CAID

July 17, 2019

Contents

1 CAID **2**
- 1.1 Dataset . . . 2
- 1.2 Evaluation metrics . . . 3
  - 1.2.1 Balanced accuracy . . . 3
  - 1.2.2 F1-score . . . 4
  - 1.2.3 MCC . . . 5
  - 1.2.4 Per target accuracy . . . 5
  - 1.2.5 Target correlation matrix . . . 6
  - 1.2.6 ROC curve . . . 7
  - 1.2.7 PR curve . . . 8
  - 1.2.8 pROC/pPR scatter plot . . . 9
  - 1.2.9 Average overall ranking . . . 9
- 1.3 Prediction bias in undefined regions . . . 10
  - 1.3.1 Accuracy correlation between datasets . . . 11
- 1.4 Consensus . . . 11
  - 1.4.1 Confusion matrix per threshold . . . 12
  - 1.4.2 Accuracy per threshold . . . 12
  - 1.4.3 Percentage of correct/incorrect classifications . . . 13
  - 1.4.4 clustermap of binary predictions correlation . . . 14
- 1.5 Fully disordered targets . . . 15
  - 1.5.1 Correctly and incorrectly classified full IDPs . . . 15
  - 1.5.2 Full IDPs ROC . . . 16

Chapter 1

CAID

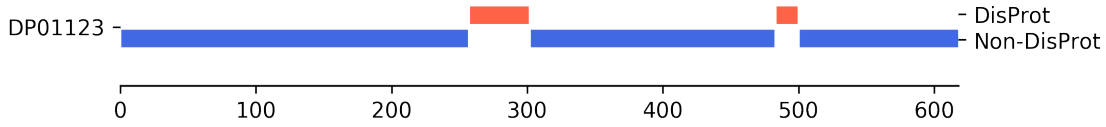
1.1 Dataset

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Disorder (CAID) is a continuous experiment where prediction methods for intrinsic disorder (ID) are blind tested on unpublished DisProt data.

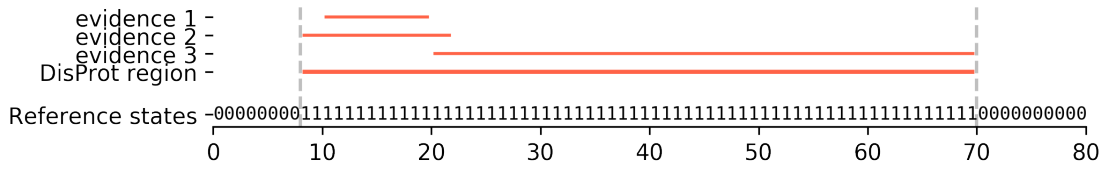
```
Accordion(children=(RadioButtons(description='Reference:', index=6, options=('new-disprot-linker', 'new
```

```
ToggleButtons(description='Plot style:', index=1, options=('Explorative', 'Slides', 'Publication'), valu
```

Current analysis is performed on the **new-disprot-all** dataset with **simple** negative definition. This means that DisProt defines order for *all its new entries* 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
*SPOT-Disorder2	11513	70732	150676	36238	0.720	0.468	0.343	0.339	0.759	0.681	0.340	0.760	610	646
*RawMSA	17363	75205	206786	37241	0.708	0.446	0.325	0.331	0.682	0.733	0.414	0.780	646	646
*SPOT-Disorder1	13693	93968	187481	40600	0.707	0.430	0.311	0.302	0.748	0.666	0.268	0.744	644	646
*AUCpreD	17070	81154	193350	37485	0.696	0.433	0.303	0.316	0.687	0.704	0.479	0.757	644	646
Predisorder	10526	111932	156268	43971	0.695	0.418	0.292	0.282	0.807	0.583	0.325	0.747	642	646
IsUnstruct	13771	102404	179587	40833	0.692	0.413	0.287	0.285	0.748	0.637	0.323	0.744	646	646
IUPred-long	19362	78821	203098	35198	0.683	0.418	0.285	0.309	0.645	0.720	0.298	0.737	645	646
VSL2B	10059	127742	153707	44234	0.680	0.391	0.266	0.257	0.815	0.546	0.301	0.732	644	646
IUPred2A-long	19974	77265	204726	34630	0.680	0.416	0.282	0.309	0.634	0.726	0.298	0.735	646	646
fIDPln	27013	41412	240579	27505	0.679	0.446	0.327	0.399	0.505	0.853	0.422	0.793	645	646
JRONN	14112	110567	171352	40448	0.675	0.394	0.259	0.268	0.741	0.608	0.302	0.724	645	646
AUCpreD-np	23312	63724	218267	31292	0.674	0.418	0.284	0.329	0.573	0.774	0.428	0.751	646	646
ESpritz-N	17054	97801	184118	37506	0.670	0.395	0.256	0.277	0.687	0.653	0.296	0.714	645	646
DisoMine	23260	65966	216025	31344	0.670	0.413	0.277	0.322	0.574	0.766	0.388	0.765	646	646
IUPred2A-short	23880	63323	218668	30724	0.669	0.413	0.278	0.327	0.563	0.775	0.313	0.741	646	646
IUPred-short	23891	63865	218054	30669	0.668	0.411	0.275	0.324	0.562	0.773	0.311	0.739	645	646
*SPOT-Disorder-Single	24220	61966	220025	30384	0.668	0.414	0.278	0.329	0.556	0.780	0.318	0.757	646	646
PyHCA	18648	96104	185887	35956	0.659	0.385	0.240	0.272	0.658	0.659	0.277	0.706	646	646
ESpritz-X	25551	60534	221385	29009	0.658	0.403	0.264	0.324	0.532	0.785	0.304	0.740	645	646
*DISOPRED-3.1	19395	92549	189442	35209	0.658	0.386	0.241	0.276	0.645	0.672	0.290	0.701	646	646
chunk	32511	32511	249479	22092	0.645	0.405	0.289	0.405	0.405	0.885	NaN	NaN	100	100
FoldUnfold	17940	106325	172346	35715	0.642	0.365	0.211	0.251	0.666	0.618	NaN	NaN	621	646
naive-pdb	6072	171438	109201	48119	0.639	0.352	0.215	0.219	0.888	0.389	0.528	0.634	630	646
naive-gene3d	6767	168792	112552	47544	0.638	0.351	0.212	0.220	0.875	0.400	0.547	0.643	639	646
MobiDB-lite	32125	40578	241341	22435	0.634	0.382	0.253	0.356	0.411	0.856	0.366	0.737	645	646
ESpritz-D	35346	24803	257116	19214	0.632	0.390	0.289	0.437	0.352	0.912	0.410	0.774	645	646
fIDPnn	38167	13462	268529	16351	0.626	0.388	0.327	0.548	0.300	0.952	0.475	0.814	645	646
DisEMBL-465	33266	46111	235338	21027	0.612	0.346	0.206	0.313	0.387	0.836	0.283	0.685	644	646
DisEMBL-HL	25527	90177	191272	28766	0.605	0.332	0.161	0.242	0.530	0.680	0.274	0.654	644	646
*S2D-2	7710	183502	97947	46583	0.603	0.328	0.163	0.202	0.858	0.348	0.229	0.654	644	646
*S2D-1	7237	191026	90423	47056	0.594	0.322	0.152	0.198	0.867	0.321	0.253	0.672	644	646
*DisPredict-2	31912	67407	214584	22692	0.588	0.314	0.147	0.252	0.416	0.761	0.250	0.637	646	646
GlobPlot	33642	62430	219489	20918	0.581	0.303	0.138	0.251	0.383	0.779	0.231	0.624	645	646
cons	4676	245470	36521	49928	0.522	0.285	0.049	0.169	0.914	0.130	0.193	0.567	646	646
DynaMine	53609	3541	278378	951	0.502	0.032	0.016	0.212	0.017	0.987	0.271	0.707	645	646
chain	45733	45733	236257	8870	0.500	0.162	0.000	0.162	0.162	0.838	NaN	NaN	100	100
DFLpred	49496	30326	251665	5108	0.493	0.113	-0.017	0.144	0.094	0.892	0.142	0.410	646	646

Where table column names mean:

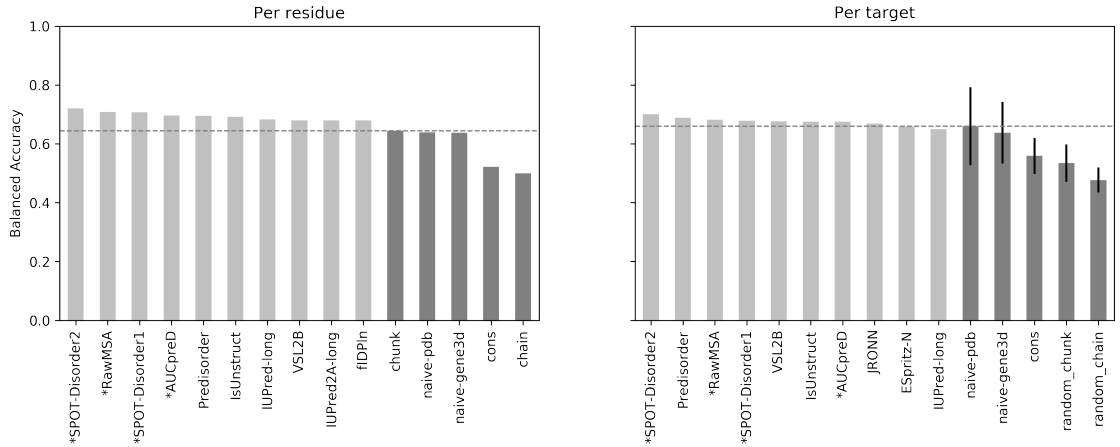
label	meaning
BAc	balanced accuracy
F1s	F1-score
MCC	Matthew's Correlation Coefficient
Pre	Precision/Selectivity
Rec	Recall/Sensitivity
Rec_n	Specificity
AUC_ROC	Area under the ROC curve
AUC_PRC	Area under the PR curve
npred	number of predicted targets
nref	number of targets in reference

1.2.1 Balanced accuracy

Comparison of predictors and baselines performance by balanced accuracy.

```
/home/marnec/.local/share/virtualenvs/caid-ICjYQIts/lib/python3.6/site-packages/ipykernel_launcher.py:3
Passing list-likes to .loc or [] with any missing label will raise
KeyError in the future, you can use .reindex() as an alternative.
```

```
See the documentation here:
https://pandas.pydata.org/pandas-docs/stable/indexing.html#deprecate-loc-reindex-listlike
```

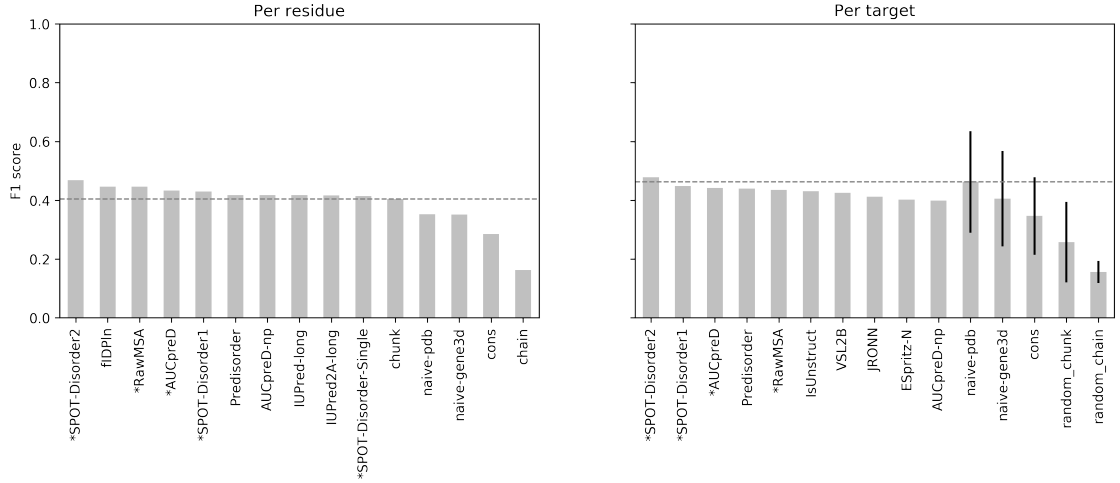


Overall (left panel) and average per-target (right panel) balanced accuracy. Light gray bars represent published prediction methods; dark gray bars represent baseline prediction methods. On the right panel standard deviation is shown as an error bar.

1.2.2 F1-score

Comparison of predictors and baselines performance by F1-score.

```
/home/marnec/.local/share/virtualenvs/caid-ICjYQIts/lib/python3.6/site-packages/ipykernel_launcher.py:3
Passing list-likes to .loc or [] with any missing label will raise
KeyError in the future, you can use .reindex() as an alternative.

See the documentation here:
https://pandas.pydata.org/pandas-docs/stable/indexing.html#deprecate-loc-reindex-listlike
```

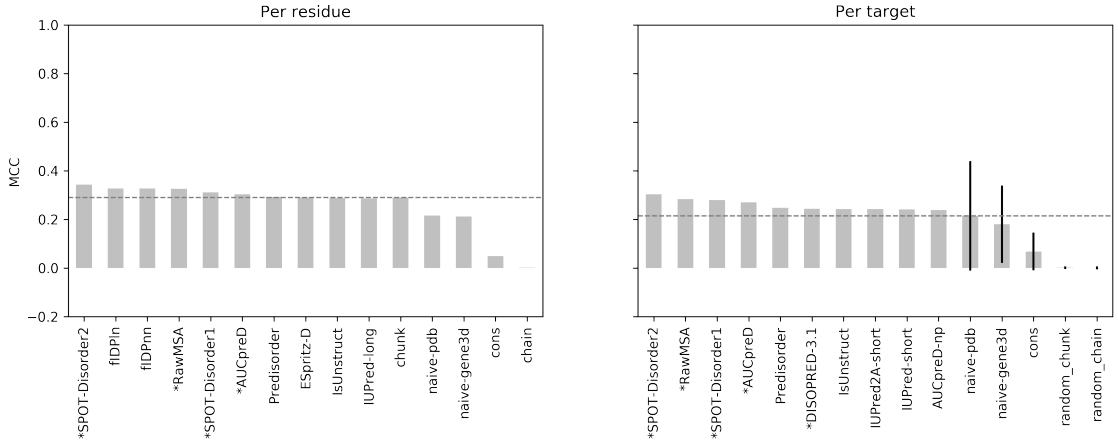


Overall (left panel) and average per-target (right panel) F1-score. Light gray bars represent published prediction methods; dark gray bars represent baseline prediction methods. On the right panel standard deviation is shown as an error bar.

1.2.3 MCC

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

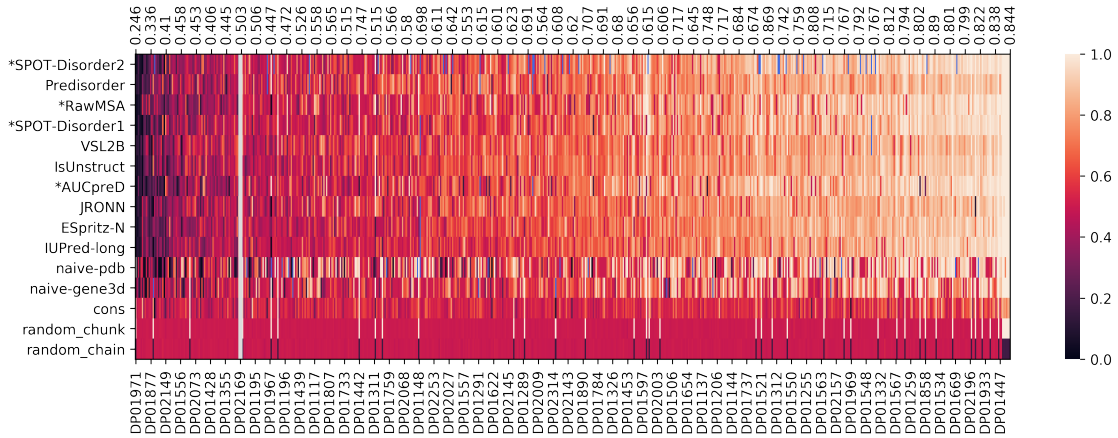
```
/home/marnec/.local/share/virtualenvs/caid-ICjYQIts/lib/python3.6/site-packages/ipykernel_launcher.py:2
Passing list-likes to .loc or [] with any missing label will raise
KeyError in the future, you can use .reindex() as an alternative.

See the documentation here:
https://pandas.pydata.org/pandas-docs/stable/indexing.html#deprecate-loc-reindex-listlike
```



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)

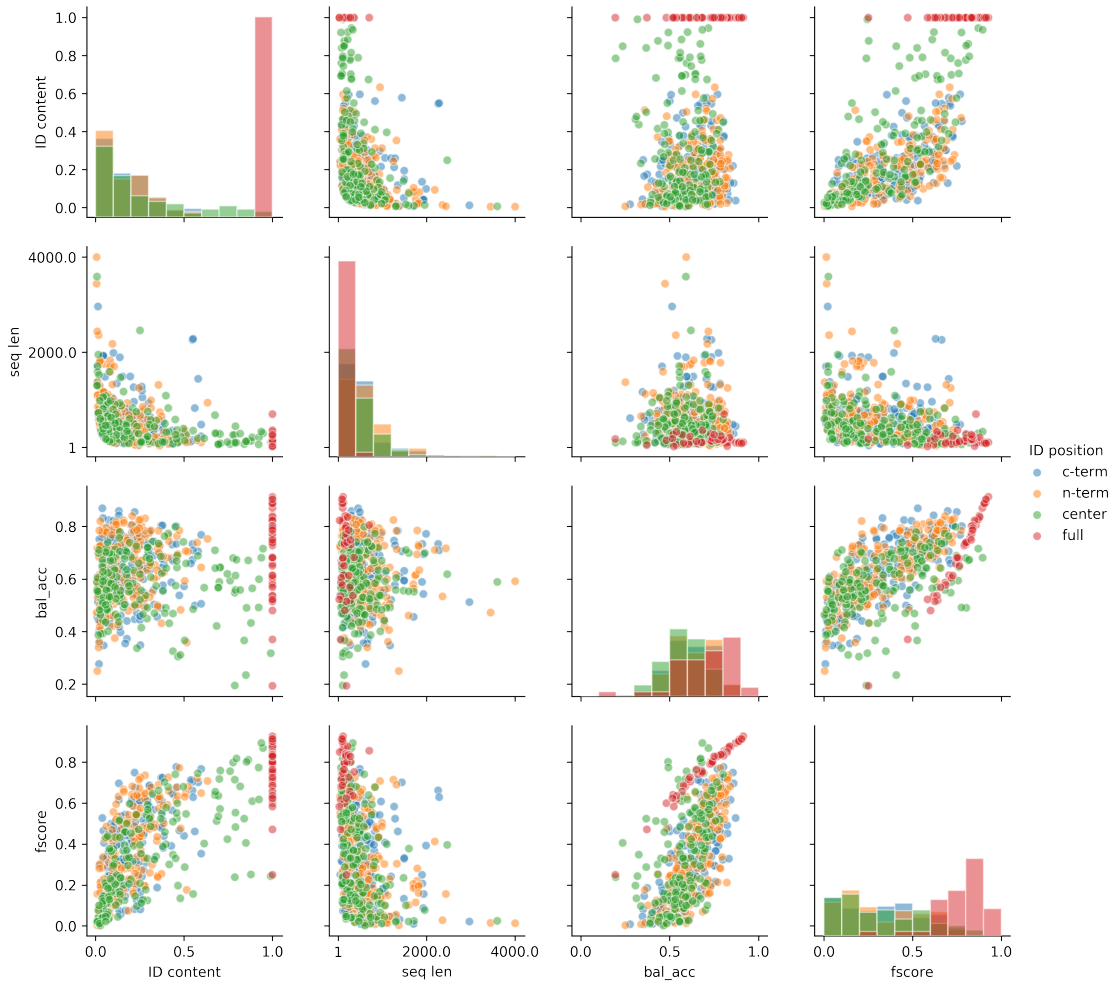


Heatmap of the predictors accuracy for each target. The higher the accuracy the lighter the color. Non-predicted targets are shown in blue. x and y axes are sorted by average accuracy over rows and columns respectively. A white semi-transparent vertical line marks the point where the average accuracy 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:

DP01971	DP01870	DP02010	DP01898	DP01278	DP01432
DP01281	DP01456	DP01128	DP01427	DP01130	DP01877
DP01494	DP01512	DP02334	DP01501	DP01203	DP01584
DP01145	DP01498	DP01462	DP01248	DP02149	DP01124
DP01844	DP01366	DP01141	DP01907	DP01407	DP01316
DP01724	DP01139	DP01883	DP01556	DP01285	DP02234
DP01474	DP01925	DP01185	DP02328	DP01878	DP01590
DP02073	DP01163	DP02324	DP02231	DP01774	DP01280
DP01978	DP01500	DP01477	DP01647	DP01428	DP01504
DP01612	DP01999	DP01772	DP01134	DP01551	DP01505
DP02086	DP01503	DP01355	DP01600	DP01309	DP01324
DP02247	DP01313	DP02168	DP01323	DP01499	DP01434
DP01749	DP01172	DP01430	DP01252	DP01887	DP02296
DP01528	DP01396	DP01914	DP01150	DP01364	DP01473
DP02326	DP01869	DP01967	DP02025	DP01489	DP01854
DP01110	DP01196	DP01762	DP01177	DP01187	DP01437
DP01339	DP01580	DP02001	DP01351	DP02301	DP01658

1.2.5 Target correlation matrix

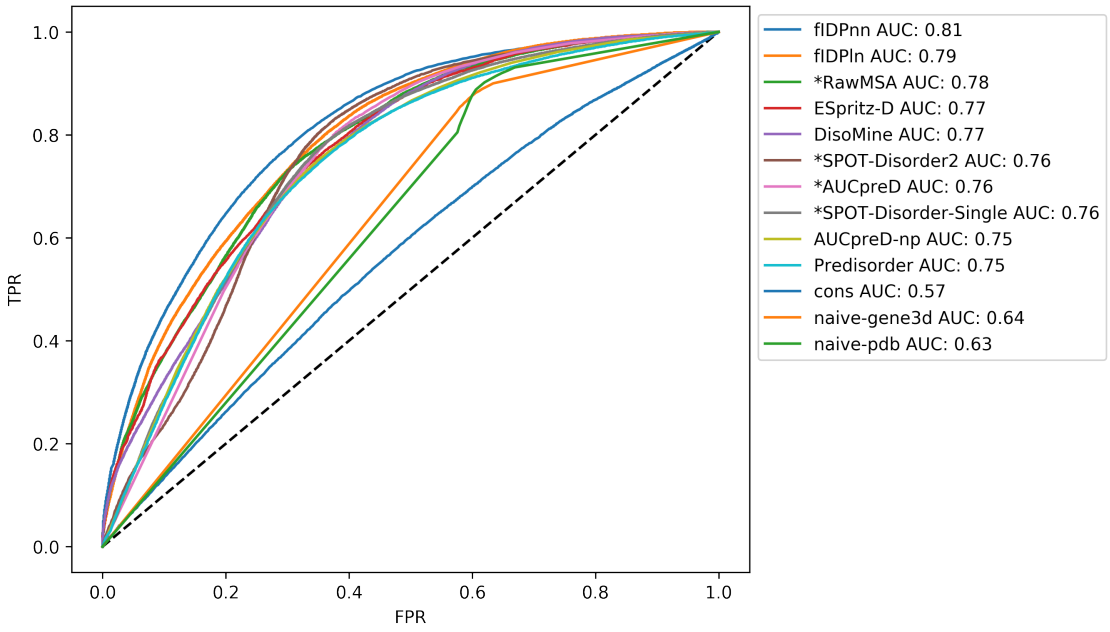
Commonly, experimental data has a bias for low disorder-content. DisProt targets have high disorder-content. Classifiers have been trained/engineered on low disorder-content. I expect difficult targets to have high disorder content. To verify is there is any correlation between target features I'm plotting 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

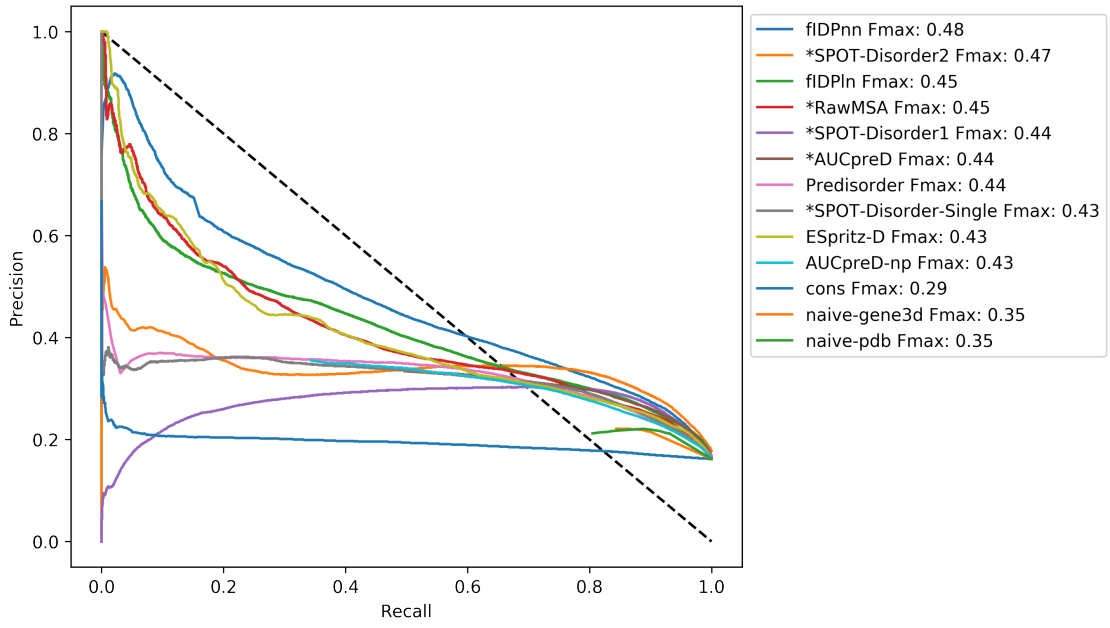


False Positive Rate (FPR) on x axis, True Positive Rate (TPR) on y axis. Methods are sorted by their Area Under the Curve (AUC). Only first ten methods plus baselines are shown.

1.2.7 PR curve

Precision Recall curve plot for predictors and baselines

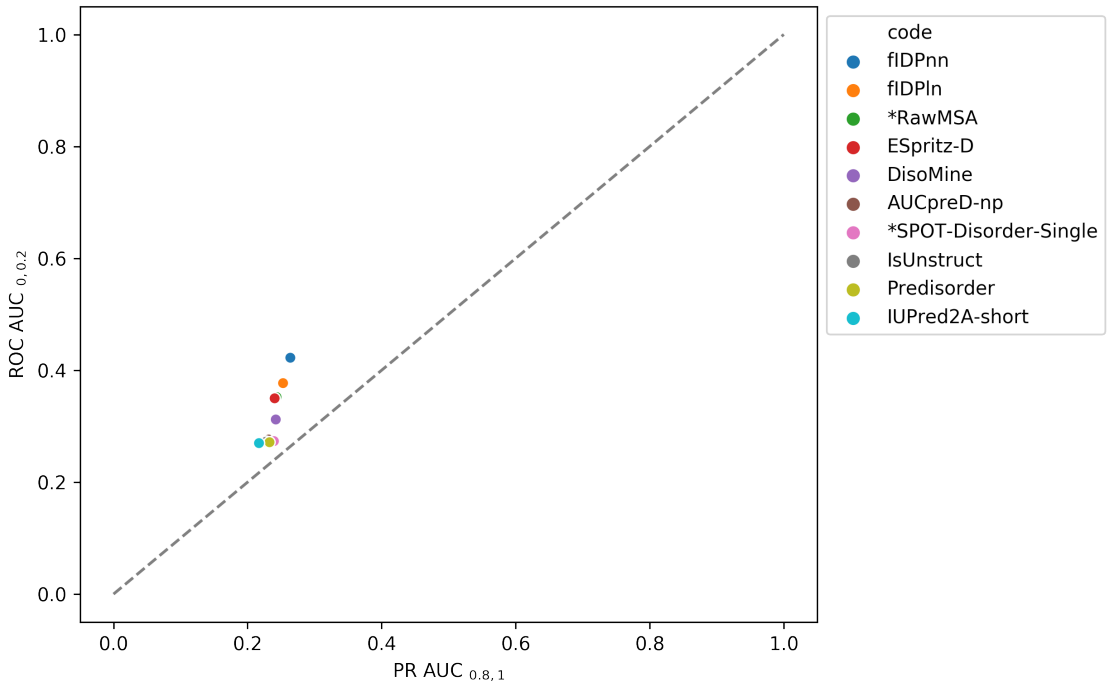
```
/home/marnec/.local/share/virtualenvs/caid-ICjYQIts/lib/python3.6/site-packages/ipykernel_launcher.py:7
  import sys
```



Recall/Sensitivity on x axis, Precision/Selectivity on y axis. Methods are sorted by their Fmax. Only first ten methods plus baselines are shown.

1.2.8 pROC/pPR scatter plot

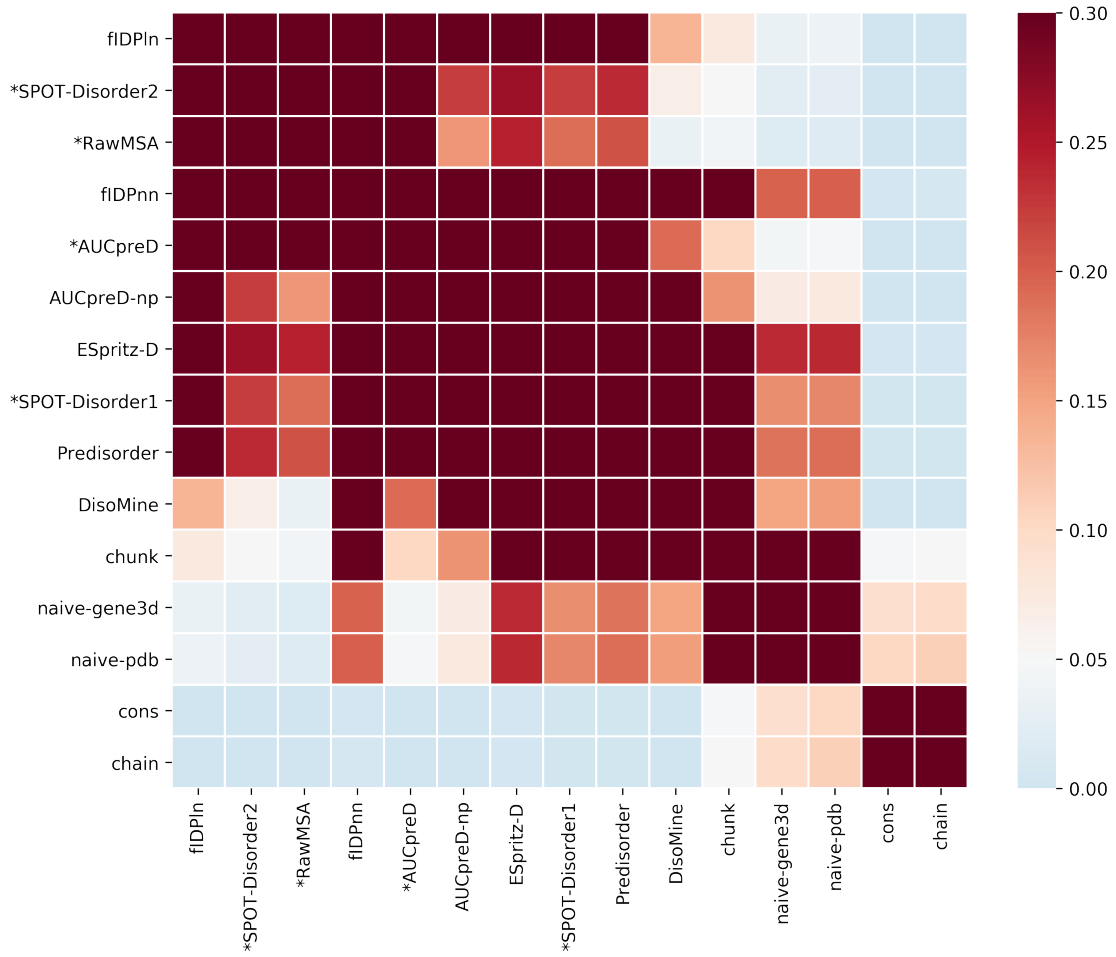
Plot of the AUCs from the ROC curve and PR curve



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.

1.3 Prediction bias in undefined regions

Comparison of accuracy on disorder regions (DisProt), structured regions (PDB) and prediction bias in undefined regions

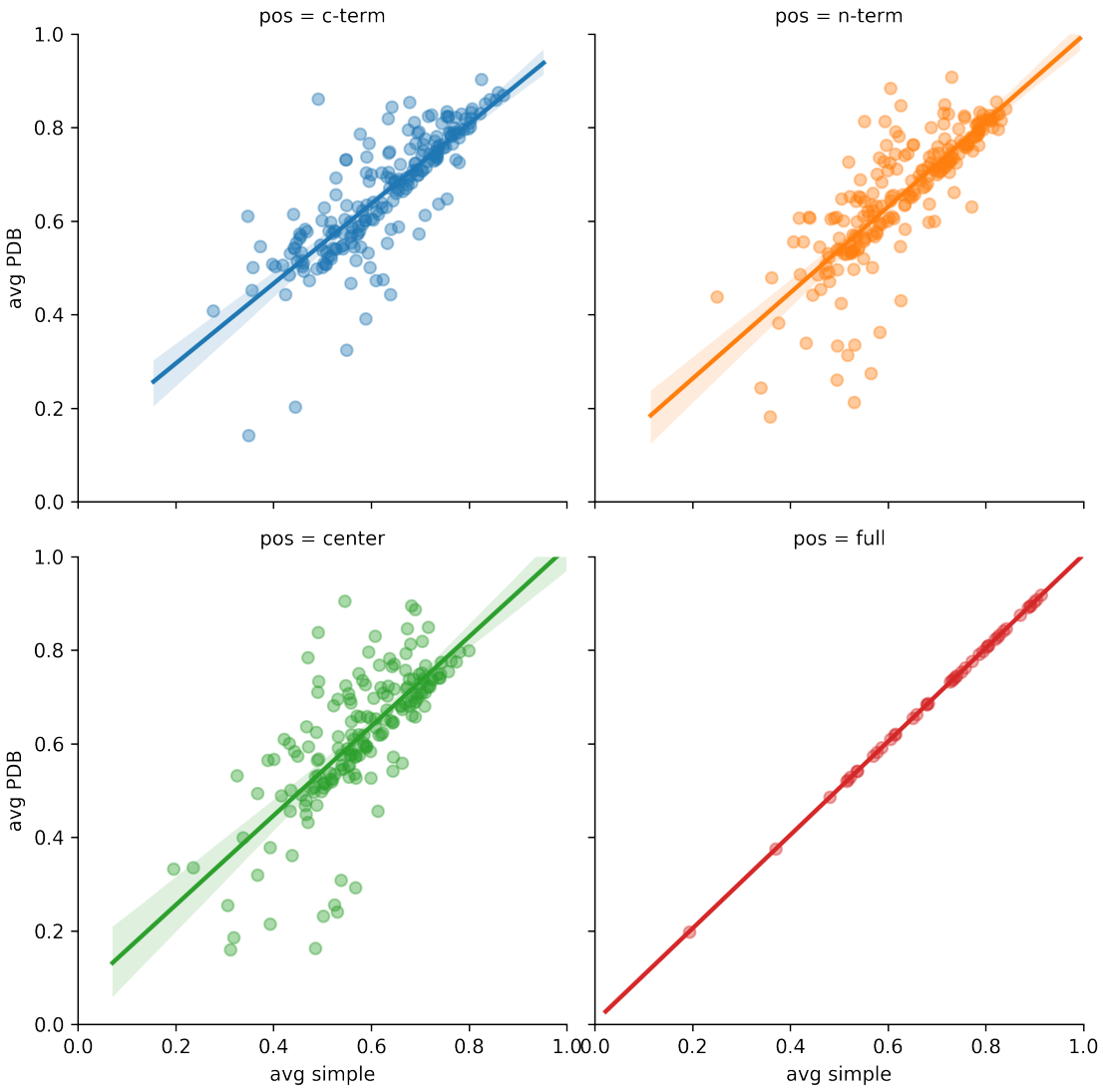
	Acc disorder	Acc order	Bias (%)
*SPOT-Disorder2	0.720	0.769	38.378
*SPOT-Disorder1	0.707	0.749	51.063
*AUCpreD	0.696	0.730	44.459
Predisorder	0.695	0.723	54.112
*RawMSA	0.708	0.689	39.400
IsUnstruct	0.692	0.704	50.708
VSL2B	0.680	0.702	59.809
IUPred-long	0.683	0.690	41.128
*DISOPRED-3.1	0.658	0.734	50.077
IUPred2A-long	0.680	0.688	40.434

Accuracy on disorder regions (DisProt), order regions (PDB) and prediction bias in undefined regions

calculated as the percentage of undefined residues predicted as disorder. Table sorted by the mean between the two accuracies.

1.3.1 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.



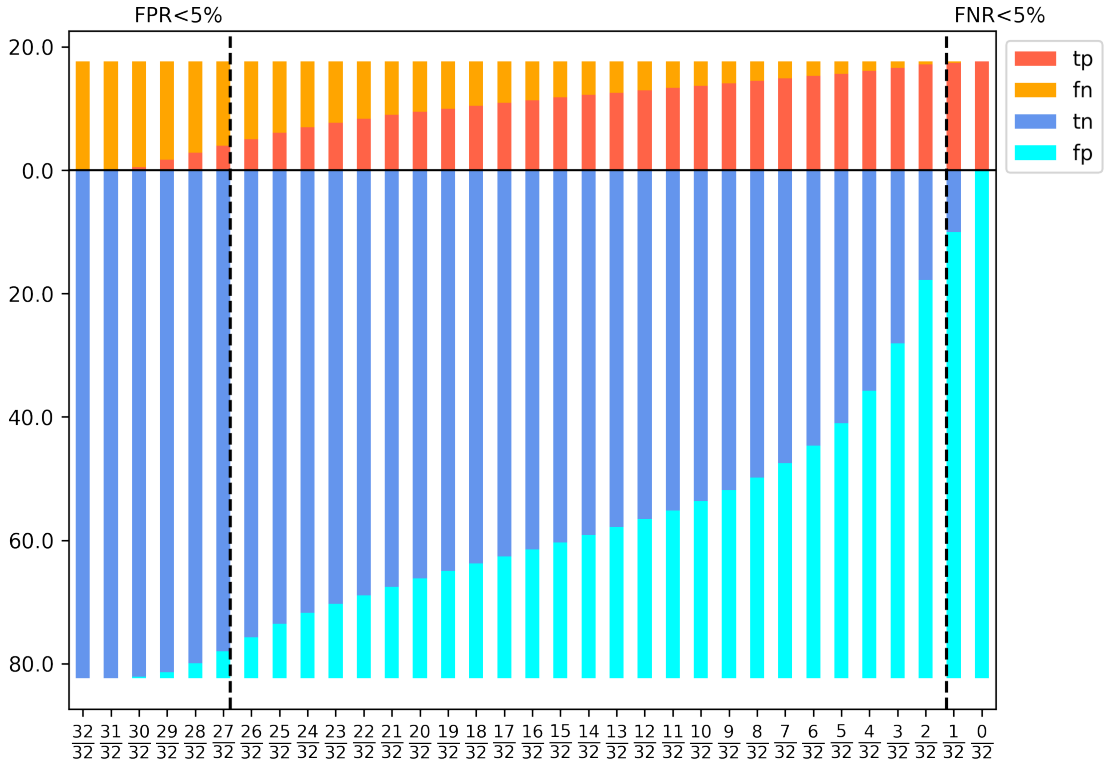
Average balanced accuracy for targets with reference negative defined by the *simple* rule on x axis. Average balanced accuracy for targets with reference negative defined by the *pdb* rule on y axis. Each panel includes only targets with a specific average disorder position (C-terminal, N-Terminal,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

1.4 Consensus

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

1.4.1 Confusion matrix per threshold

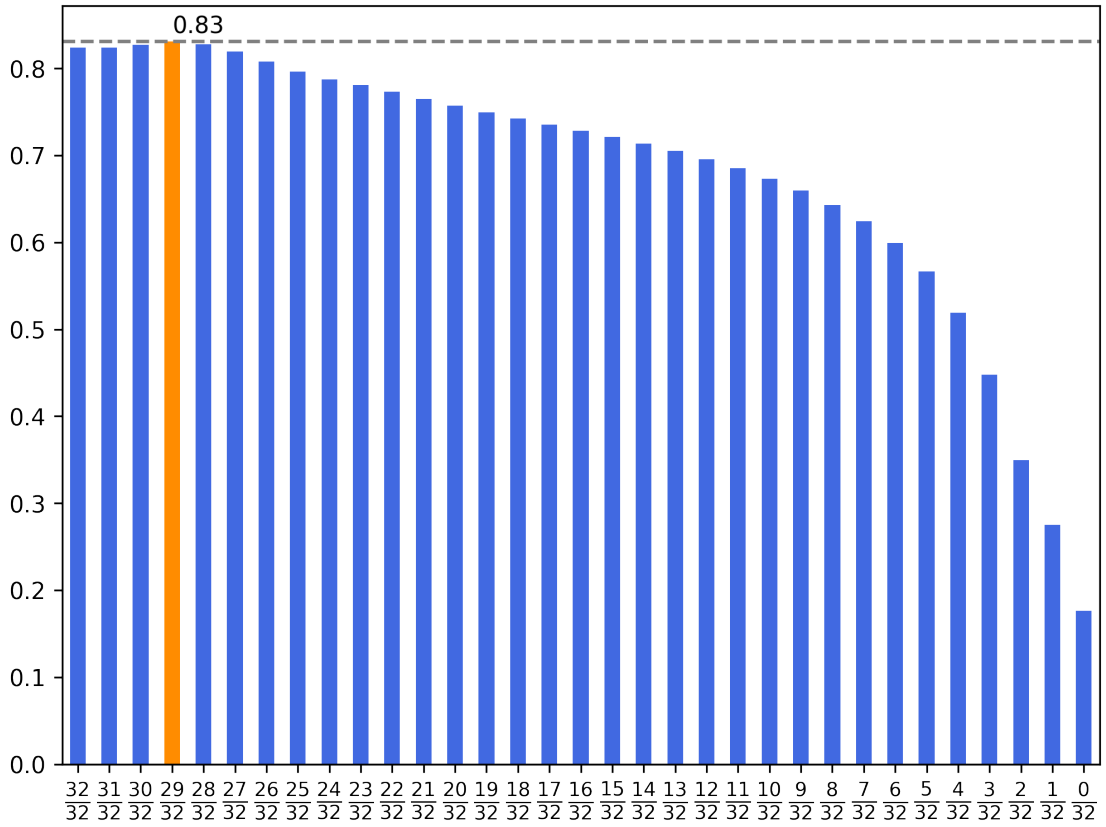
Predicted and actual positive and negatives for each threshold on the consensus score.



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

1.4.2 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.

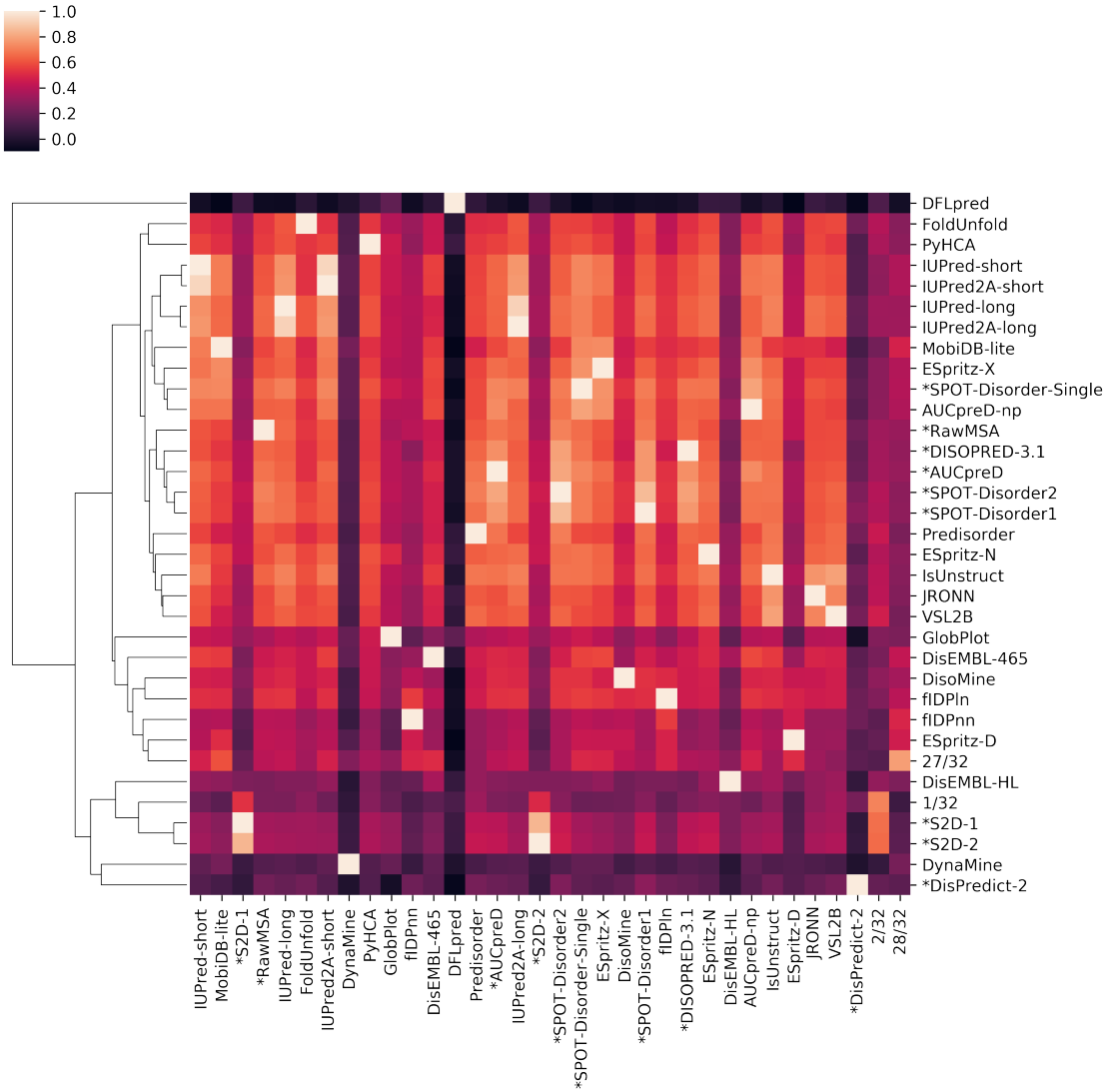
1.4.3 Percentage of correct/incorrect classifications

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

	DisProt		PDB		Undefined	
	TP	FN	TN	FP	TN	FP
IUPred-short	56.2	43.8	94.7	5.3	66.2	33.8
MobiDB-lite	41.1	58.9	98.6	1.4	77.2	22.8
*S2D-1	86.7	13.3	47.3	52.7	22.3	77.7
*RawMSA	68.2	31.8	93.1	6.9	60.6	39.4
IUPred-long	64.5	35.5	92.5	7.5	58.9	41.1
FoldUnfold	66.6	33.4	82.5	17.5	48.8	51.2
IUPred2A-short	56.3	43.7	94.9	5.1	66.3	33.7
DynaMine	1.7	98.3	100.0	0.0	97.9	2.1
PyHCA	65.8	34.2	84.8	15.2	53.7	46.3
GlobPlot	38.3	61.7	90.0	10.0	70.0	30.0
fIDPnn	30.0	70.0	98.7	1.3	93.0	7.0
DisEMBL-465	38.7	61.3	95.2	4.8	76.1	23.9
DFLpred	9.4	90.6	89.0	11.0	89.4	10.6
Predisorder	80.7	19.3	82.4	17.6	41.8	58.2
*AUCpreD	68.7	31.3	95.5	4.5	54.0	46.0
IUPred2A-long	63.4	36.6	92.8	7.2	59.6	40.4
*S2D-2	85.8	14.2	50.4	49.6	24.7	75.3
*SPOT-Disorder2	75.9	24.1	95.1	4.9	45.8	54.2
*SPOT-Disorder-Single	55.6	44.4	97.7	2.3	65.3	34.7
ESpritz-X	53.2	46.8	95.6	4.4	67.5	32.5
DisoMine	57.4	42.6	91.0	9.0	67.3	32.7
*SPOT-Disorder1	74.8	25.2	94.2	5.8	48.8	51.2
fIDPln	50.5	49.5	94.6	5.4	79.3	20.7
*DISOPRED-3.1	64.5	35.5	93.9	6.1	49.9	50.1
ESpritz-N	68.7	31.3	86.6	13.4	51.6	48.4
DisEMBL-HL	53.0	47.0	74.4	25.6	63.8	36.2
AUCpreD-np	57.3	42.7	96.5	3.5	65.1	34.9
IsUnstruct	74.8	25.2	86.0	14.0	49.3	50.7
ESpritz-D	35.2	64.8	95.6	4.4	88.4	11.6
JRONN	74.1	25.9	81.7	18.3	47.3	52.7
VSL2B	81.5	18.5	77.2	22.8	40.0	60.0
*DisPredict-2	41.6	58.4	81.6	18.4	72.5	27.5

1.4.4 clustermap of binary predictions correlation

Correlation of binary states between predictors.



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.

1.5 Fully disordered targets

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

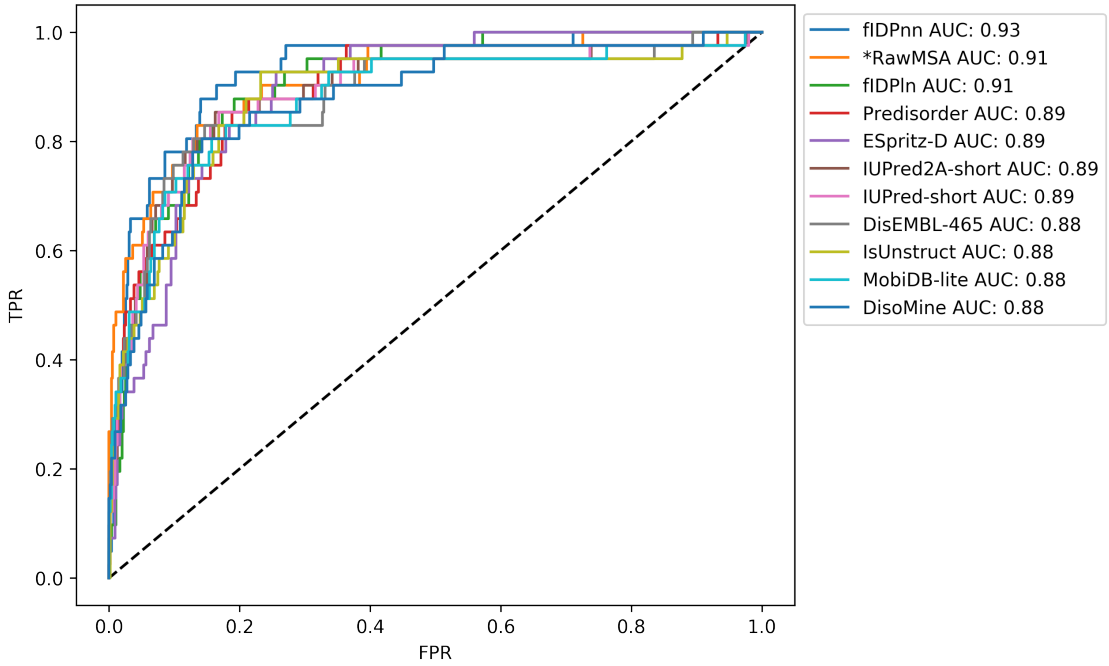
1.5.1 Correctly and incorrectly classified full IDPs

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

Actual	Positives		Negatives	
Predicted	TP	FN	FP	TN
GlobPlot	44	0	0	602
DisEMBL-HL	44	0	0	602
DisEMBL-465	44	0	0	602
random_chain	44	0	0	602
IUPred-short	44	0	1	601
random_chunk	44	0	1	601
IUPred2A-short	44	0	2	600
ESpritz-N	44	0	4	598
PyHCA	44	0	5	597
*DISOPRED-3.1	44	0	6	596
ESpritz-X	44	0	6	596
*SPOT-Disorder-Single	44	0	7	595
AUCpreD-np	44	0	11	591
Predisorder	44	0	13	589
IsUnstruct	44	0	14	588
*AUCpreD	44	0	15	587
VSL2B	44	0	24	578
*SPOT-Disorder2	44	0	28	574
*SPOT-Disorder1	44	0	31	571
*S2D-2	44	0	33	569
DisoMine	44	0	55	547
*S2D-1	44	0	57	545
cons	44	0	160	442
*DisPredict-2	43	1	18	584
*RawMSA	43	1	20	582
fIDPln	43	1	41	561
naive-pdb	43	1	138	464
IUPred2A-long	42	2	6	596
IUPred-long	42	2	6	596
JRONN	42	2	6	596
fIDPnn	42	2	16	586
ESpritz-D	42	2	52	550
naive-gene3d	42	2	159	443
FoldUnfold	41	3	156	446
MobiDB-lite	37	7	2	600
DFLpred	13	31	0	602
DynaMine	13	31	0	602

1.5.2 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.

	Acc disorder	Acc order	Bias (%)
*SPOT-Disorder2	0.720	0.769	38.378
*SPOT-Disorder1	0.707	0.749	51.063
*AUCpreD	0.696	0.730	44.459
Predisorder	0.695	0.723	54.112
*RawMSA	0.708	0.689	39.400
IsUnstruct	0.692	0.704	50.708
VSL2B	0.680	0.702	59.809
IUPred-long	0.683	0.690	41.128
*DISOPRED-3.1	0.658	0.734	50.077
IUPred2A-long	0.680	0.688	40.434