CAID-binding

June 3, 2019

Contents

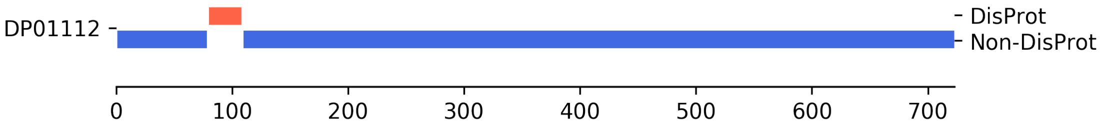
Chapter 1

CAID

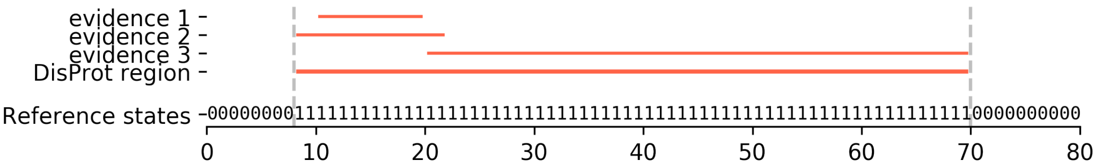
1.1 Dataset

Critical Assessment of Intrinsic Disorder (CAID) is a continuous experiment where prediction methods for intrinsic disorder (ID) are blind tested on unpublished DisProt data.

Current analysis is performed on the **new-disprot-binding** dataset with **simple** negative definition. This means that DisProt defines order for *new-disprot-binding* and *the contrary of DisProt defines order*.



DisProt entries can have annotation covering the same sequence space. In these cases different evidences are merged in a unique continuous region. DisProt merged regions define positive cases (labeled as 1) in reference states. Remaining states are considered negatives.



1.2 Evaluation metrics

Metrics evaluating prediction scores are calculated applying **Default** thresholds to prediction scores. Table is sorted by descending value of BAc column

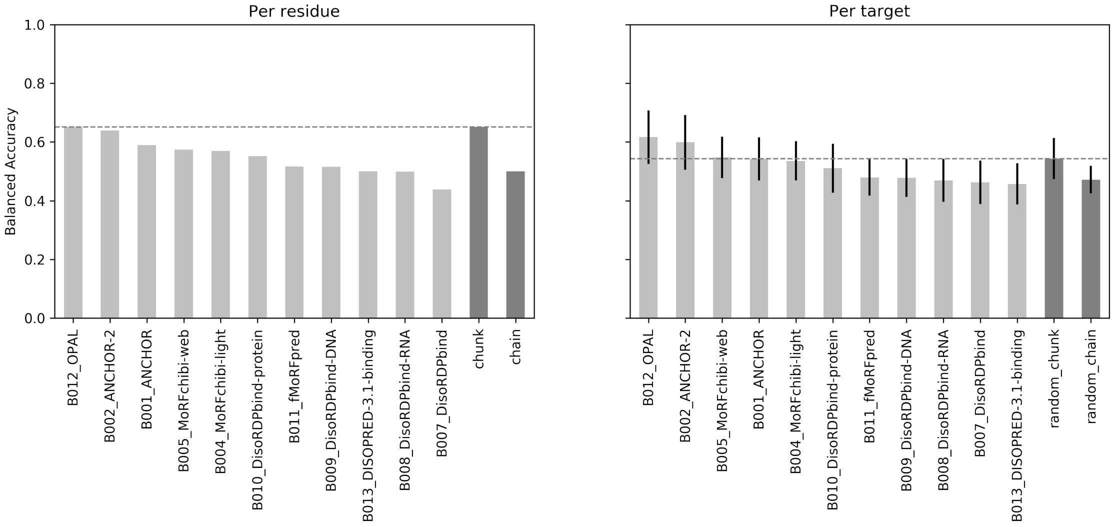
	FN	FP	TN	TP	BAc	F1s	MCC	Pre	Rec	Rec_n	AUC_PRC	AUC_ROC	npred	nref
chunk	12418.41	12418.41	92853.59	9044.59	0.652	0.421	0.303	0.421	0.421	0.882	NaN	NaN	100.0	100.0
B012_OPAL	4592.00	50479.00	54323.00	16604.00	0.651	0.376	0.226	0.248	0.783	0.518	0.348	0.717	232.0	233.0
B002_ANCHOR-2	9325.00	30195.00	75077.00	12138.00	0.639	0.381	0.222	0.287	0.566	0.713	0.278	0.701	233.0	233.0
B001_ANCHOR	12151.00	26906.00	78366.00	9312.00	0.589	0.323	0.148	0.257	0.434	0.744	0.250	0.653	233.0	233.0
B005_MoRFchibi-web	17013.00	5071.00	99731.00	4183.00	0.574	0.275	0.214	0.452	0.197	0.952	0.335	0.703	232.0	233.0
B004_MoRFchibi-light	17493.00	3537.00	101265.00	3703.00	0.570	0.260	0.227	0.511	0.175	0.966	0.359	0.718	232.0	233.0
B010_DisoRDPbind-protein	17146.00	10292.00	94980.00	4317.00	0.552	0.239	0.121	0.296	0.201	0.902	0.265	0.689	233.0	233.0
B011_fMoRFpred	20435.00	1734.00	103538.00	1028.00	0.516	0.085	0.081	0.372	0.048	0.984	0.208	0.546	233.0	233.0
B009_DisoRDPbind-DNA	19854.00	4767.00	100505.00	1609.00	0.515	0.116	0.051	0.252	0.075	0.955	0.197	0.521	233.0	233.0
B013_DISOPRED-3.1-binding	21463.00	0.00	105272.00	0.00	0.500	0.000	0.000	0.000	0.000	1.000	0.242	0.620	233.0	233.0
chain	17829.00	17829.00	87443.00	3634.00	0.500	0.169	-0.000	0.169	0.169	0.831	NaN	NaN	100.0	100.0
B008_DisoRDPbind-RNA	20239.00	6289.00	98983.00	1224.00	0.499	0.084	-0.004	0.163	0.057	0.940	0.150	0.425	233.0	233.0
B007_DisoRDPbind	6795.00	84954.00	20318.00	14668.00	0.438	0.242	-0.113	0.147	0.683	0.193	NaN	NaN	233.0	233.0

Where table column names mean:

label	meaning
BAc	balanced accuracy
F1s	F1-score
MCC	Matthew's Correlation Coefficient
Pre	Precision/Selectivity
Rec	Recall/Sensitivity
Rec_n	Specificity
AUC_ROC	Area under the ROC curve
AUC_PRC	Area under the PR curve
npred	number of predicted targets
nref	number of targets in reference

1.2.1 Balanced accuracy

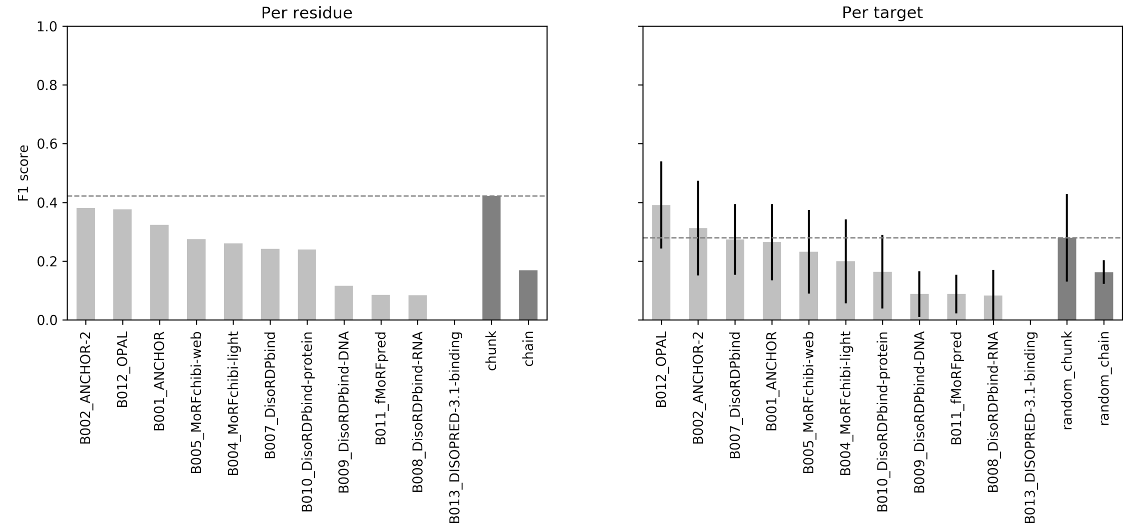
Comparison of predictors and baselines performance by balanced accuracy.



Overall (left panel) and average per-target (right panel) balanced accuracy. Light gray bars represent published prediction methods; dark gray bars represent baseline prediction methods. On the right panel standard deviation is shown as an error bar.

1.2.2 F1-score

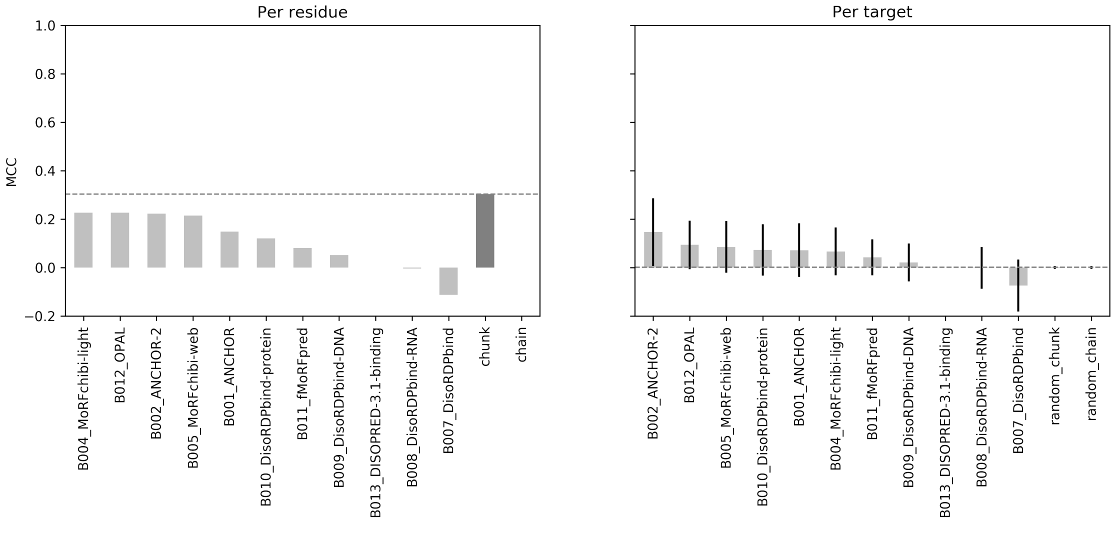
Comparison of predictors and baselines performance by F1-score.



Overall (left panel) and average per-target (right panel) F1-score. Light gray bars represent published prediction methods; dark gray bars represent baseline prediction methods. On the right panel standard deviation is shown as an error bar.

1.2.3 MCC

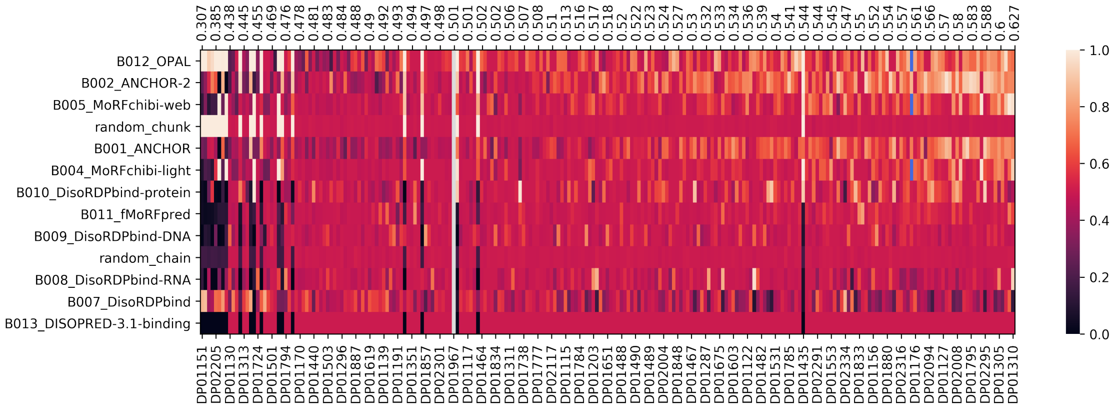
Comparison of predictors and baselines performance by Matthew's Correlation Coefficient.



Overall (left panel) and average per-target (right panel) MCC. Light gray bars represent published prediction methods; dark gray bars represent baseline prediction methods. On the right panel standard deviation is shown as an error bar.

1.2.4 Per target accuracy

Balanced accuracy score for each target for each prediction methods (including baslines)



Heatmap of the predictors accuracy for each target. The higher the accuracy the lighter the color. Non-predicted targets are shown in blue. x and y axes are sorted by average accuracy over rows and columns respectively. A white semi-transparent vertical line marks the point where the average accuracy scores for a target is below (left) or above (right) 0.5. Accuracy score approaches 0.5 for a random classifier. Accuracy < 0.5 indicates anti-correlation between predicted and reference classes. Targets with an average accuracy score < 0.5 are:

DP01151
DP01776
DP01313
DP01870
DP01794
DP01556
DP01503
DP01436
DP01619
DP01141
DP01351
DP01527

DP01471
DP01146
DP01474
DP01181
DP01185
DP01972
DP01319
DP01528
DP01578
DP01941
DP01643
DP01622

DP01942
DP01130
DP02066
DP01501
DP01787
DP01440
DP01499
DP01887
DP01462
DP01191
DP02326
DP02301

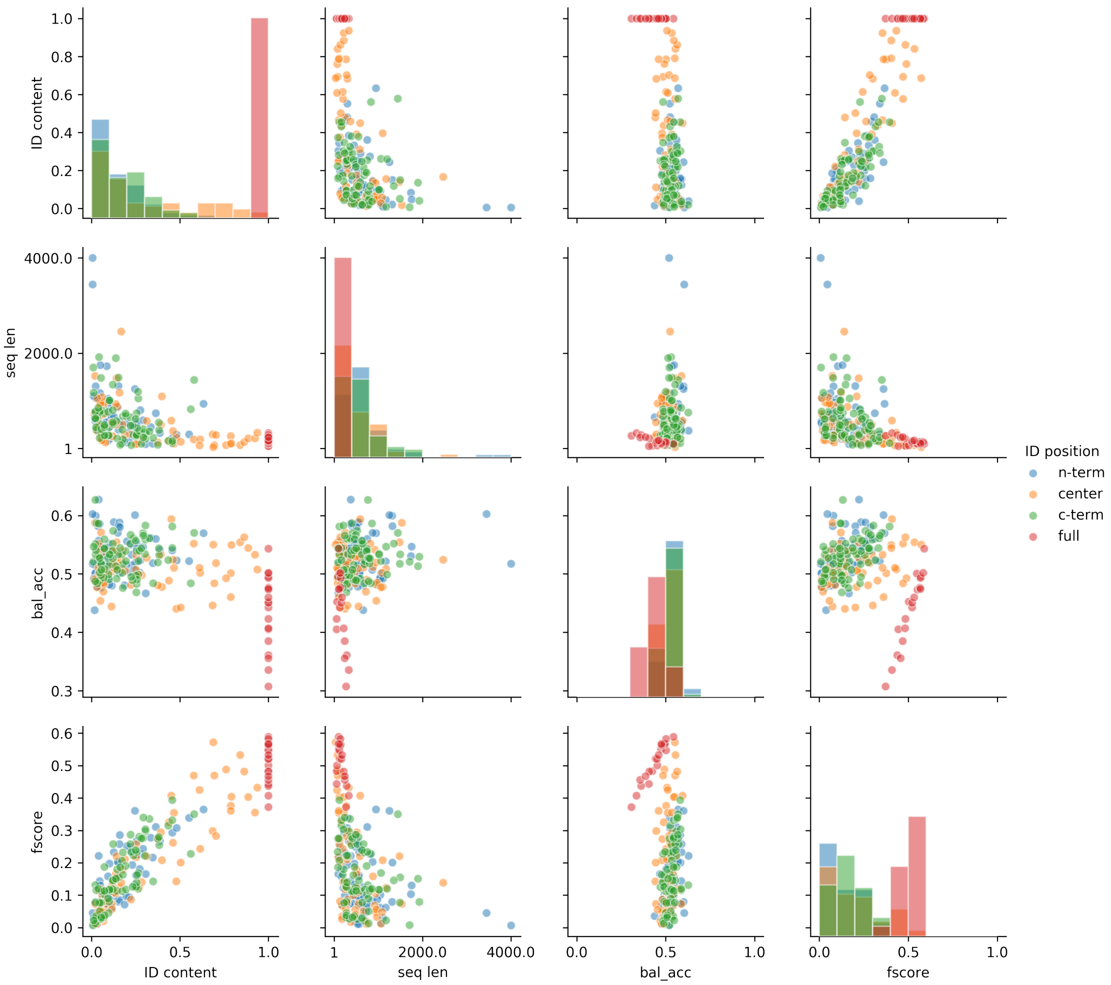
DP01521
DP01456
DP01148
DP01199
DP02120
DP01437
DP01288
DP01128
DP01325
DP01163
DP01147
DP01334

DP02205
DP01498
DP01724
DP01159
DP01170
DP01999
DP01296
DP01323
DP01139
DP01786
DP01857
DP02327

DP01512
DP01293
DP02078
DP01876
DP01990
DP01426
DP01507
DP01616
DP01893
DP01366
DP01500
DP01475

1.2.5 Target correlation matrix

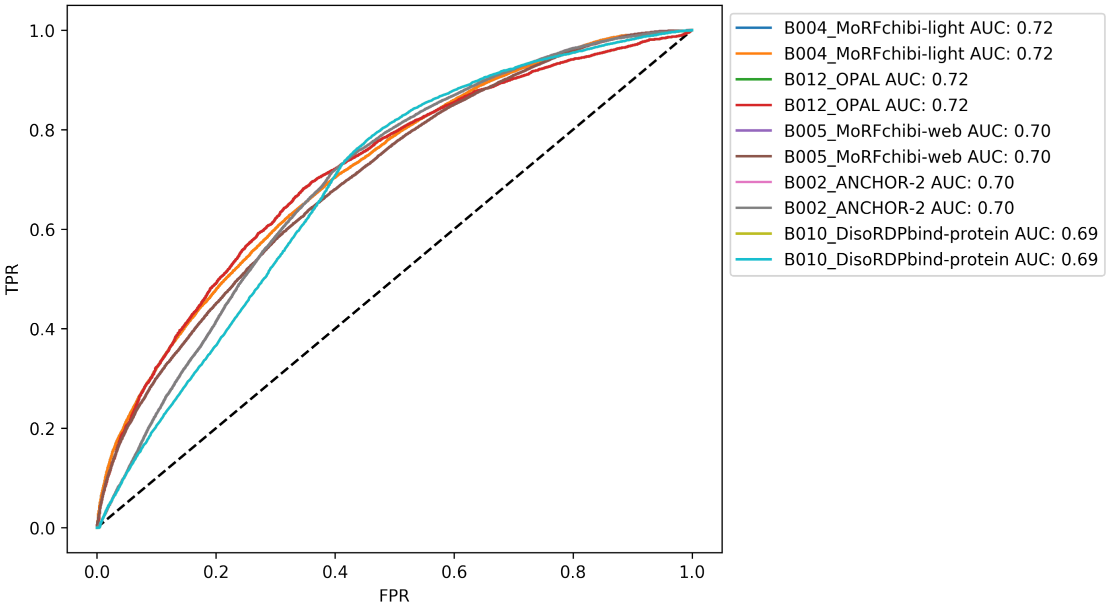
Commonly, experimental data has a bias for low disorder-content. DisProt targets have high disorder-content. Classifiers have been trained/engineered on low disorder-content. I expect difficult targets to have high disorder content. To verify is there is any correlation between target features I'm plotting 4 selected features against each other (Balanced accuracy, F1-Score, Sequence length and ID content). A fifth feature (ID position) divides the datasets in subsets. ID position is calculated as the average of the indexes of disordered residues along the sequence. A correlation is observed in a subset if its points gather around a diagonal.



Correlation matrix of Balanced accuracy, F1-Score, Sequence length and ID content. Average position of disorder is color-coded. Figure matrix is symmetrical. Plots along diagonal axis display single variables distributions. No meaningful correlation is observed.

1.2.6 ROC curve

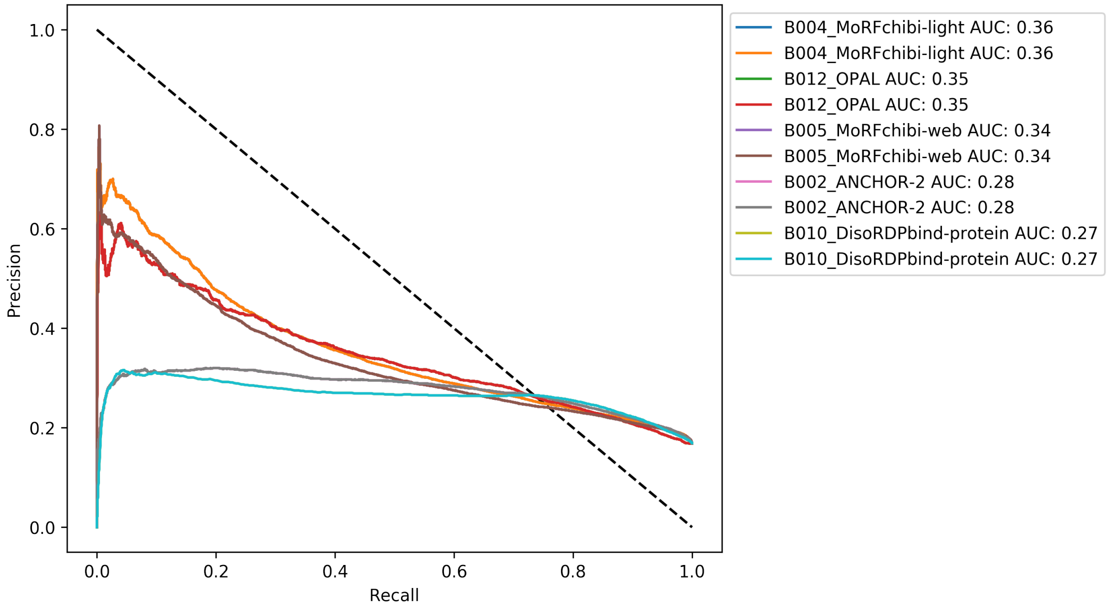
Receiver Operator Characteristic plot for predictors and baselines



False Positive Rate (FPR) on *x* axis, True Positive Rate (TPR) on *y* axis. Methods are sorted by their Area Under the Curve (AUC). Only first ten methods plus baselines are shown.

1.2.7 PR curve

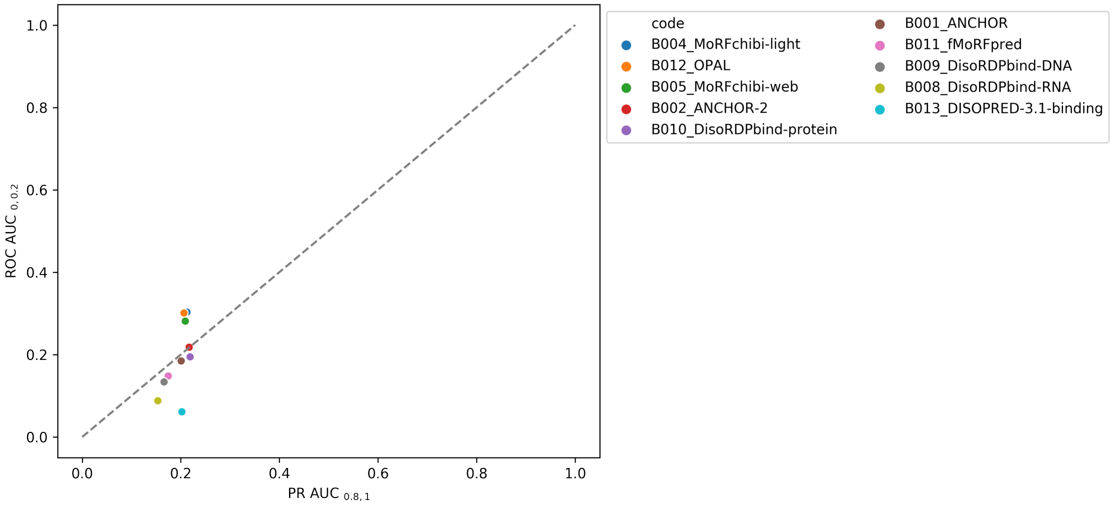
Precision Recall curve plot for predictors and baselines



Recall/Sensitivity on *x* axis, Precision/Selectivity on *y* axis. Methods are sorted by their Area Under the Curve (AUC). Only first ten methods plus baselines are shown.

1.2.8 pROC/pPR scatter plot

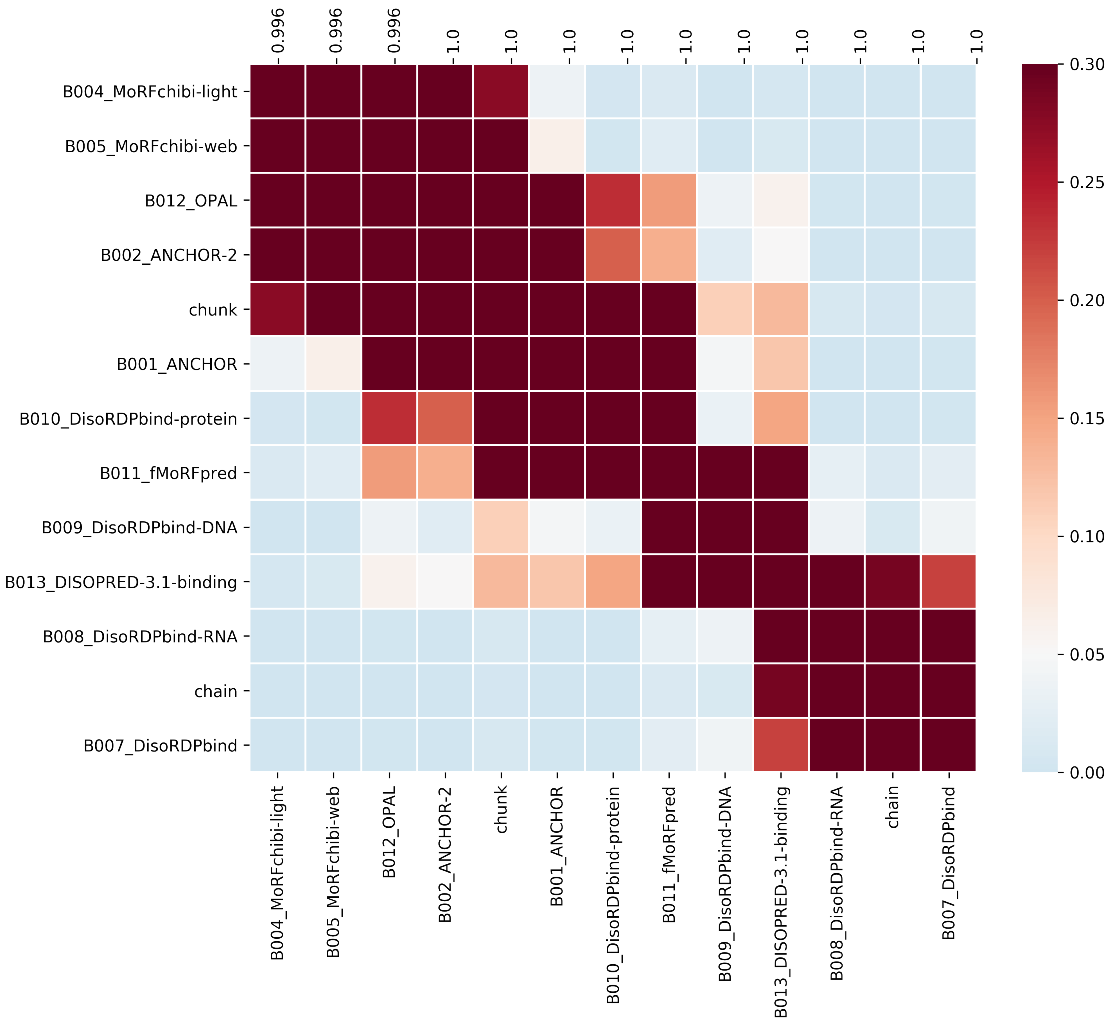
Plot of the AUCs from the ROC curve and PR curve



ROC AUC on the x axis, Precision-Recall (PR) AUC on the y axis. ROC AUC is calculated including ROC curve points with x values (FPR) between 0 and 0.2. PR AUC is calculated including PR curve points with x values (Recall) between 0.8 and 1. Both AUCs are then rescaled to the [0, 1] range.

1.2.9 Average overall ranking

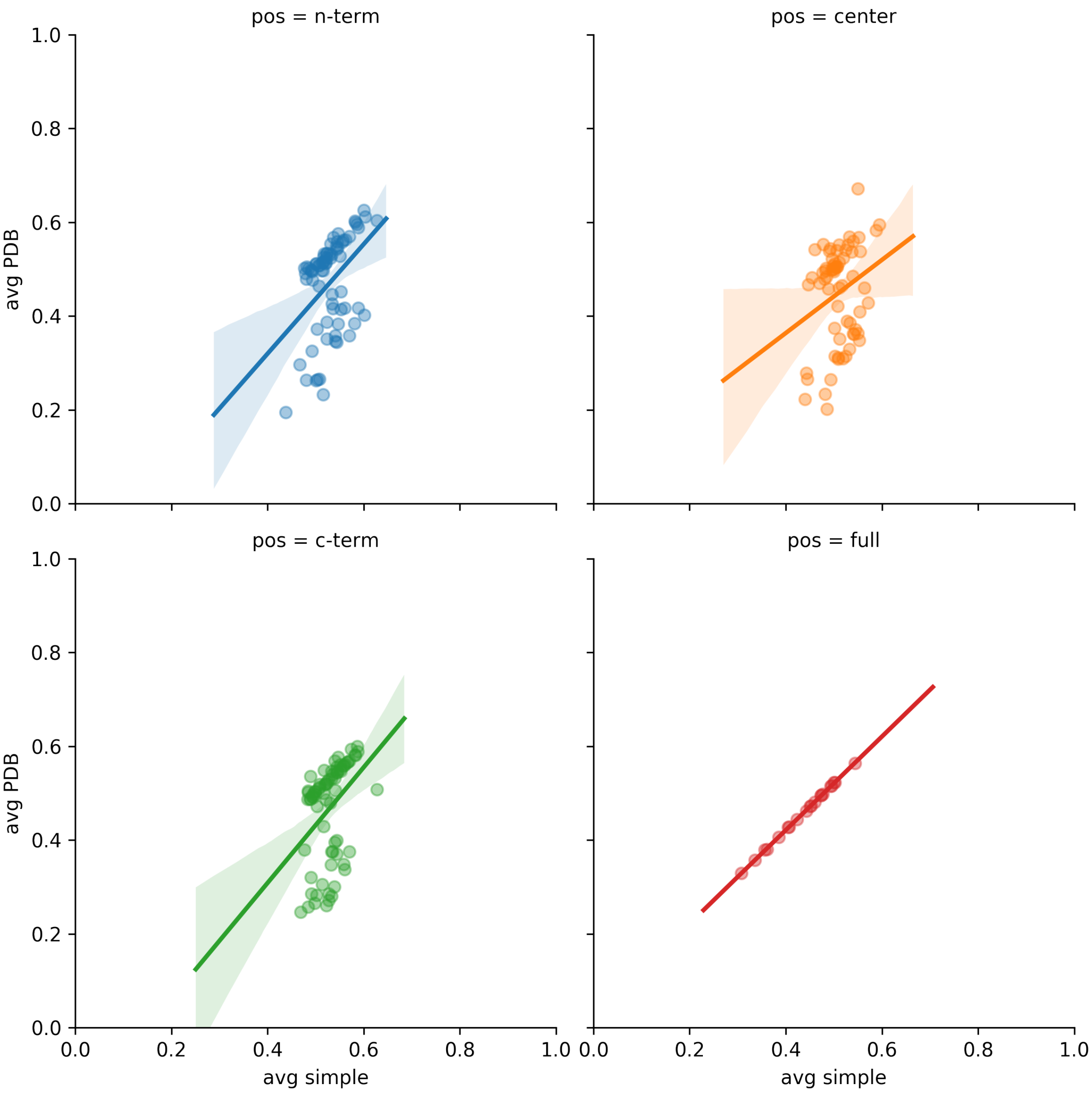
Predictor ranking and ranking statistical significance. Predictors are ranked on average rank from metrics scores: Balanced Accuracy, MCC, Precision, Recall, F1-score, F1-scores on negatives, Precision on negatives, Specificty, ROC AUC, PR AUC.



Heatmap of the p-value associated to the statistical significance of the difference between ranking distributions. Coloramap is centered on 0.05 so that any pvalue above 0.05 is **red**-ish. Red color indicates that the ranking difference between two predictors is not statistically significant. Top tick labels of *x* axis display prediction coverage for each predictor.

1.2.10 Accuracy correlation between datasets

Per target average balanced accuracy correlation between *simple* and *pdb* negative definition. Datasets is divided by average disorder position in targets.



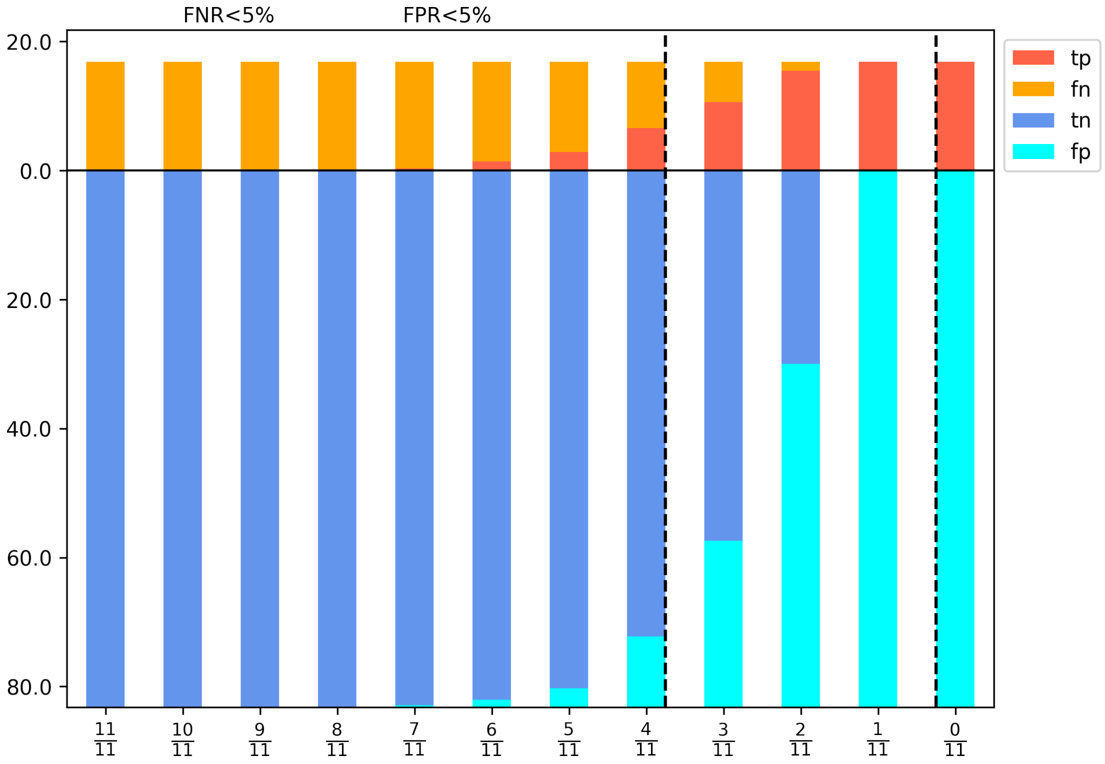
Average balanced accuracy for targets with reference negative defined by the *simple* rule on *x* axis. Average balanced accuracy for targets with reference negative defined by the *pdb* rule on *y* axis. Each panel includes only targets with a specific average disorder position (C-terminal, N-Terminal, central, full-disorder)

1.3 Consensus

Consensus among all prediction methods was calculated as the fraction of positive predictions per residue.

1.3.1 Confusion matrix per threshold

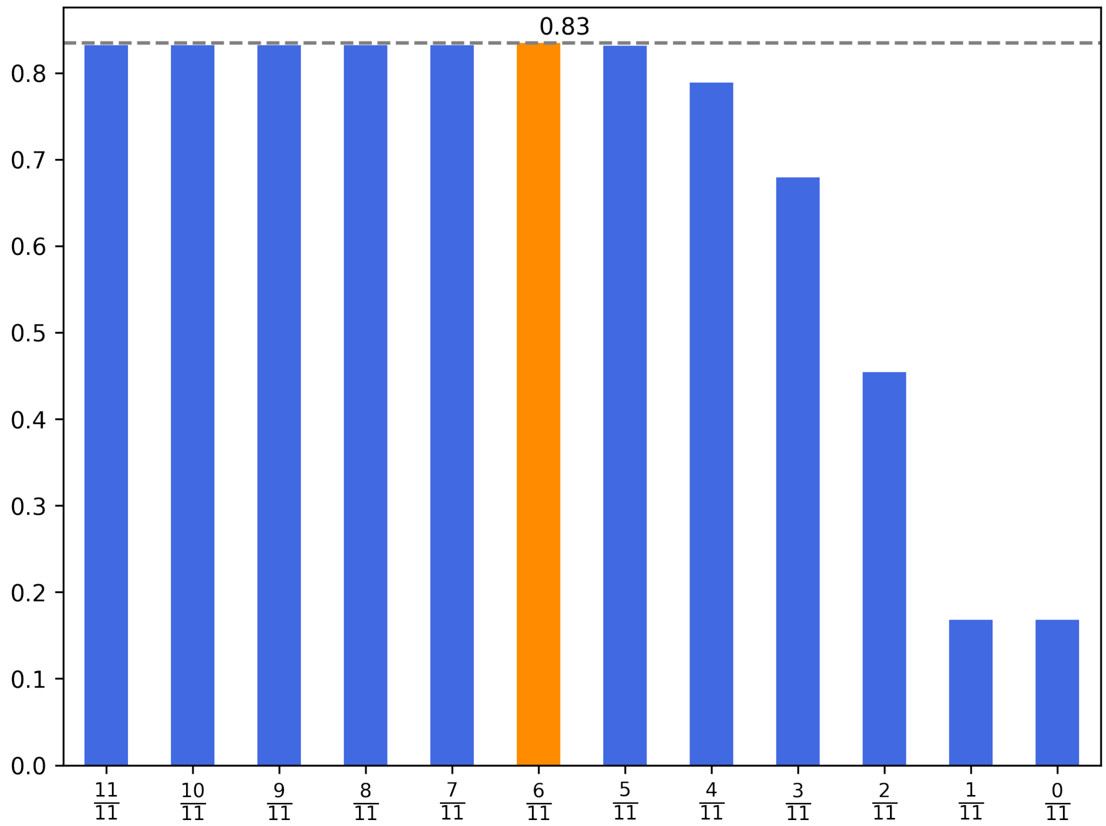
Predicted and actual positive and negatives for each threshold on the consensus score.



Percentage of correct and wrong assignment of positives (above 0) and negatives (below 0) for each threshold of the consensus.

1.3.2 Accuracy per threshold

Balanced accuracy score for each threshold of the consensus.



Accuracy distribution for each consensus threshold. Bar of max threshold is highlighted in orange.

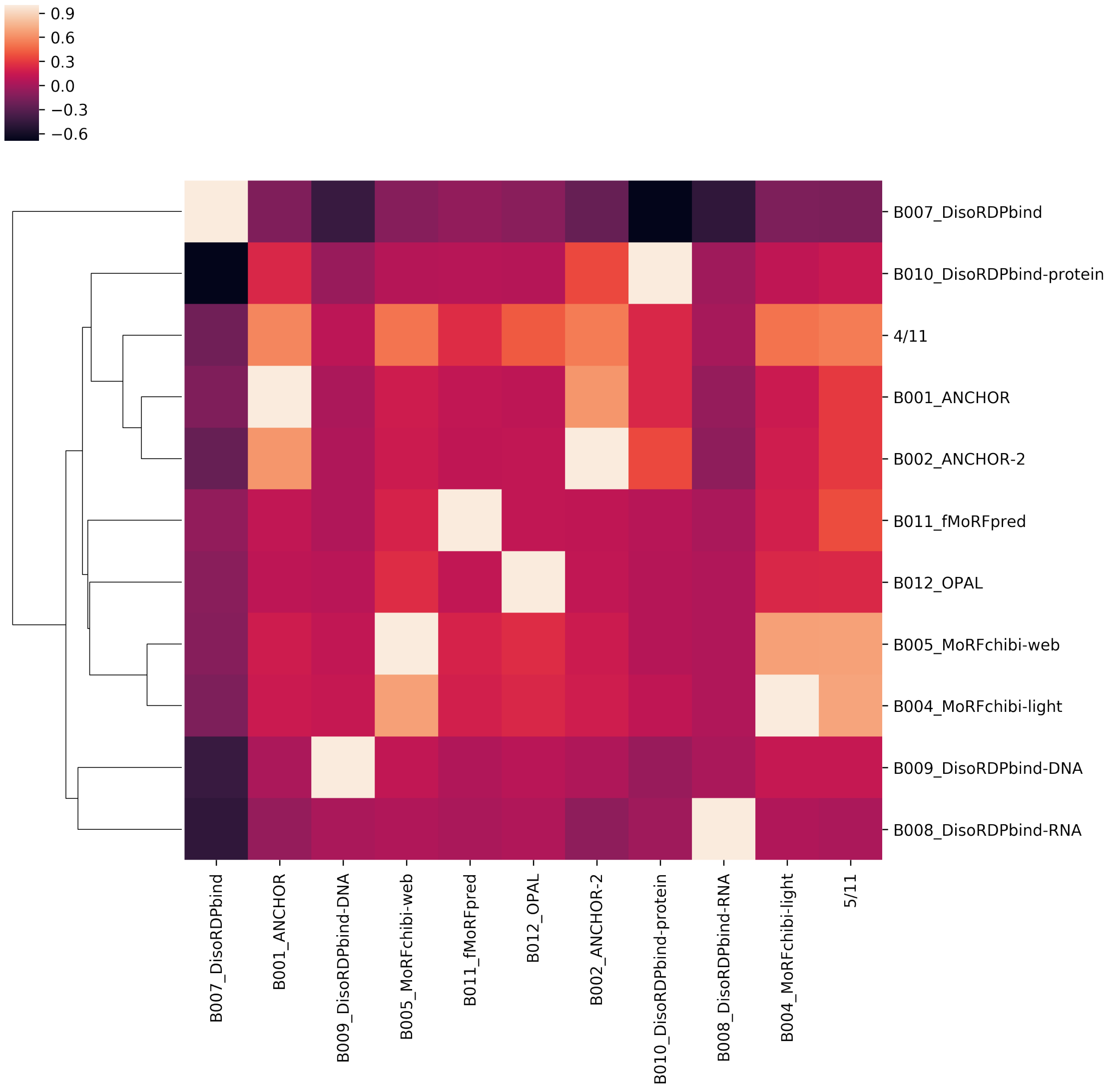
1.3.3 Percentage of correct/incorrect classifications (missing)

Percentage of correct and incorrect classifications for nositives (defined by DisProt), negatives (defined by PDB) and undefined residues for each predictor.

1.3.4 clustermap of binary predictions correlation

Correlation of binary states between predictors.

```
/home/marnec/Projects/CAID/caid/.venv/lib/python3.6/site-packages/scipy/stats/stats.py:3399: PearsonRCo
  warnings.warn(PearsonRConstantInputWarning())
```



Heatmap of the correlation of binary prediction states for each couple of predictors. Pearson R is calculated between all predictions. Clustering is based on Euclidean distance calculated over an array (column) of R correlation coefficients.

1.4 Fully disordered targets

Statistics calculated for the subset of targets that are reported as completely disordered in DisProt.

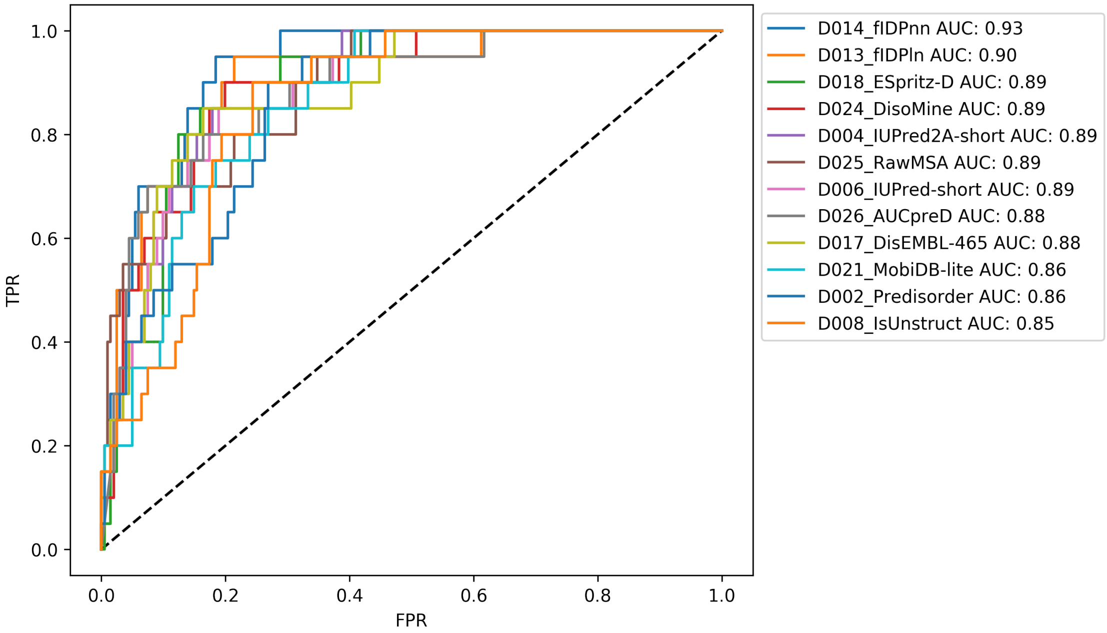
1.4.1 Correctly and incorrectly classified full IDPs

Number of correctly and incorrectly classified full IDPs with a prediction tolerance of 5%.

Actual	Positives		Negatives	
Predicted	TP	FN	FP	TN
B011_fMoRFpred	20	0	0	213
random_chain	20	0	0	213
random_chunk	20	0	0	213
B005_MoRFchibi-web	20	0	6	207
B004_MoRFchibi-light	20	0	7	206
B007_DisoRDPbind	20	0	26	187
B012_OPAL	20	0	54	159
B001_ANCHOR	19	1	0	213
B002_ANCHOR-2	18	2	1	212
B010_DisoRDPbind-protein	16	4	1	212
B009_DisoRDPbind-DNA	14	6	1	212
B008_DisoRDPbind-RNA	11	9	0	213
B013_DISOPRED-3.1-binding	0	20	0	213

1.4.2 Full IDPs ROC

ROC for the classification power of Full IDPs. Average disorder scores for each target is compared to full IDPs (positives) and partial IDPs (negatives). 5% prediction tolerance is applied.



FPR on the *x* axis, TPR on the *y* axis. Methods are sorted by their AUC. Only first 12 methods are shown.