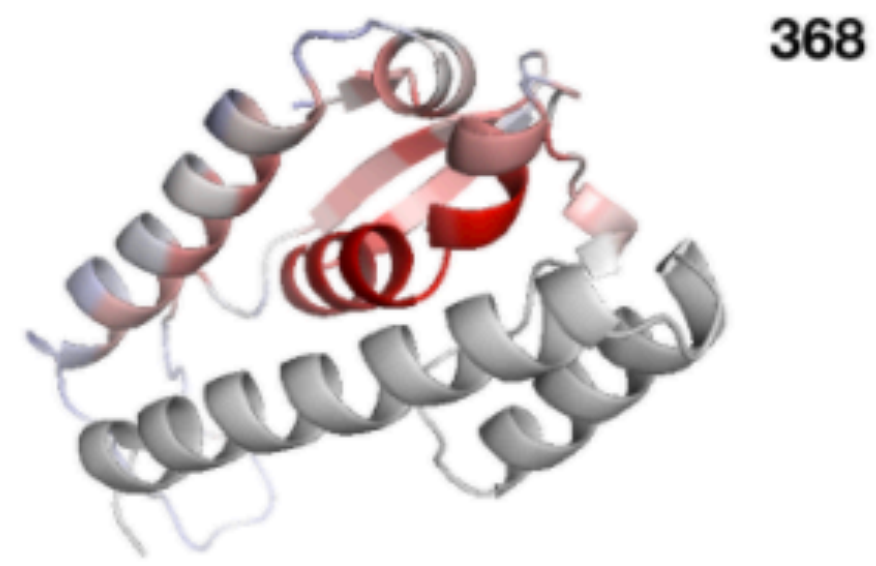
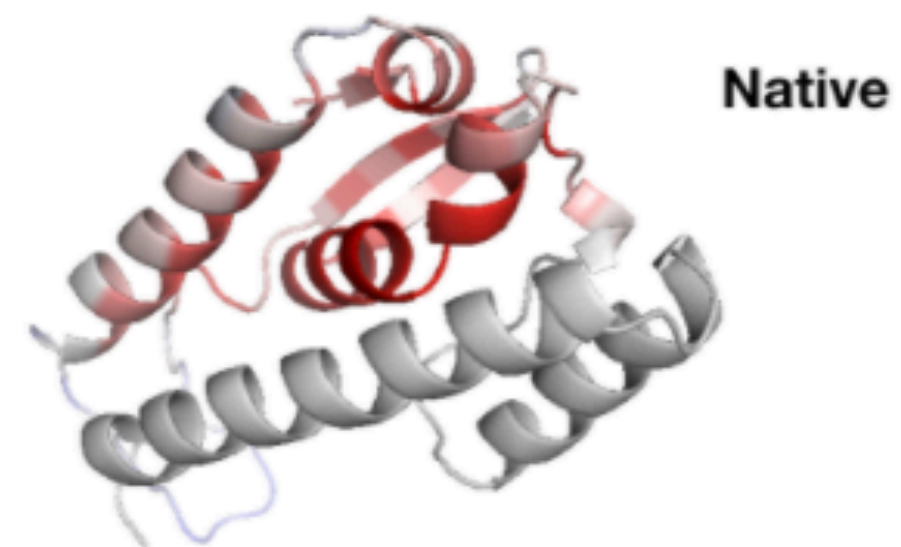
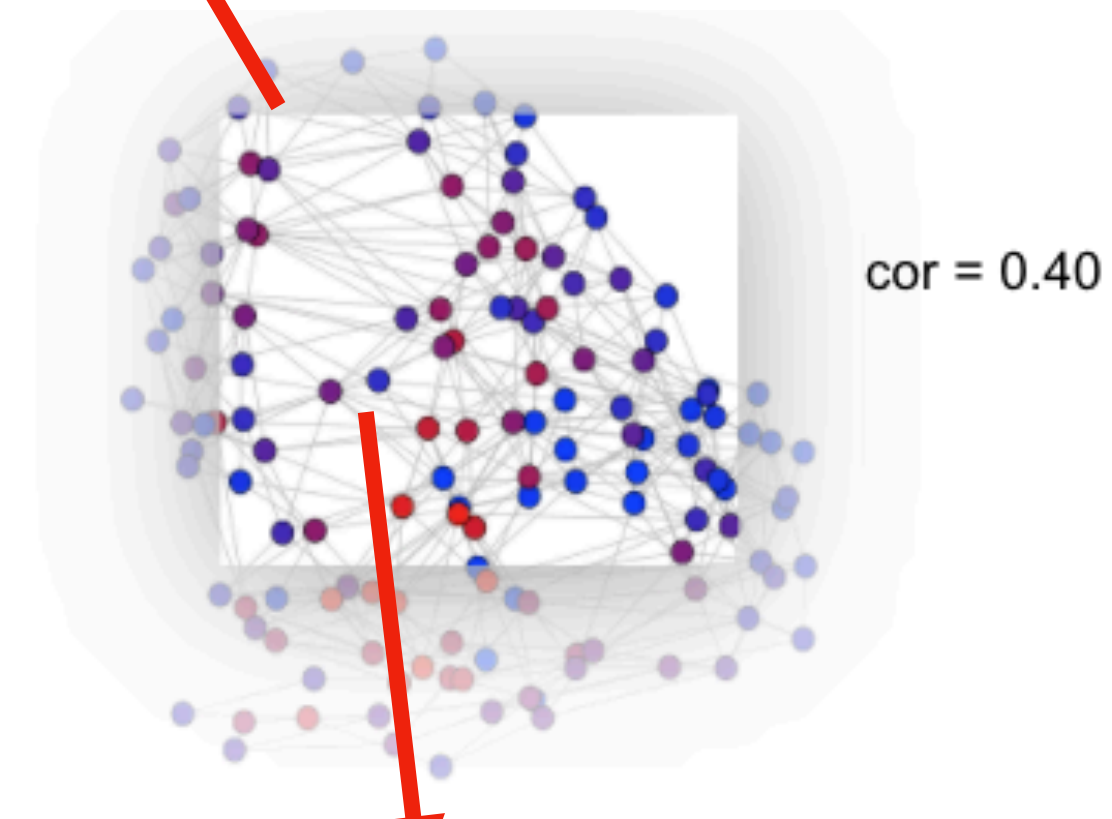
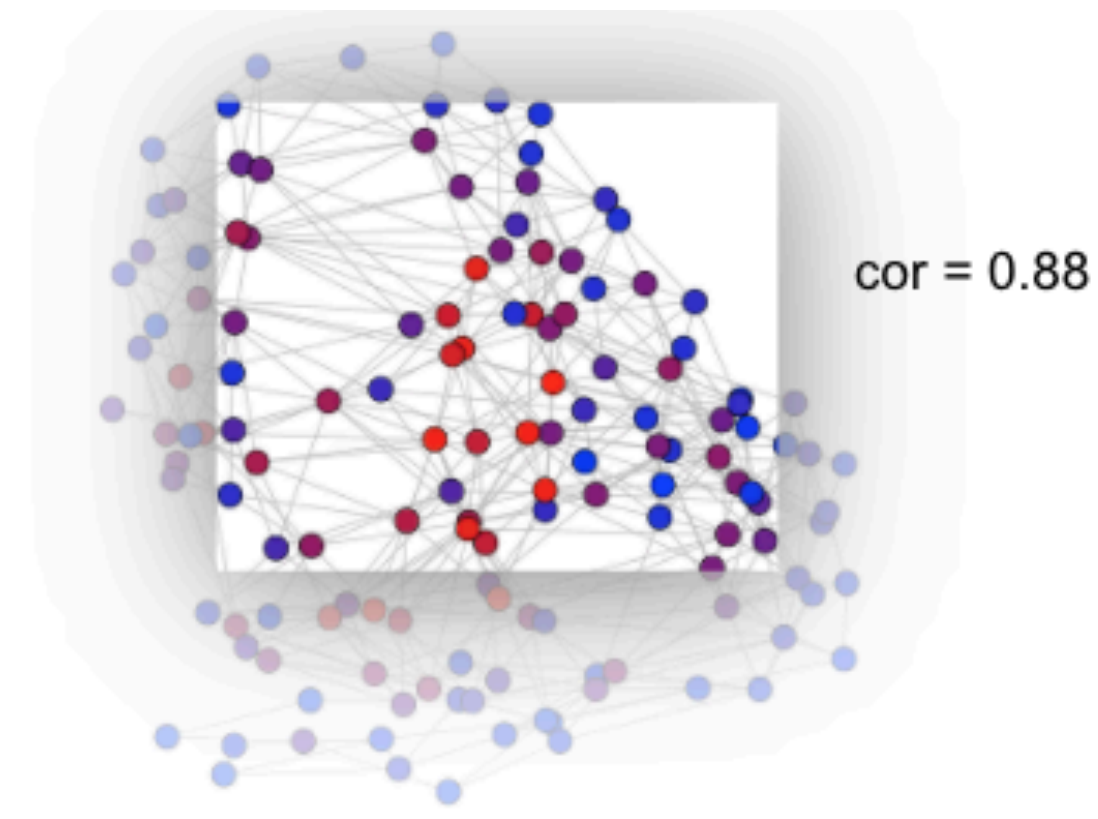
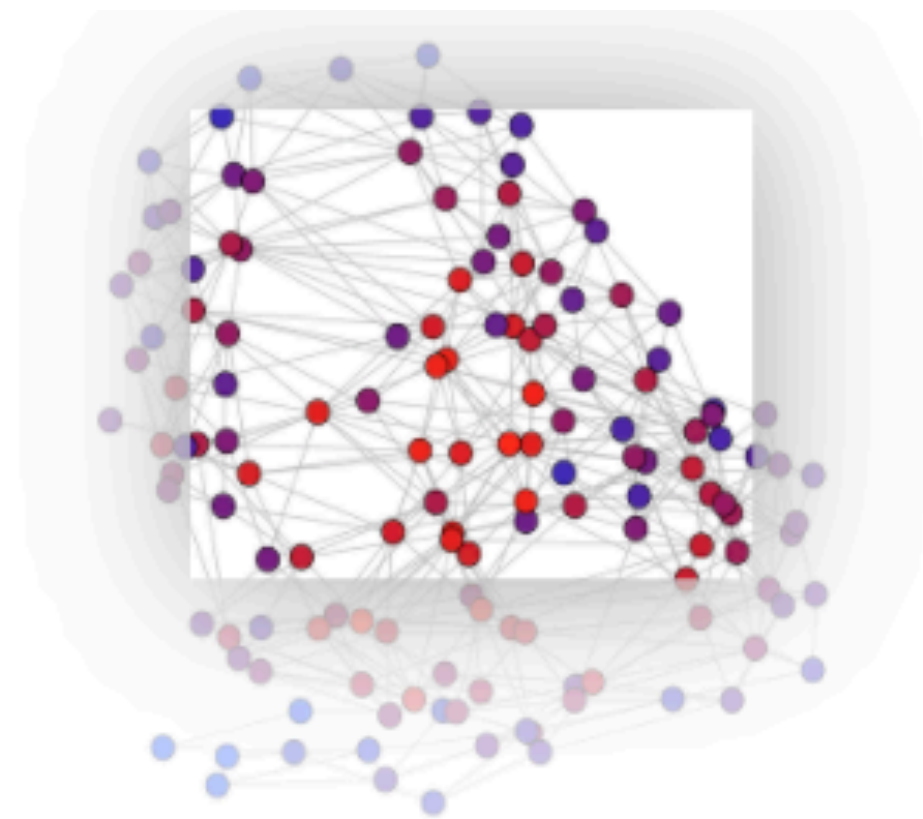
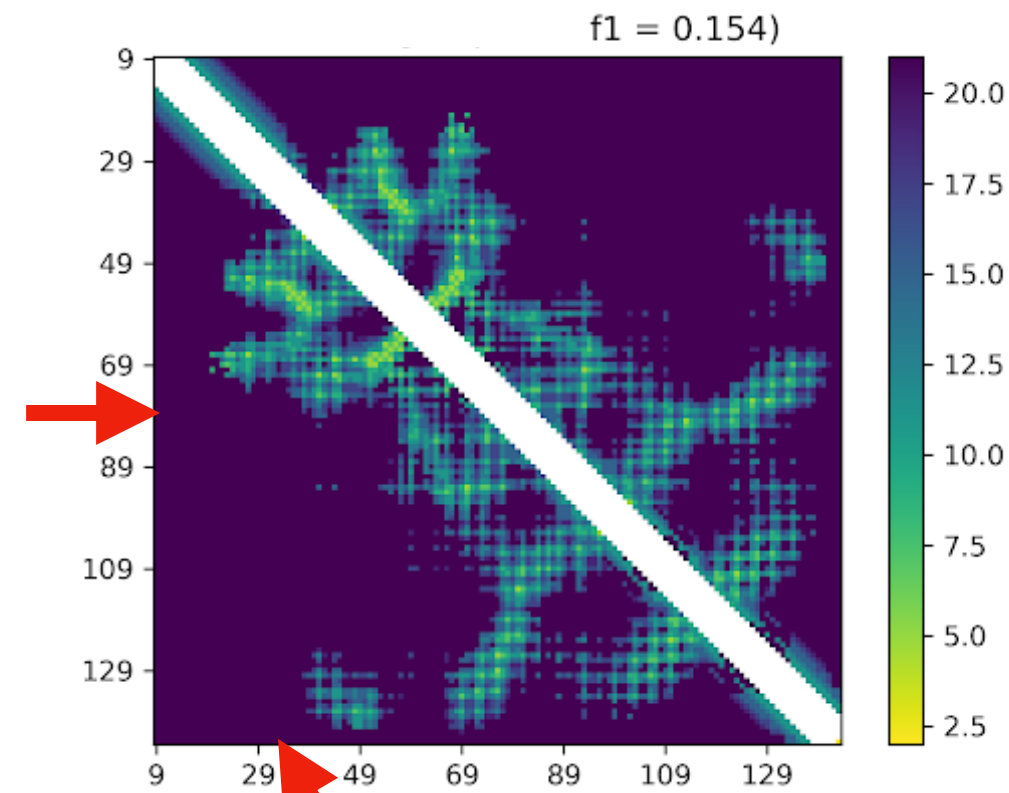
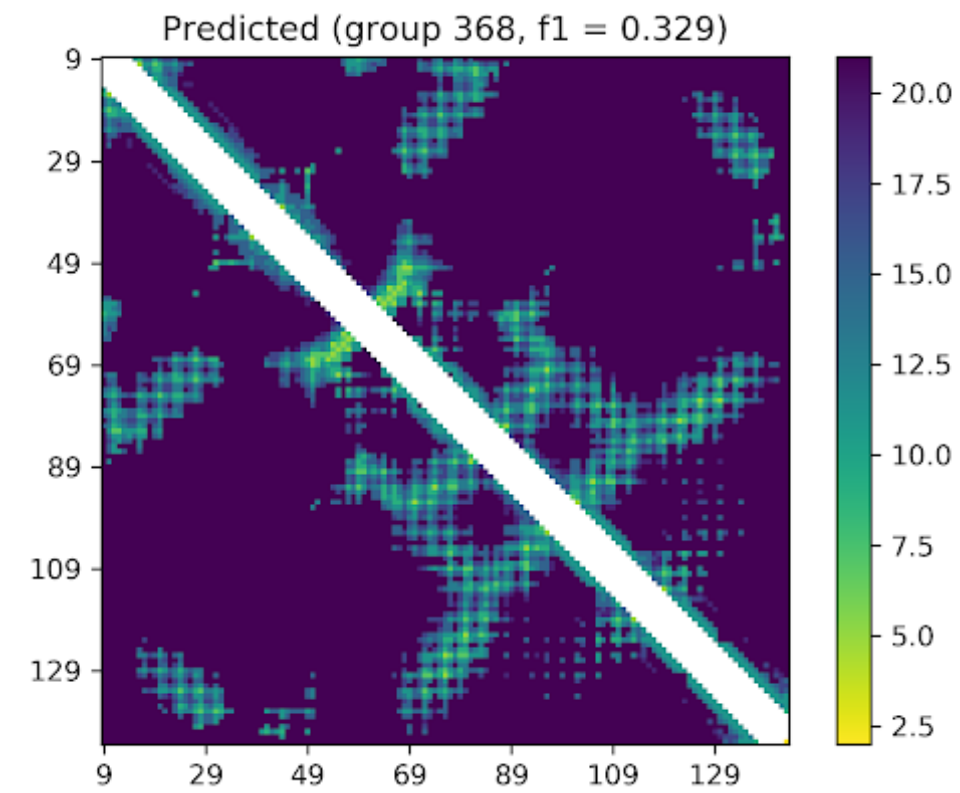
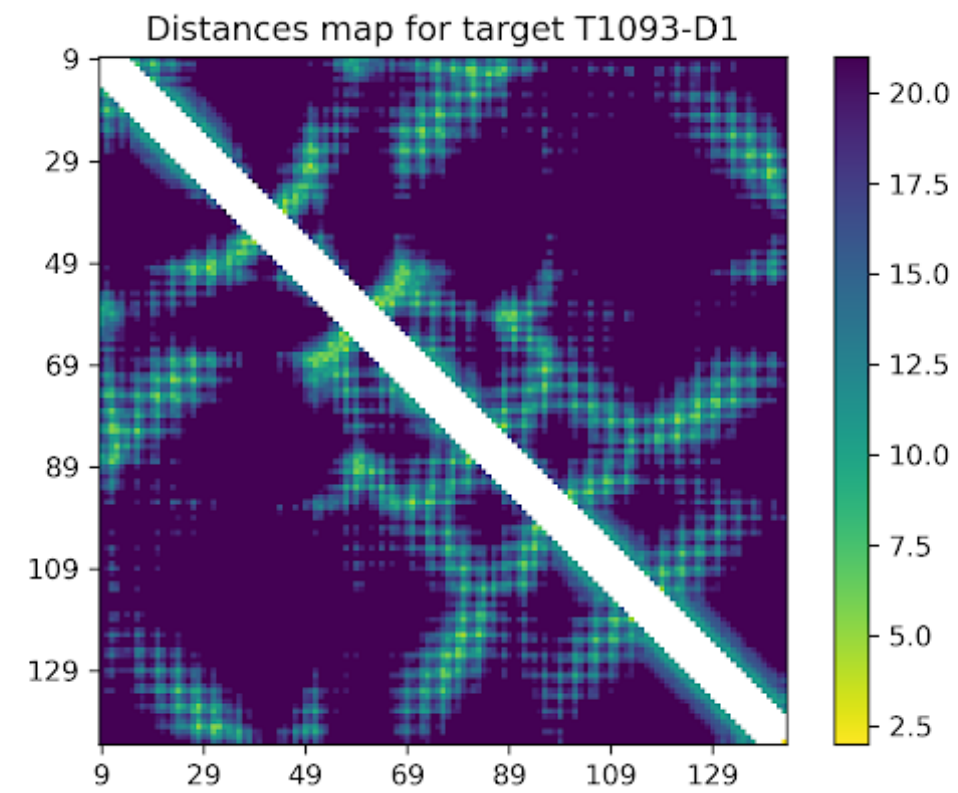
Groups



T1094-D2 - FM $\log(\text{Neff} / \text{len}) = 0.31$

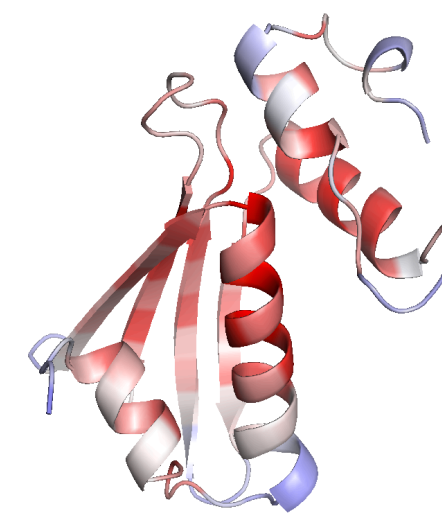
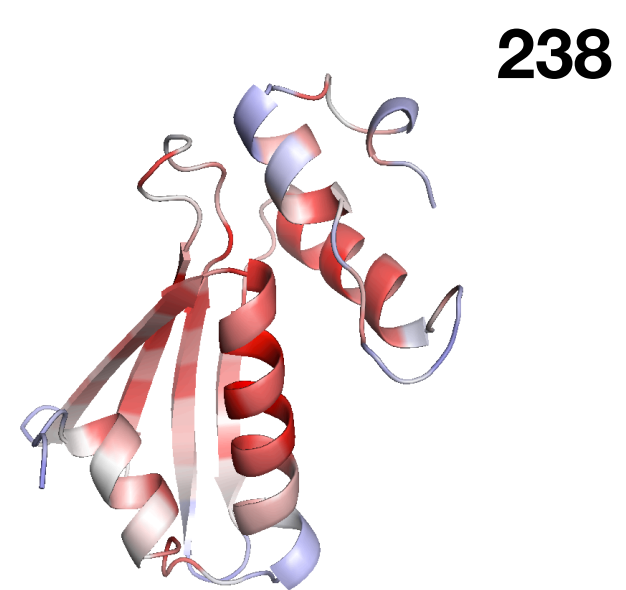
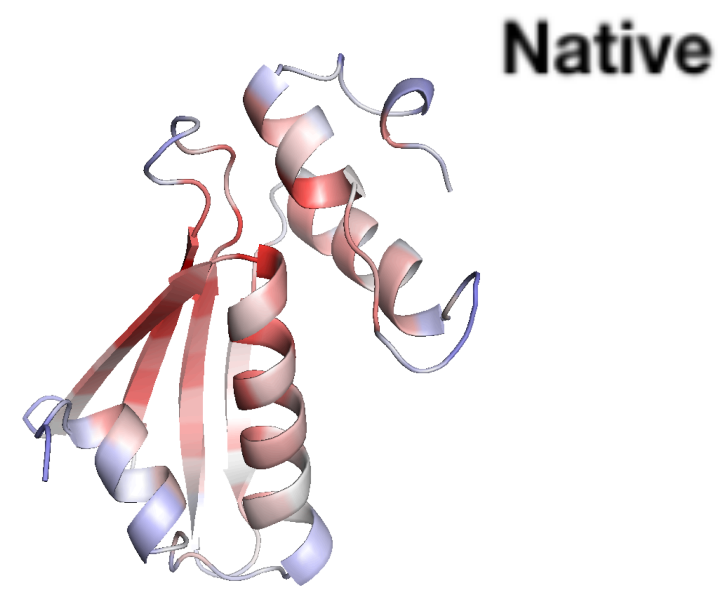
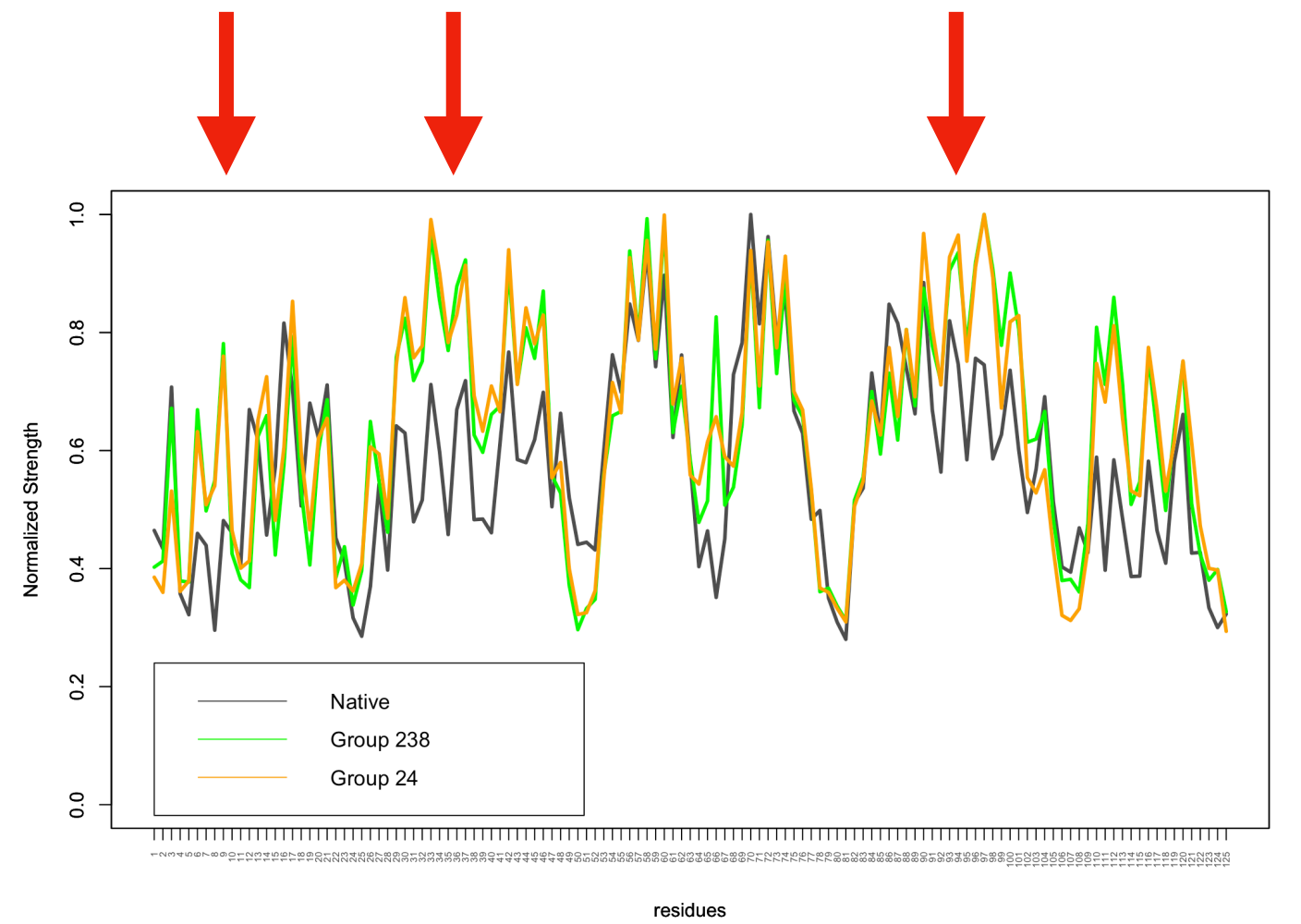
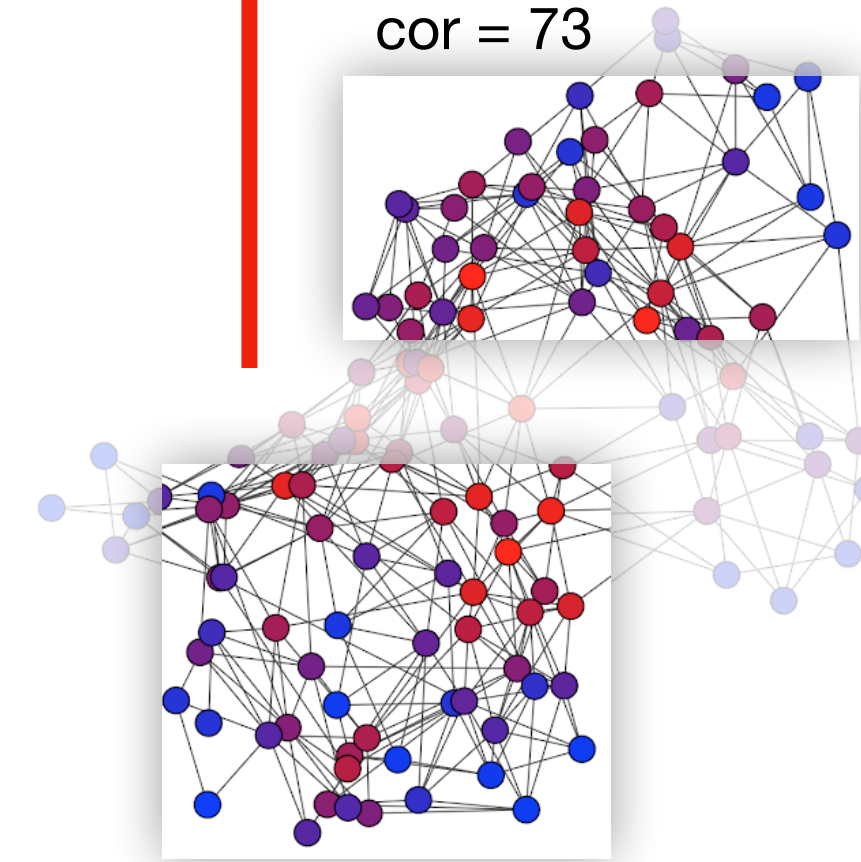
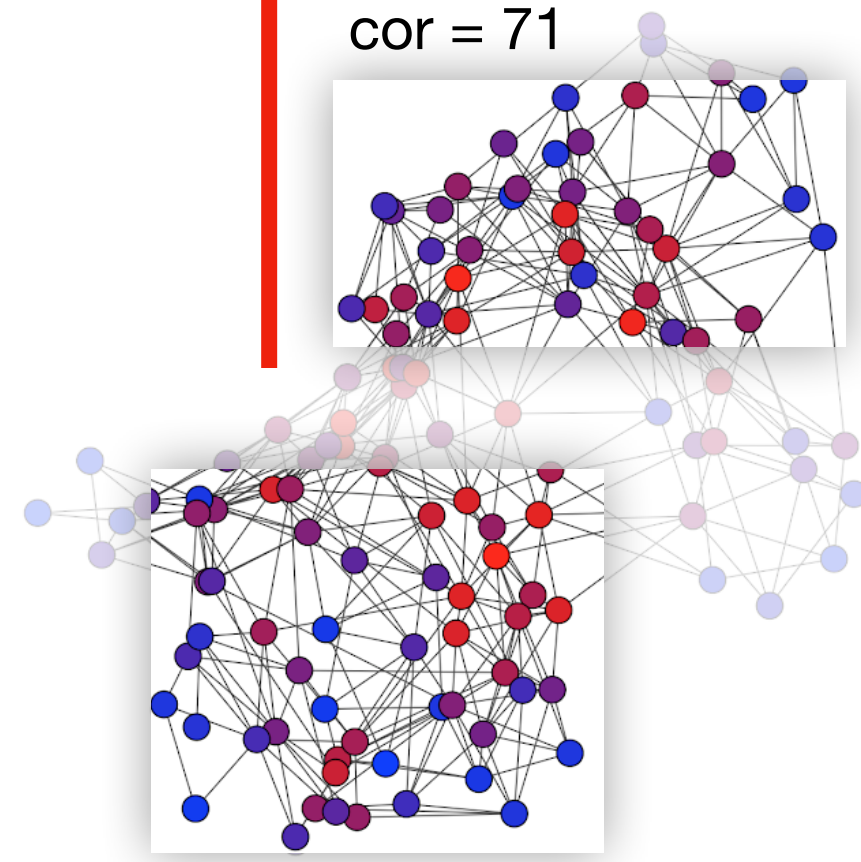
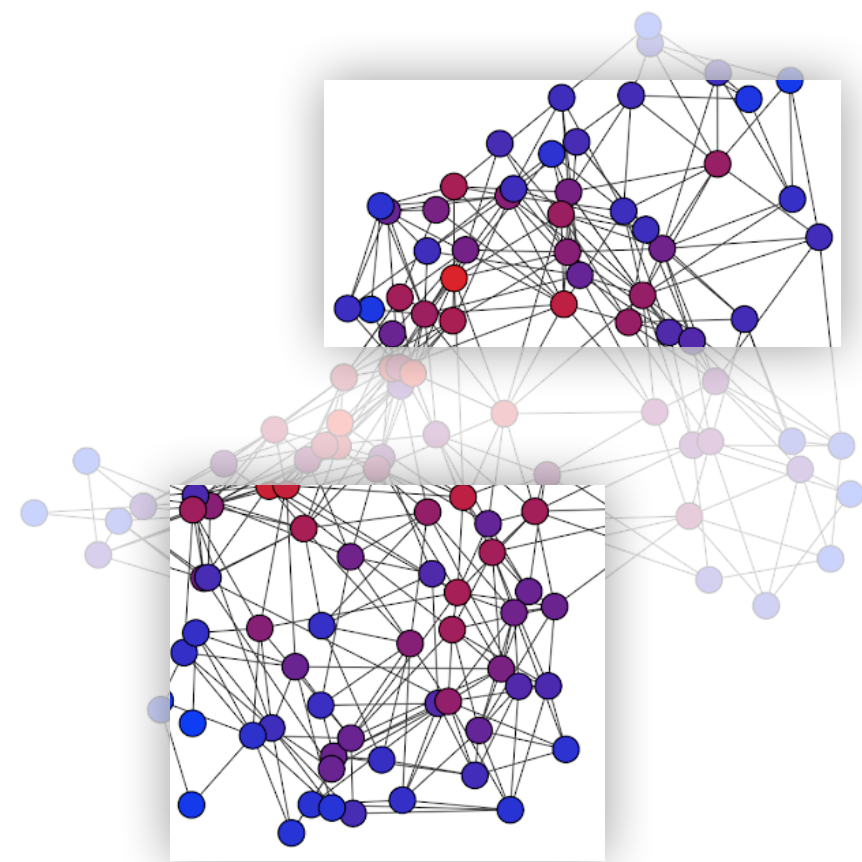
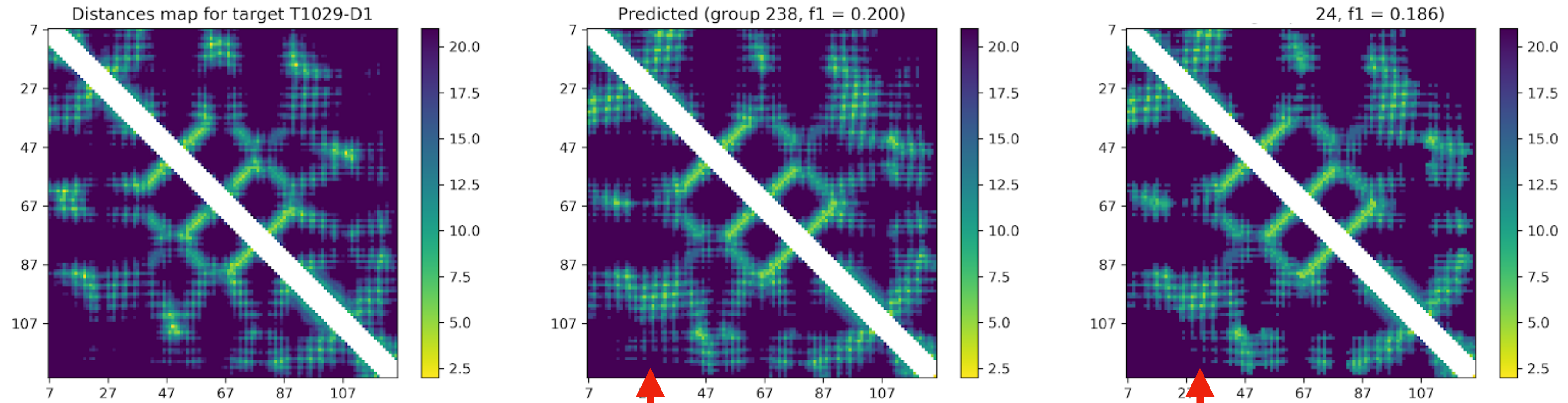


T1093-D1 - FM $\log(\text{Neff} / \text{len}) = 0.11$

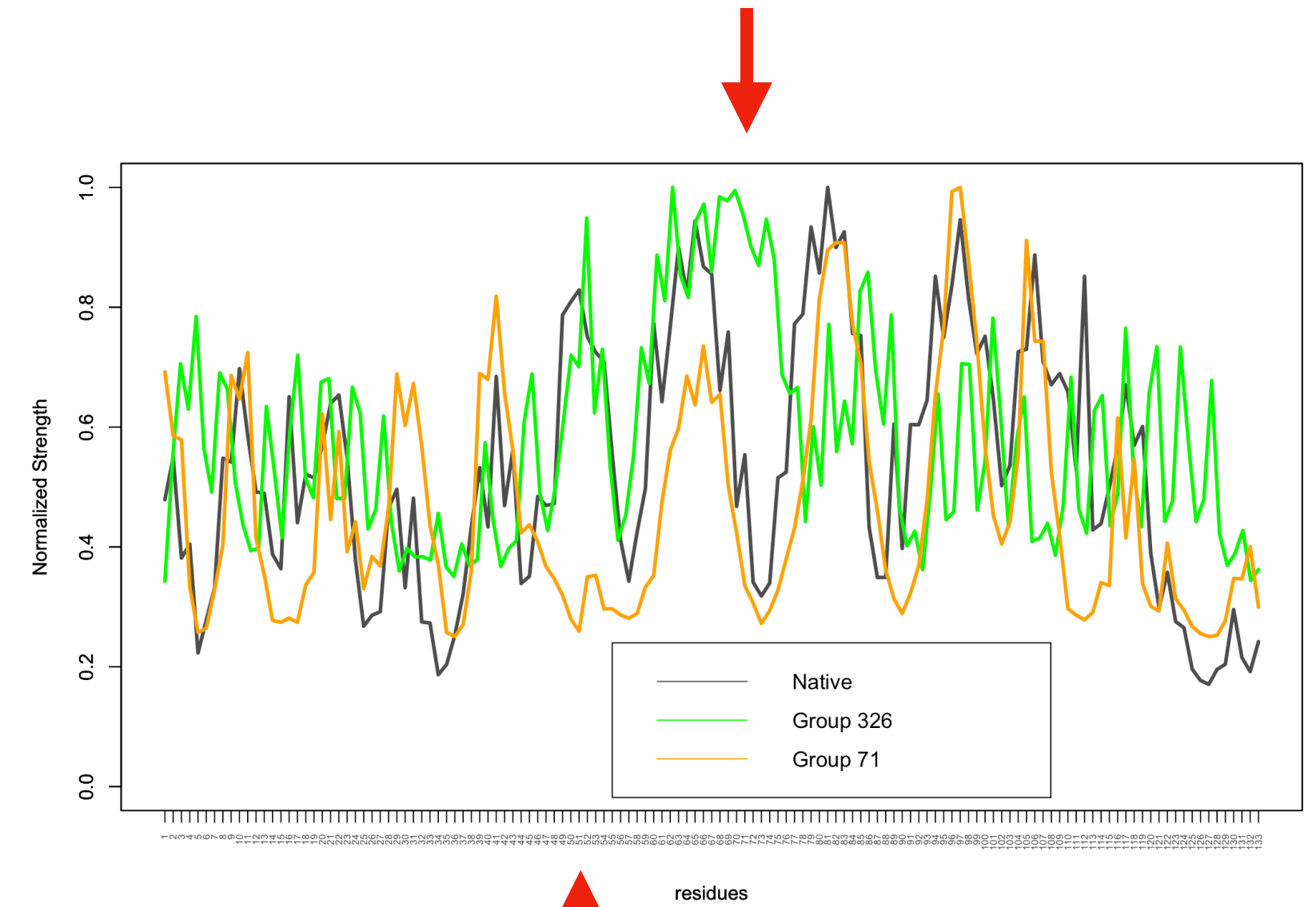
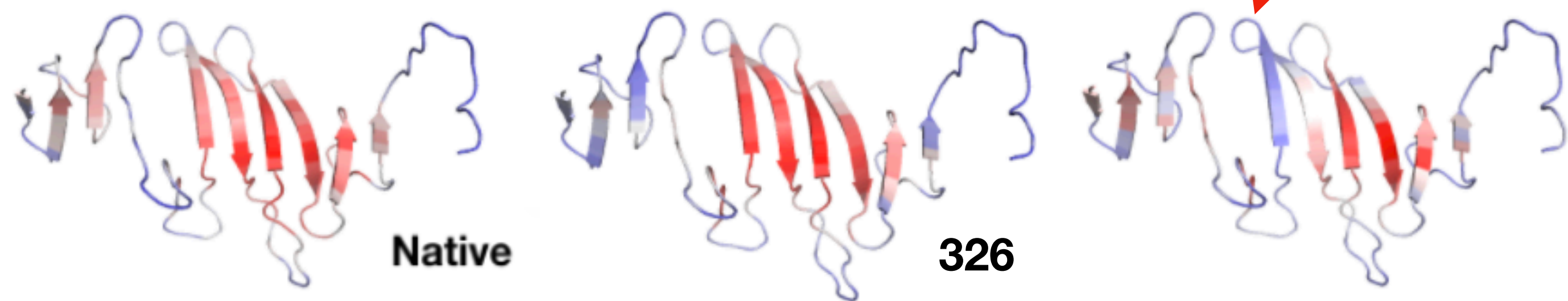
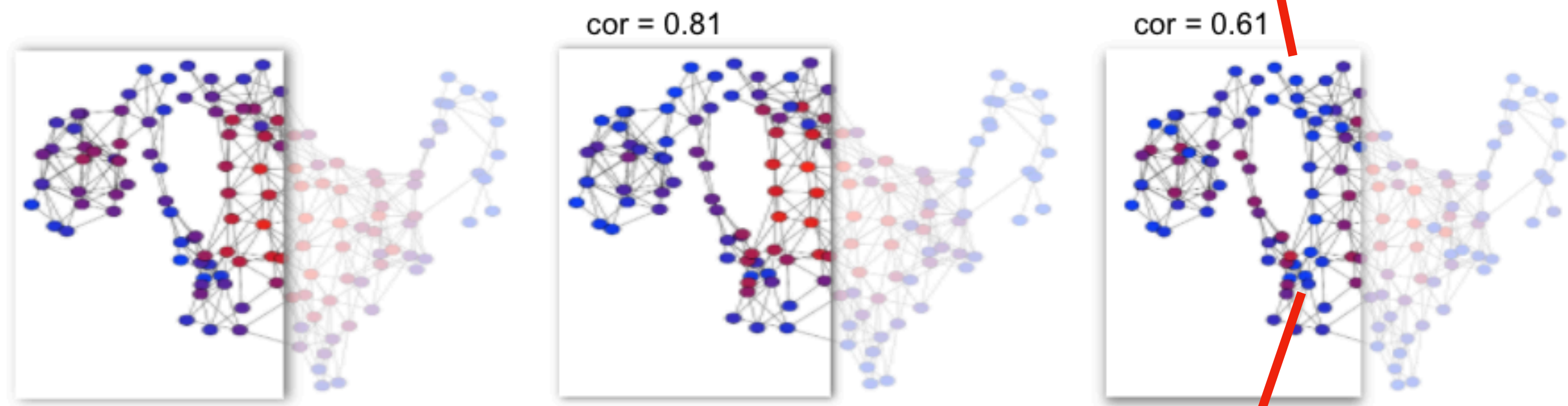
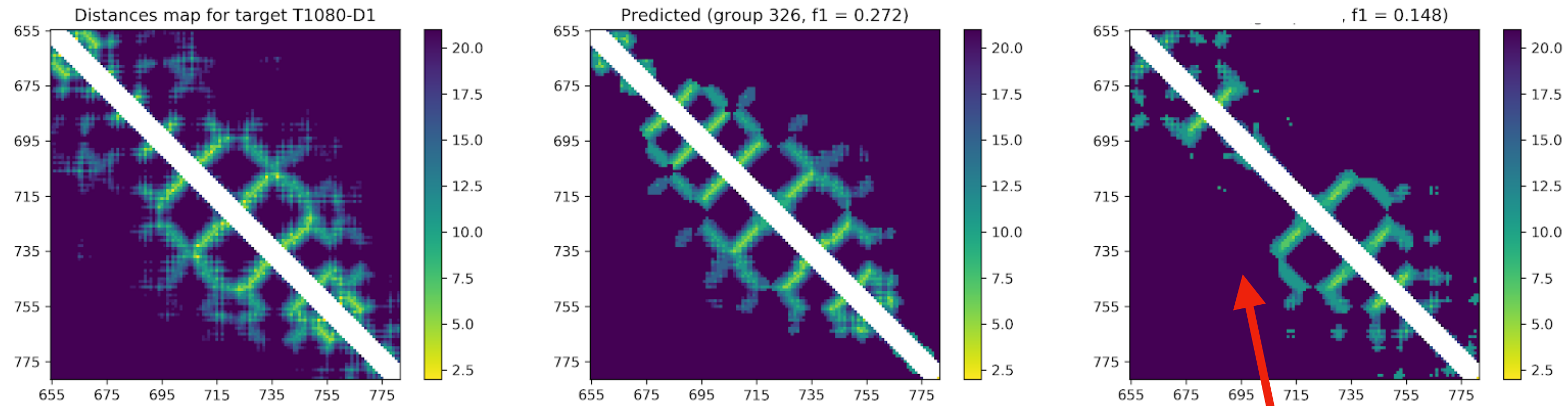


N-term helix-loop-helix
Loop

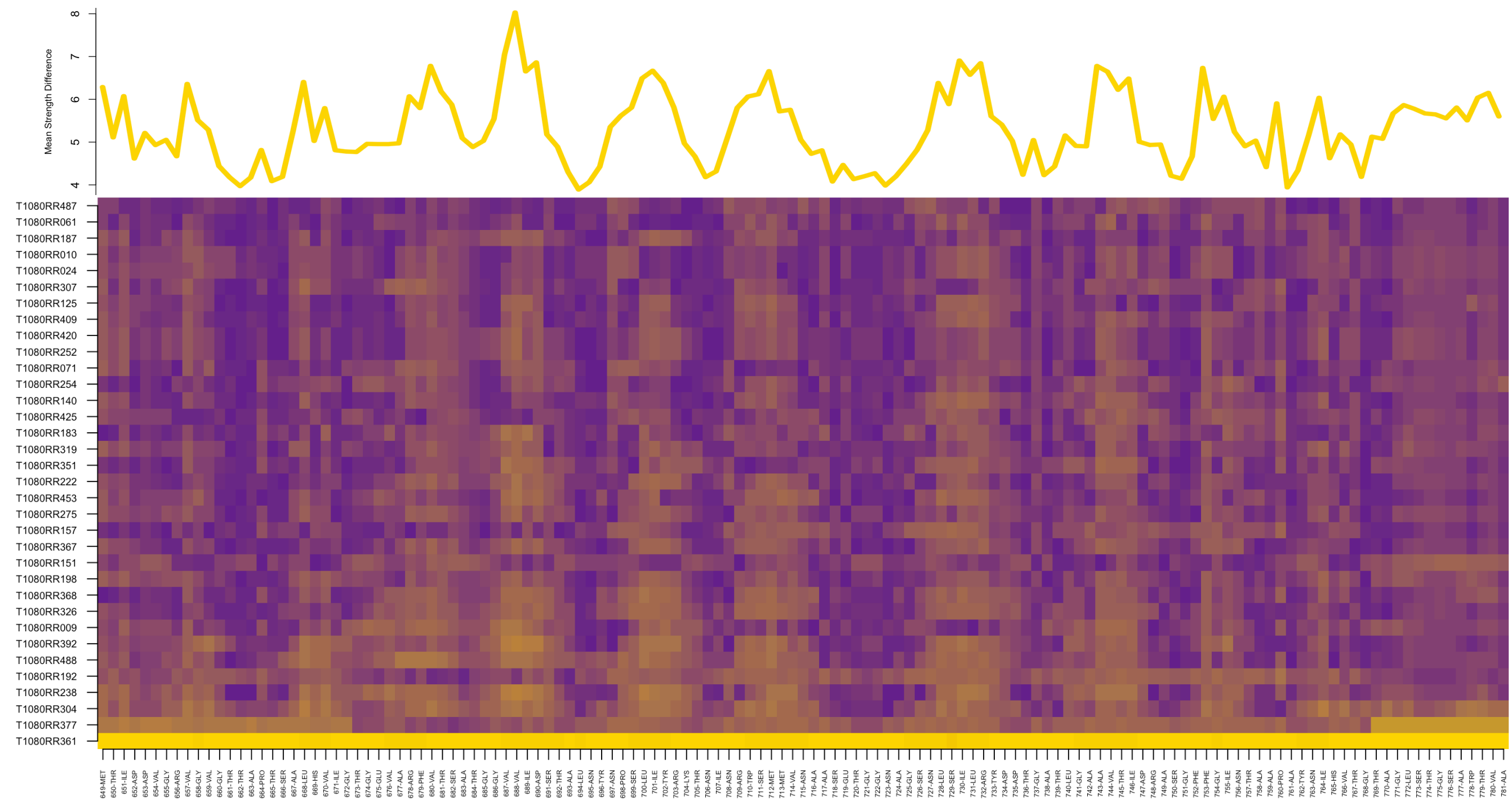
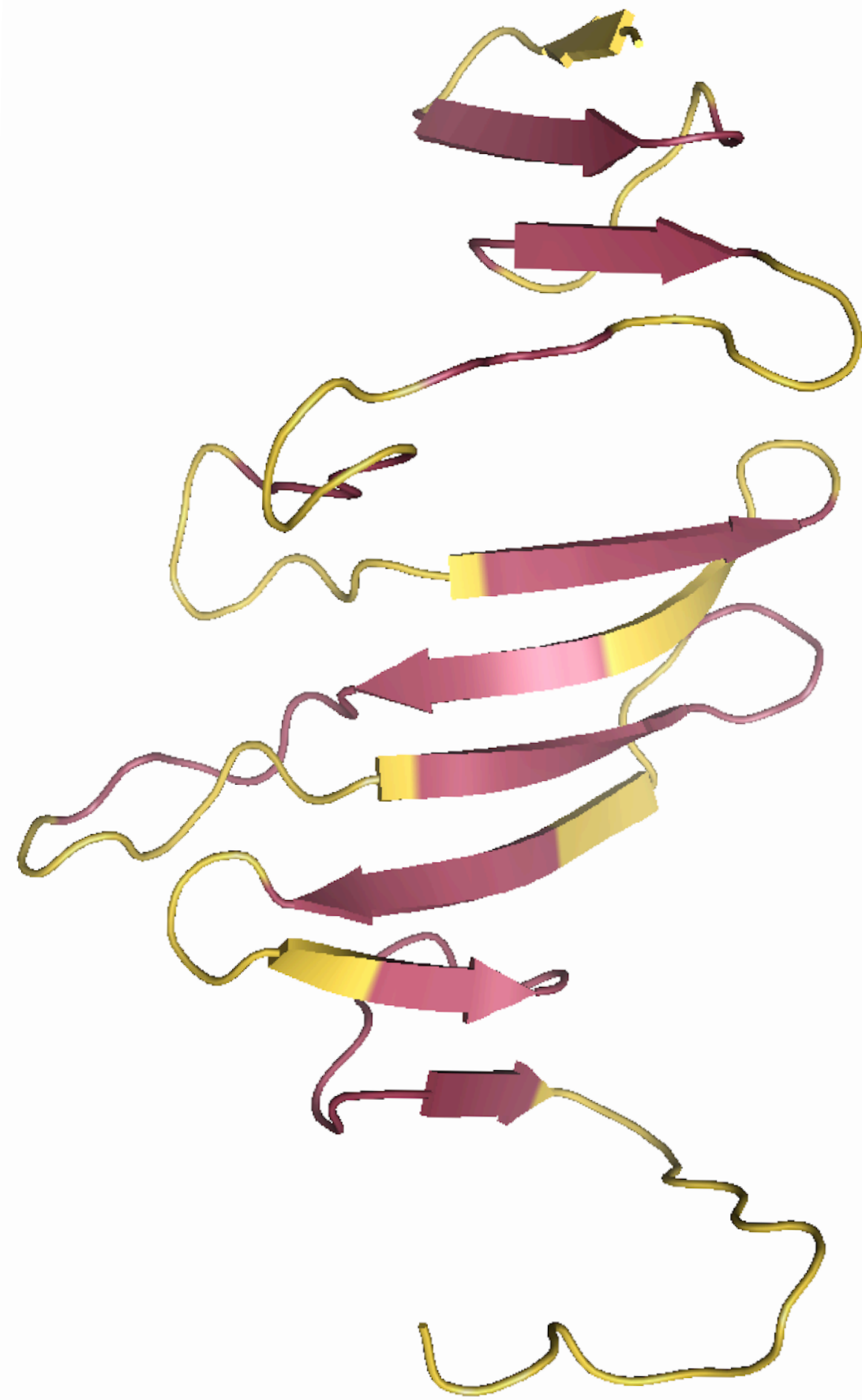
T1029-D1 - FM $\log(\text{Neff} / \text{len}) = 1.84$



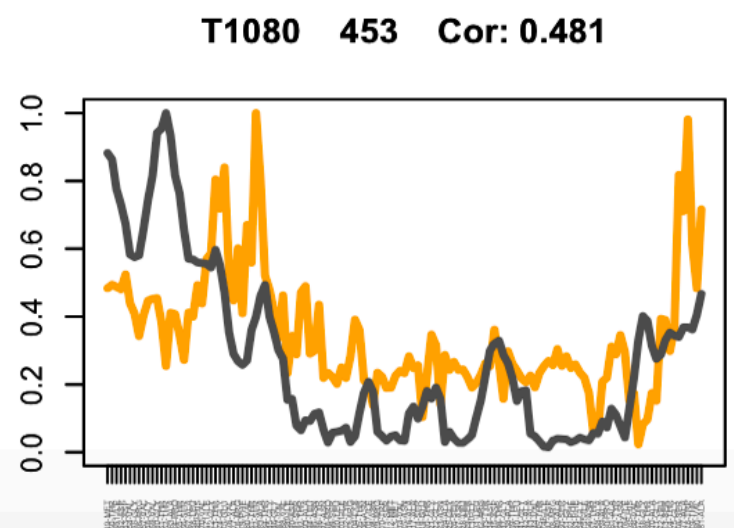
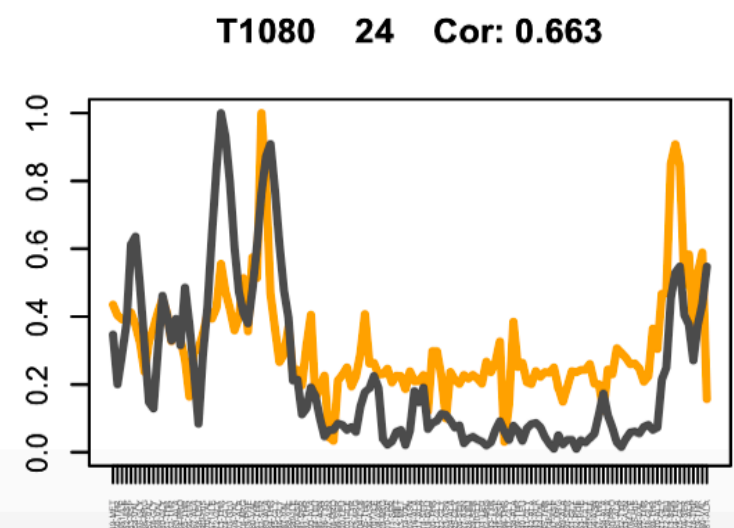
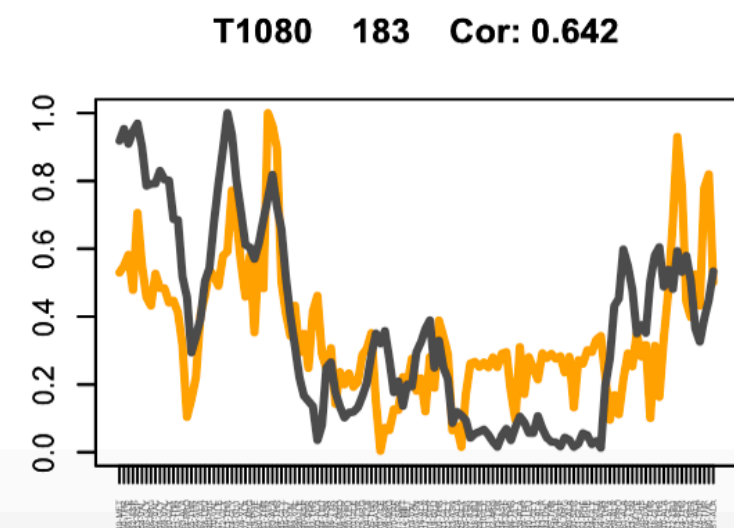
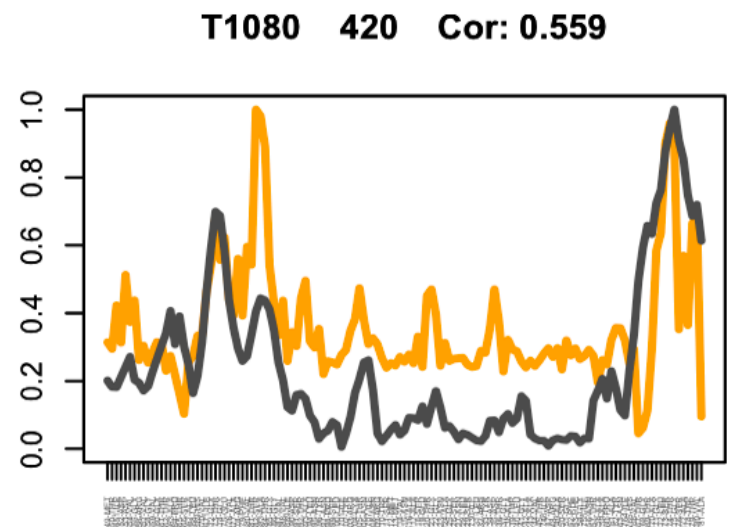
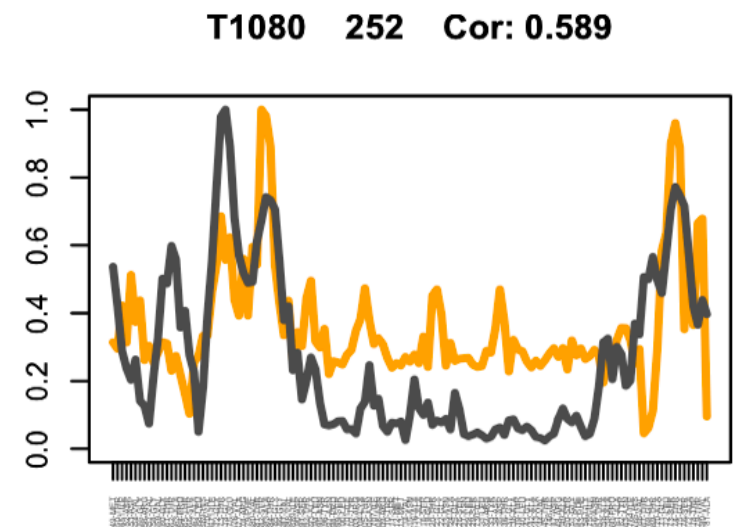
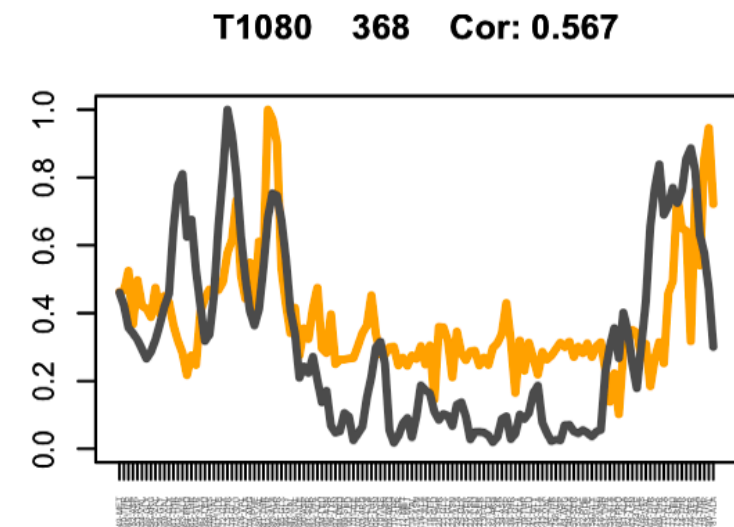
T1080-D1 - FM/TBM



T1080-D1 - FM/TBM



**Strength
Prediction vs Target**

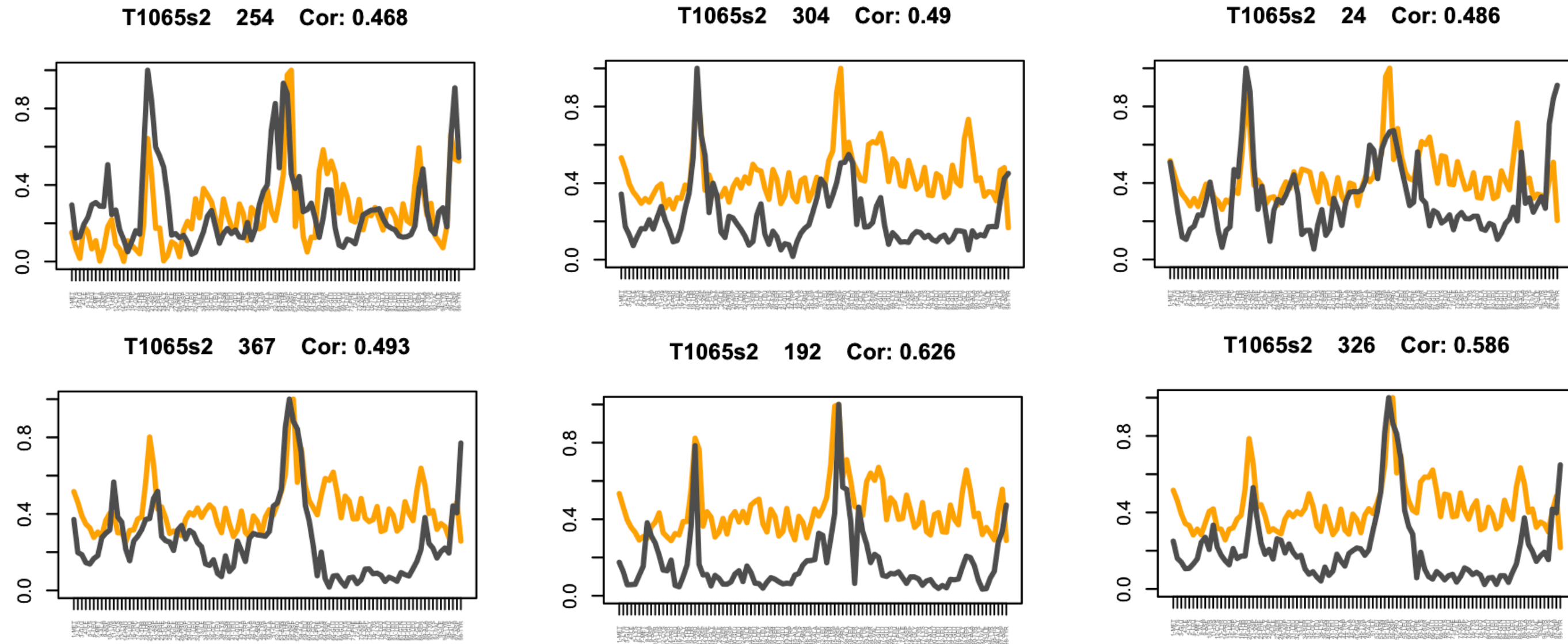


**Shortest path
prediction vs Target**

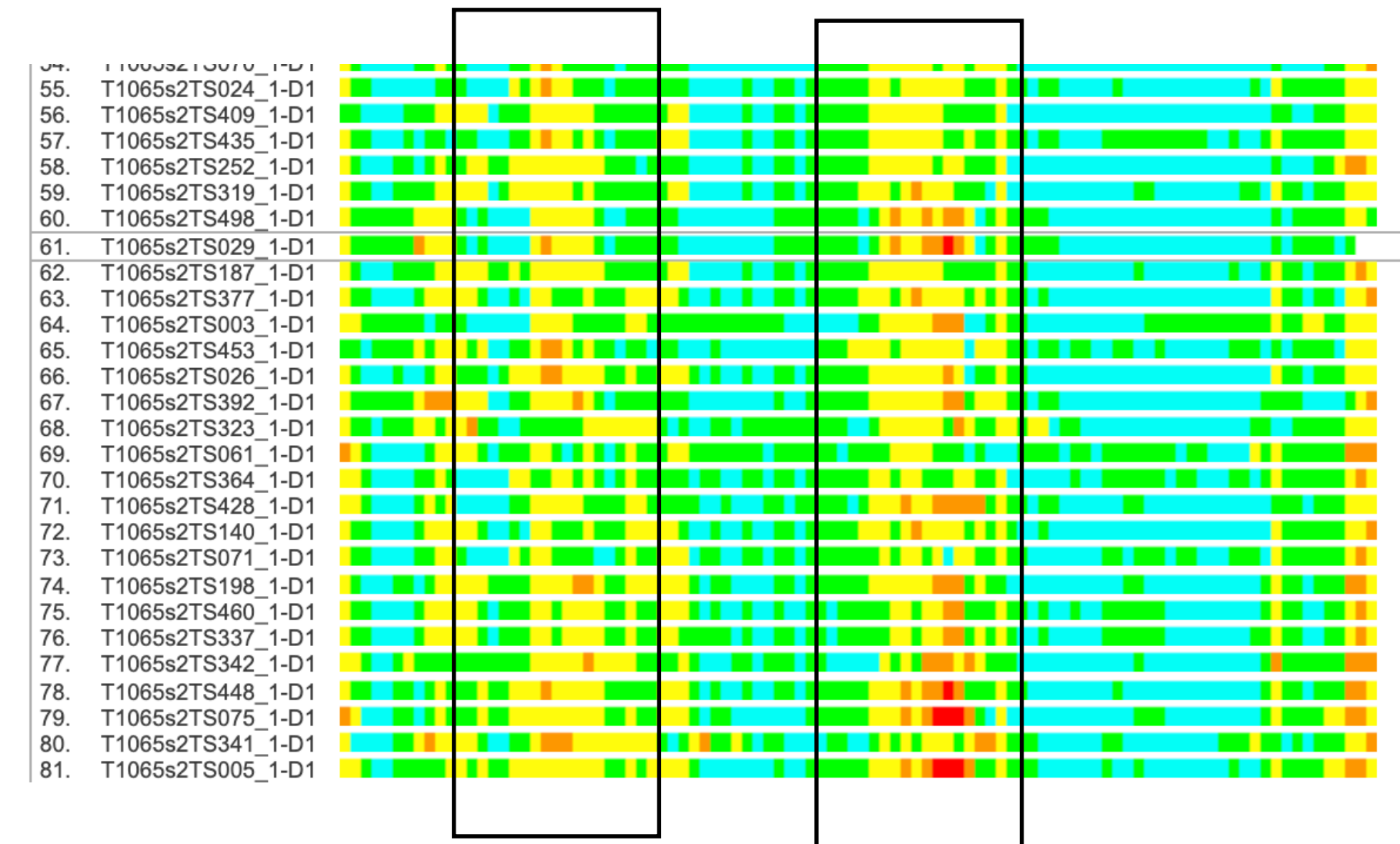
VS

RMSD

Errors in distance (Δ shortest path) vs errors in models (rmsd)

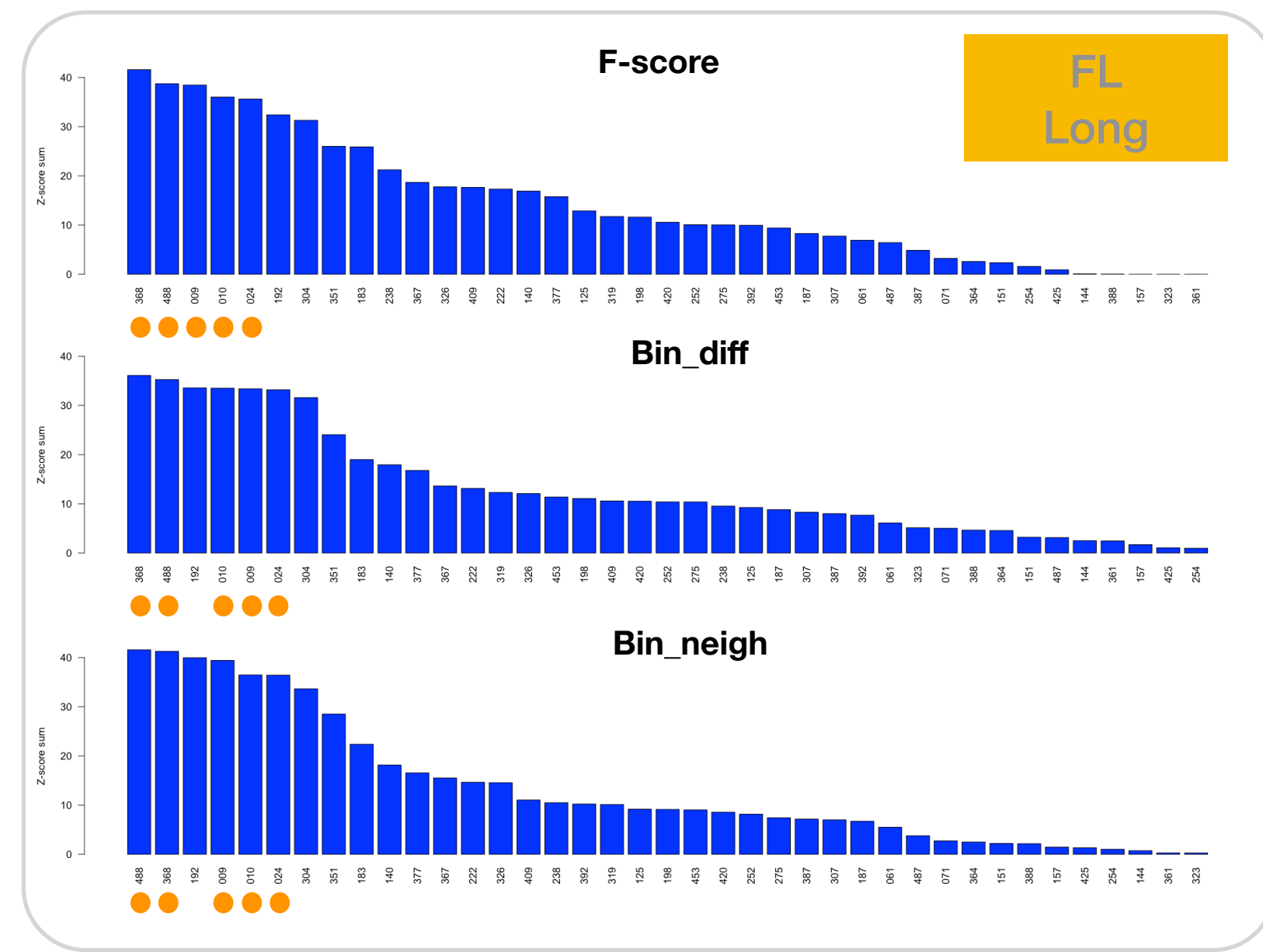
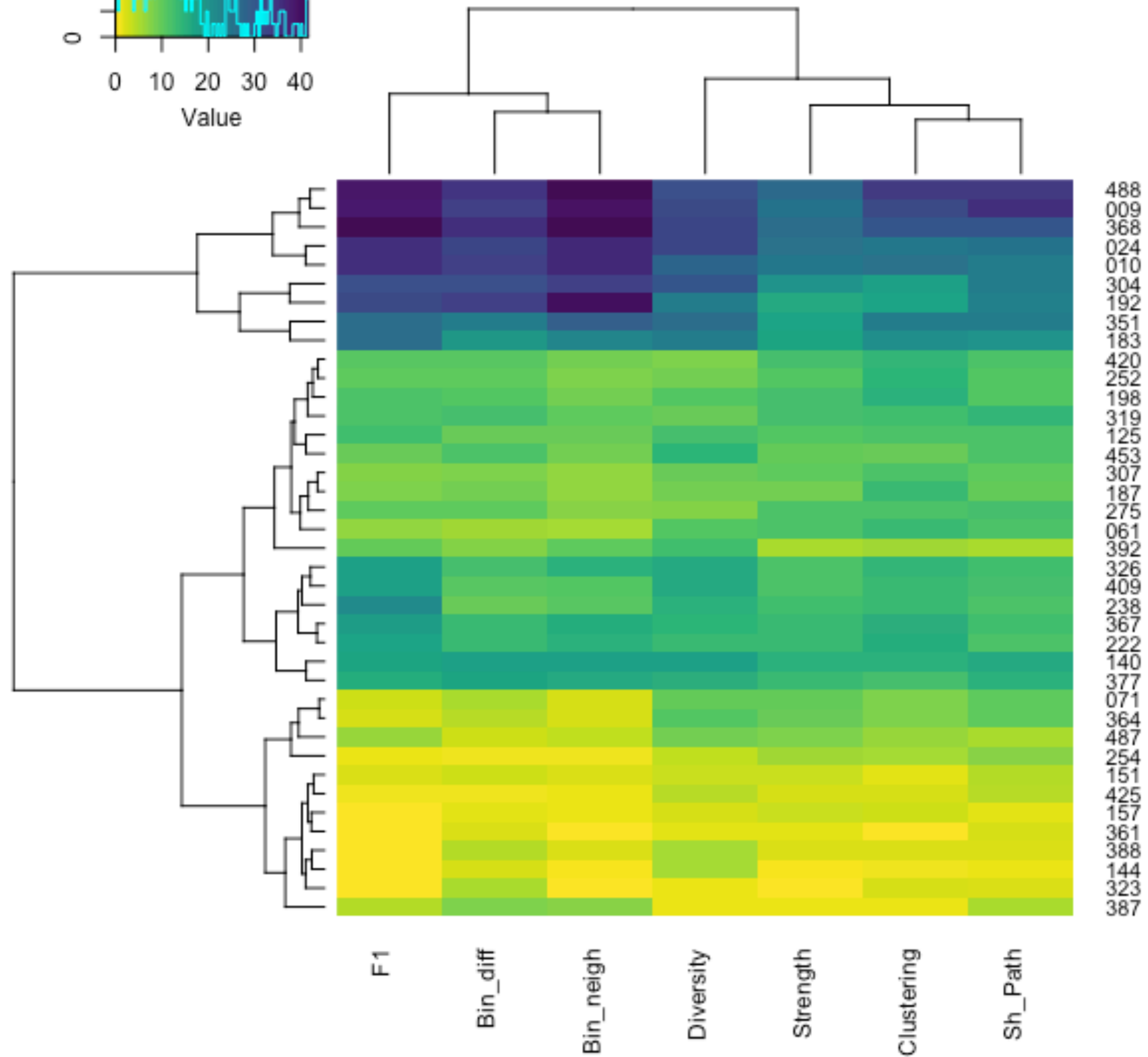
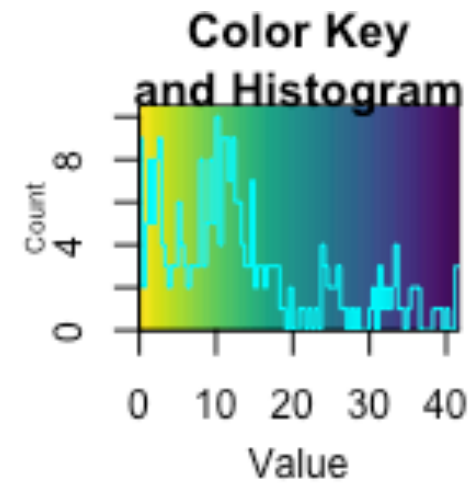
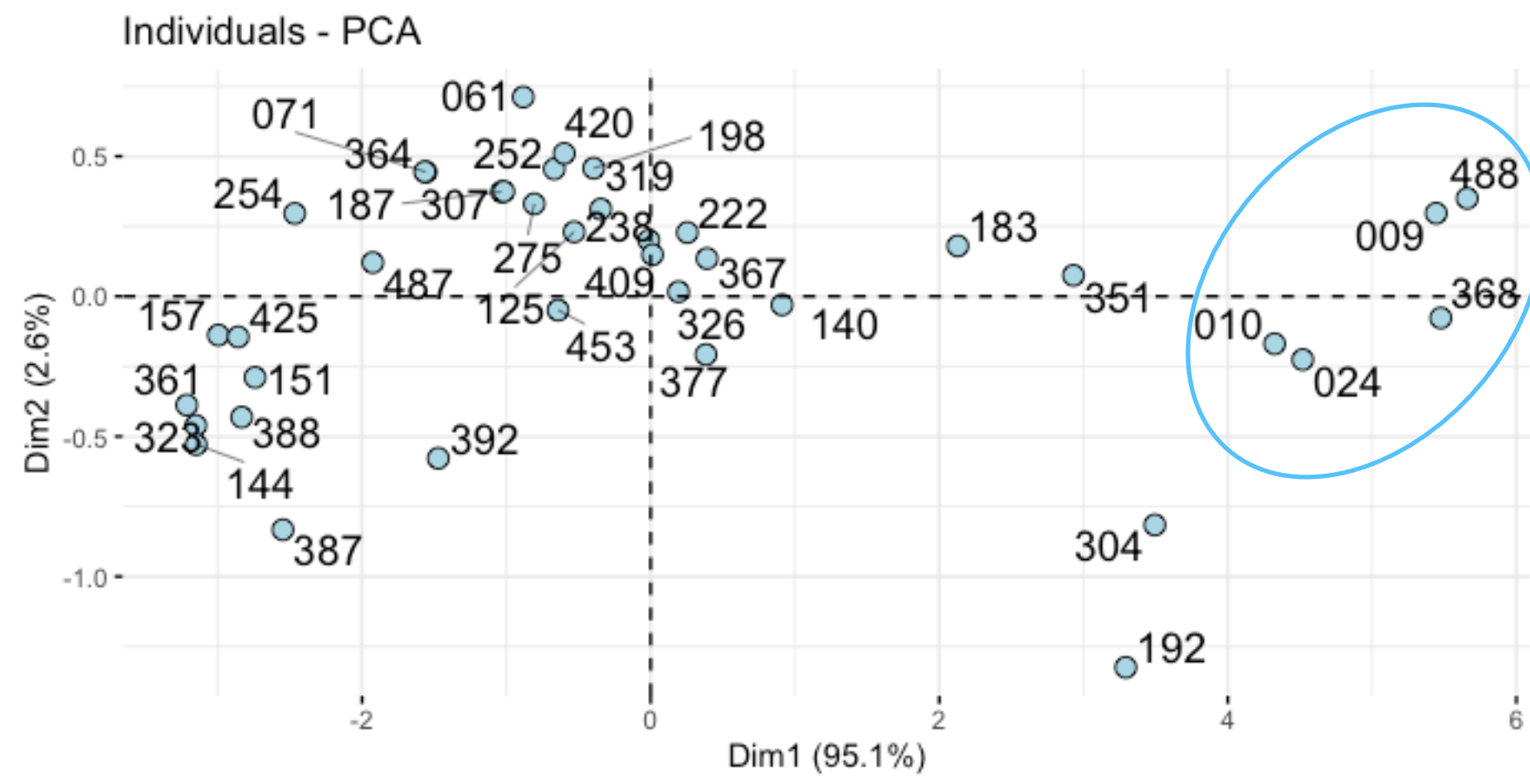


T1065s2-D1 - FM/TBM

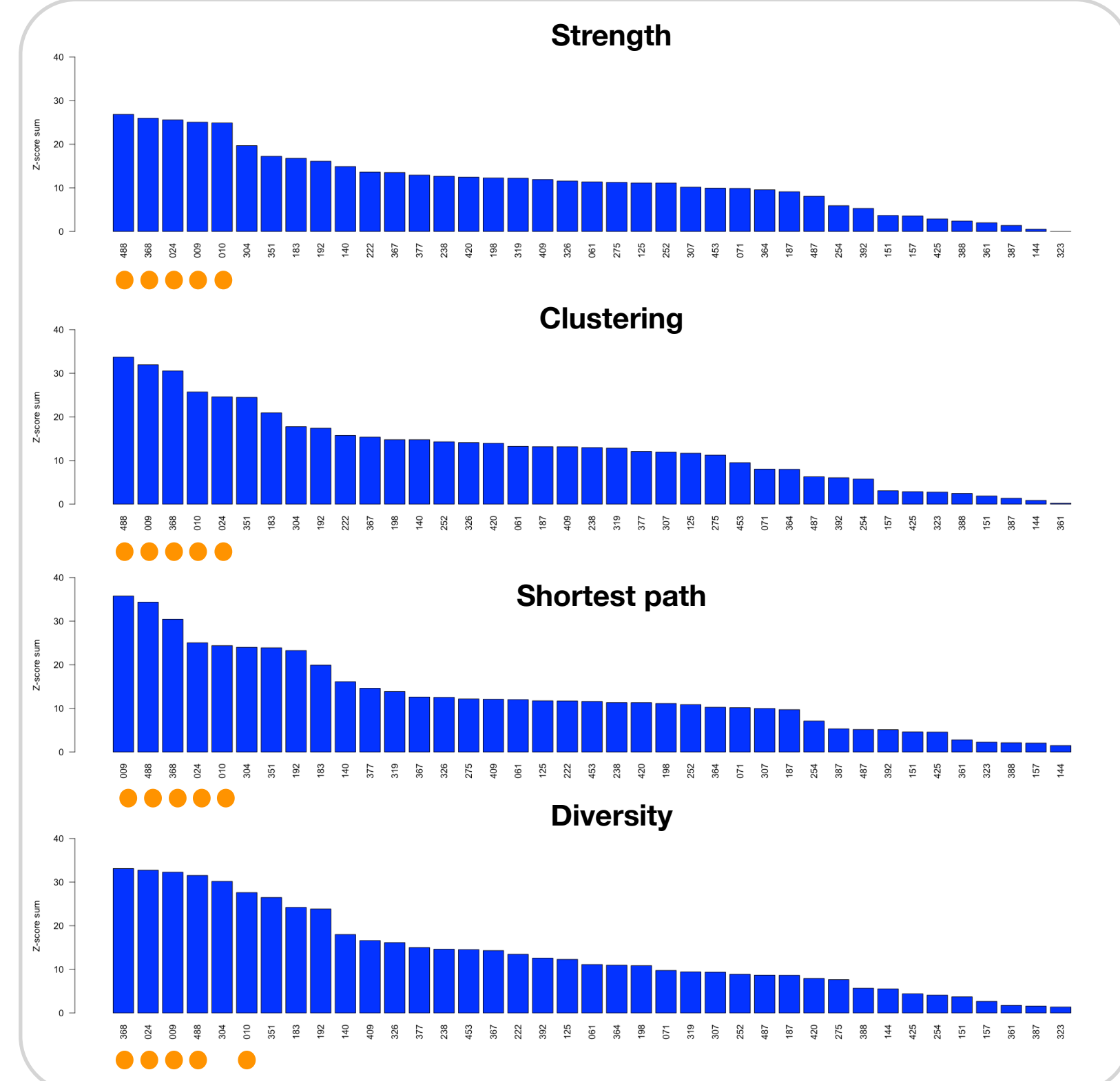


Performances

sum(z-scores>0)

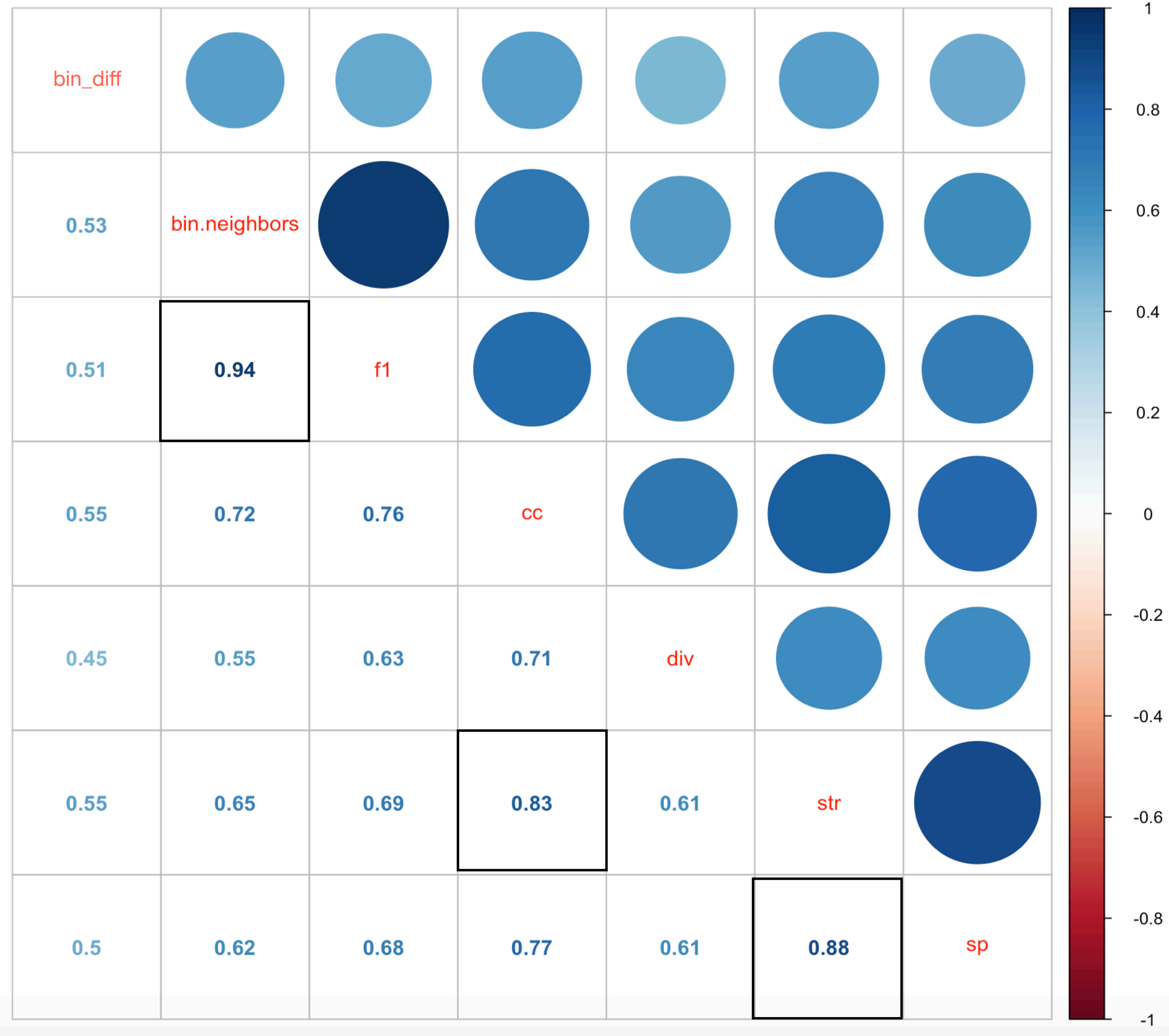


i-j pair



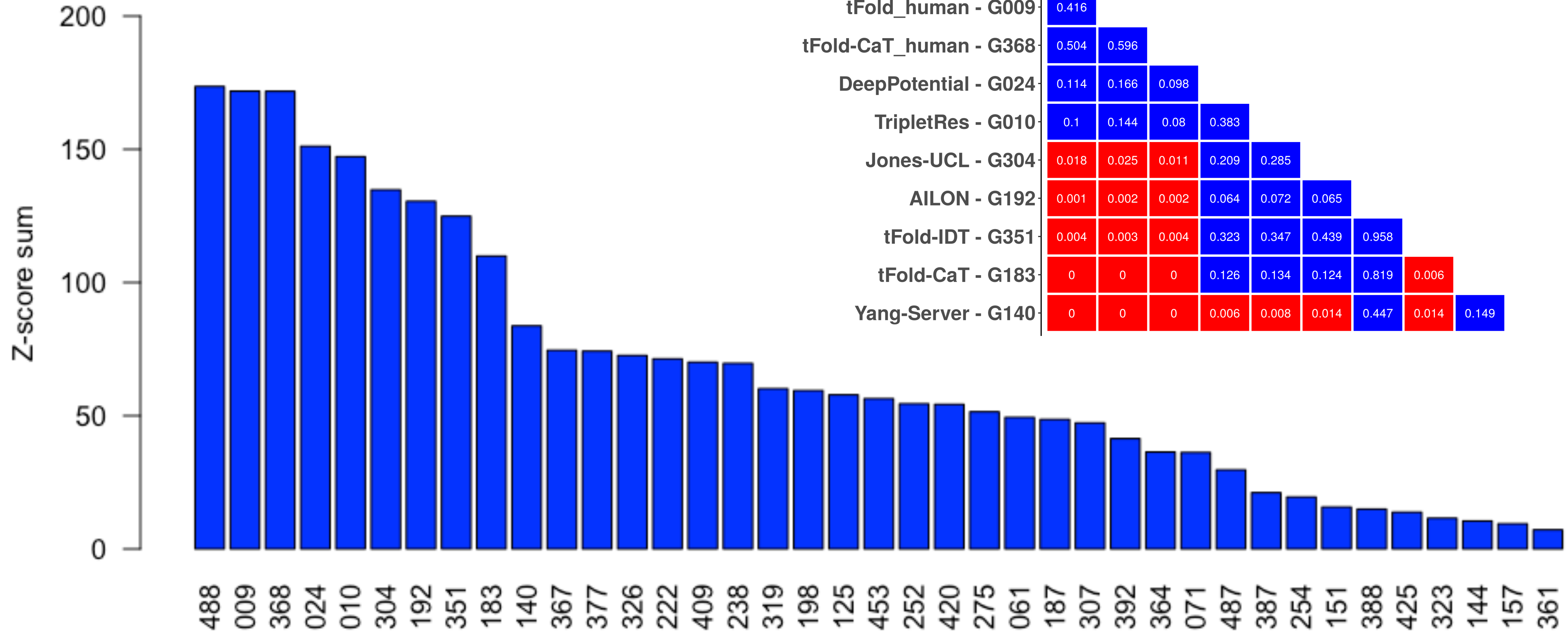
graph-based

Correlation between metrics



$$\text{Cumulative Z-score} = z_{F1} + z_{\text{Bin_diff}} + z_{cc} + z_{sp} + z_{div}$$

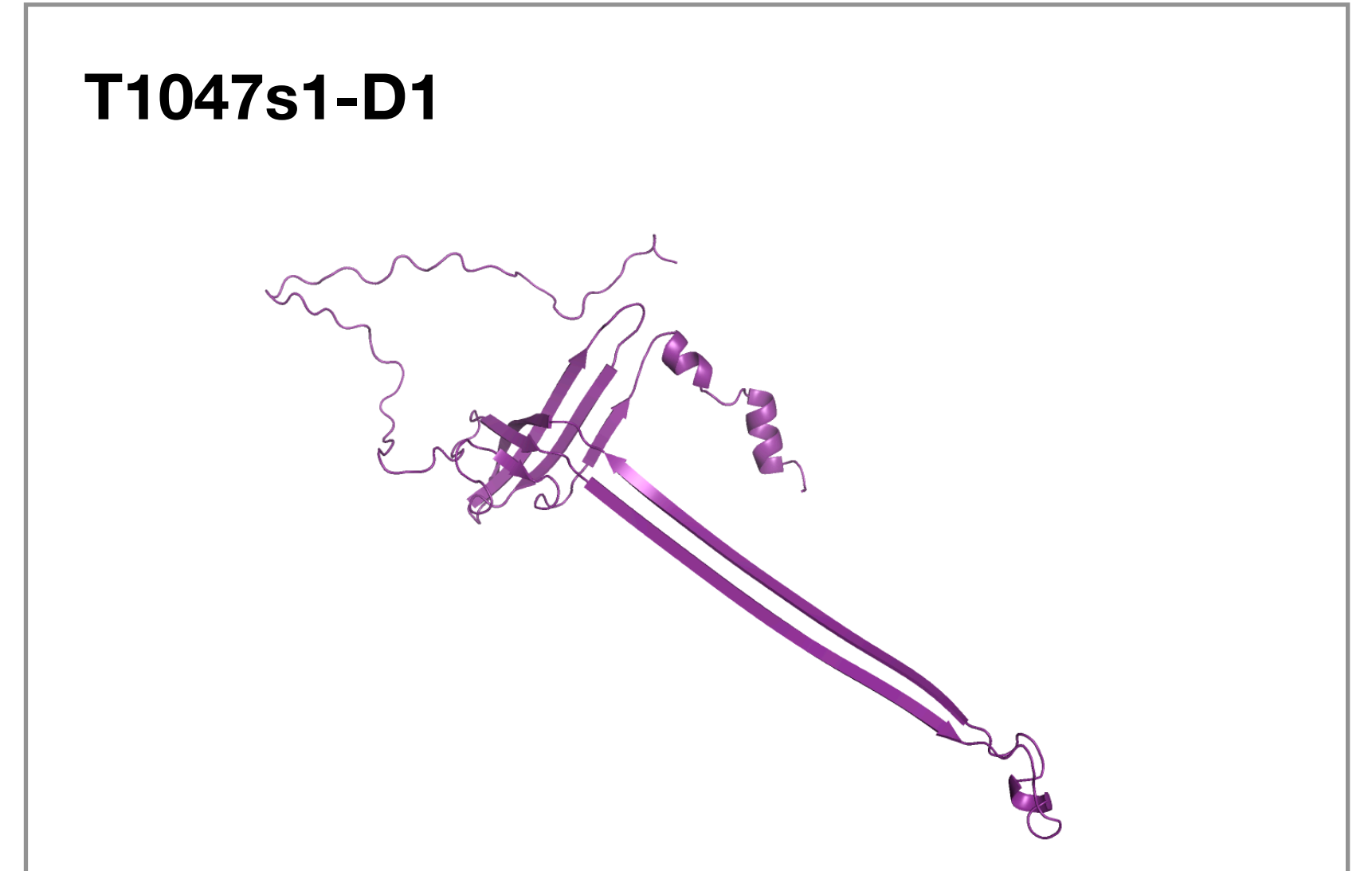
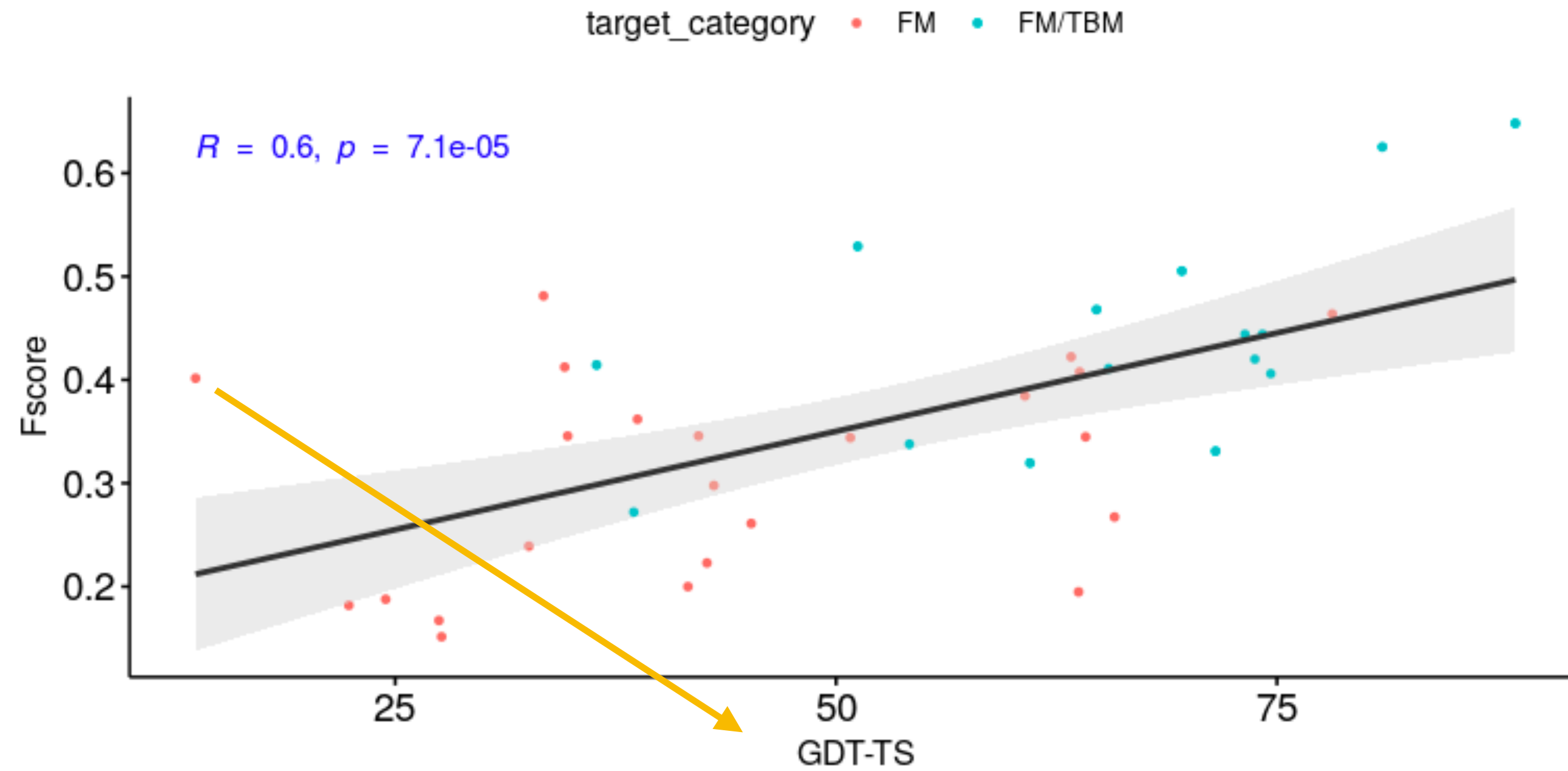
Ranking - Distances Long FL



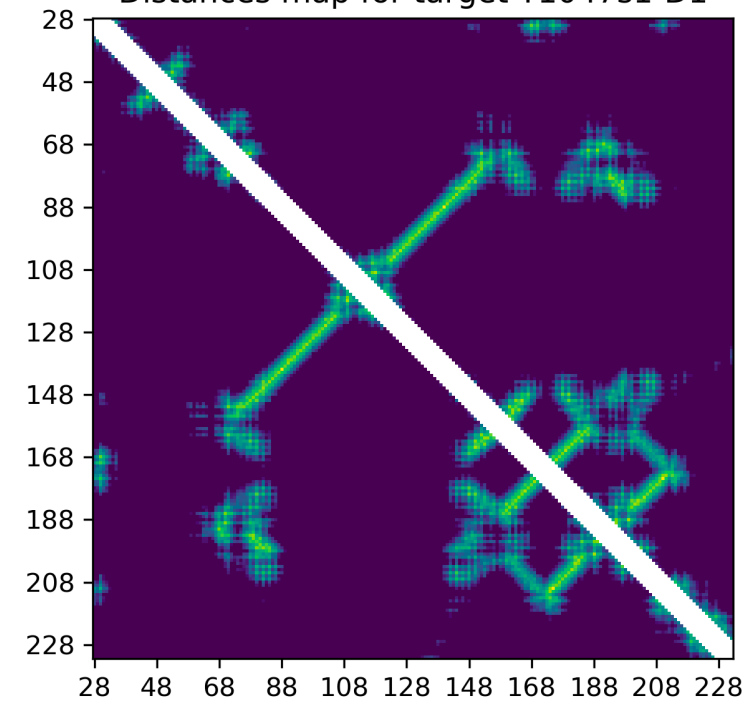
Head to head & paired t-test

	G488	G009	G368	G024	G010	G304	G192	G351	G183	G140
tFold-IDT_human - G488										
tFold_human - G009	0.416									
tFold-CaT_human - G368	0.504	0.596								
DeepPotential - G024	0.114	0.166	0.098							
TripletRes - G010	0.1	0.144	0.08	0.383						
Jones-UCL - G304	0.018	0.025	0.011	0.209	0.285					
AILON - G192	0.001	0.002	0.002	0.064	0.072	0.065				
tFold-IDT - G351	0.004	0.003	0.004	0.323	0.347	0.439	0.958			
tFold-CaT - G183	0	0	0	0.126	0.134	0.124	0.819	0.006		
Yang-Server - G140	0	0	0	0.006	0.008	0.014	0.447	0.014	0.149	

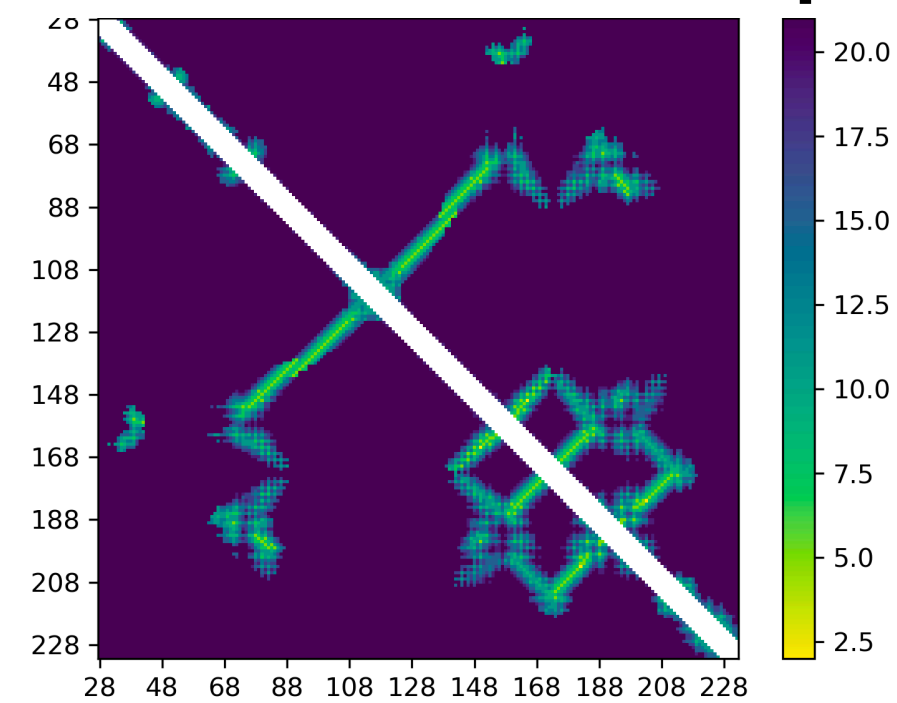
Performance in distance prediction vs performance in TS



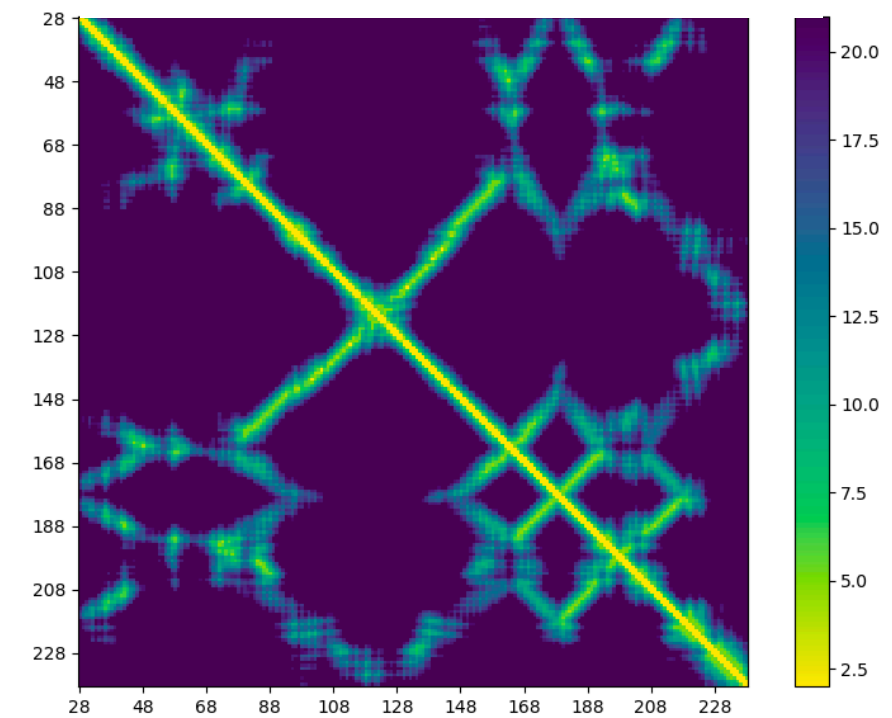
Distances map for target T1047s1-D1



Predicted distance map



Submitted model 1



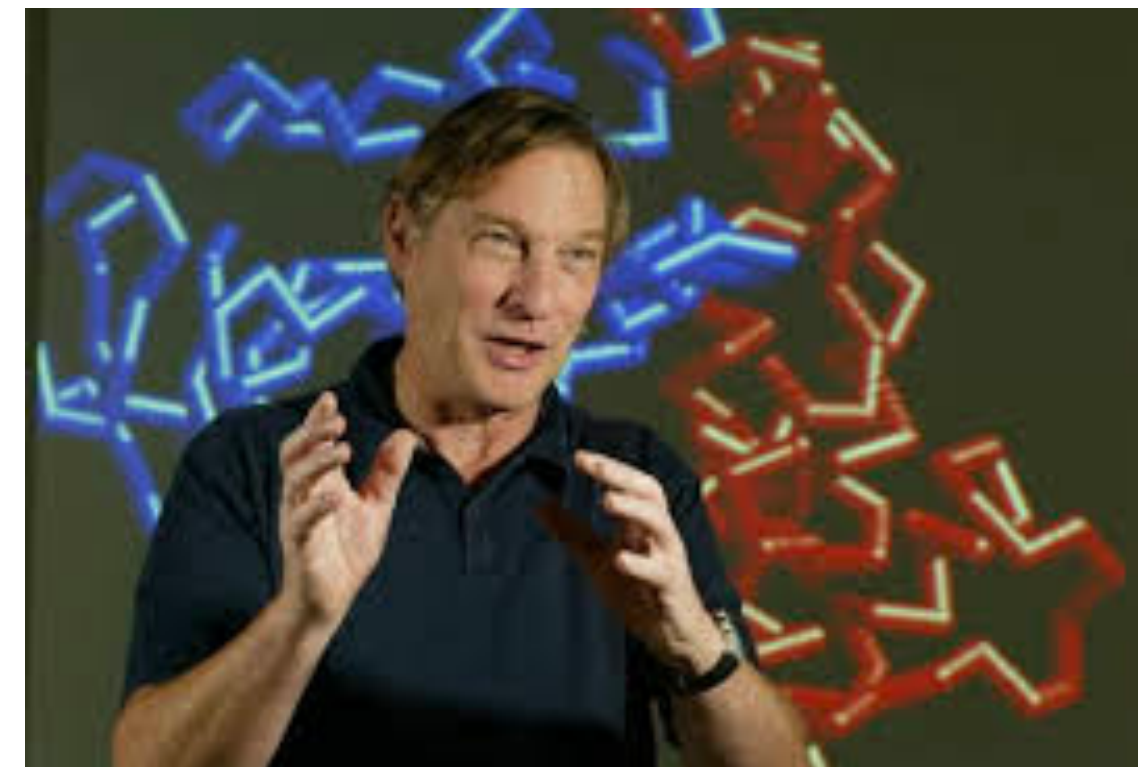
Conclusions

- Top 10 participants reached 70% average precision in contact prediction
- Apparent no progress in FM targets compared to CASP13 likely due to higher target difficulty and different content of secondary structure elements
- Both in contact and distance, very high quality predictions were made despite very low number of homologous sequences available
- Graph-based metrics might be helpful in interpreting/mapping distance predictions to local model quality
- Groups **488, 009, 368 (TFold_IDT_human, TFold_human, TFold_Cat_human)** and **024, 010 (DeepPotenital, TripletRes)** are consistently top ranking according to different metrics both in contact and distance predictions

Acknowledgements



Alfonso Valencia



John Mout



Camila Pontes



Victoria Ruiz-Serra



Edoardo Milanetti



Krzysztof Fidelis



Andriy kryshfovych