CAID

July 17, 2019

Contents

Chapter 1

CAID

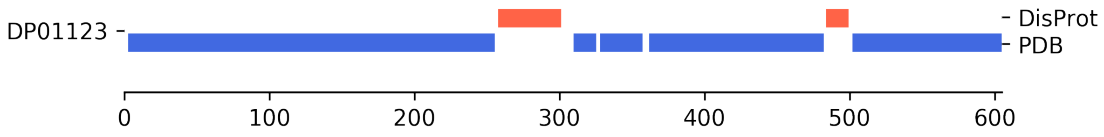
1.1 Dataset

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Disorder (CAID) is a continuous experiment where prediction methods for intrinsic disorder (ID) are blind tested on unpublished DisProt data.

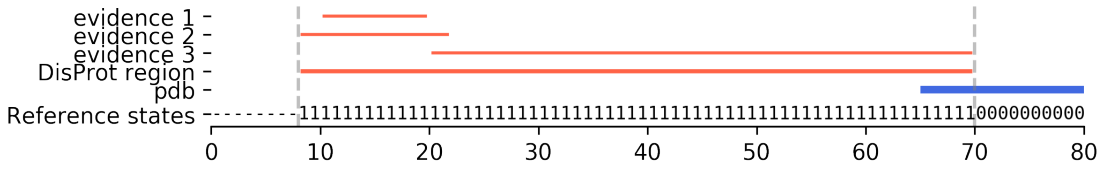
```
Accordion(children=(RadioButtons(description='Reference:', index=6, options=('new-disprot-linker', 'new
```

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

Current analysis is performed on the **new-disprot-all** dataset with **pdb** negative definition. This means that DisProt defines order for *all its new entries* and *pdb 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 undefined.



1.2 Evaluation metrics

Metrics evaluating prediction scores are calculated applying **ROC** 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
naive-pdb	6072	0	109201	48119	0.944	0.941	0.917	1.000	0.888	1.000	0.963	0.957	630	646
*SPOT-Disorder2	8102	9718	90397	39649	0.867	0.817	0.727	0.803	0.830	0.903	0.895	0.931	610	646
*SPOT-Disorder1	8514	12209	98272	45779	0.866	0.815	0.721	0.789	0.843	0.889	0.872	0.927	644	646
*AUCpreD	10325	12505	96350	44230	0.848	0.795	0.689	0.780	0.811	0.885	0.878	0.919	644	646
*SPOT-Disorder-Single	11250	13448	97105	43354	0.836	0.778	0.666	0.763	0.794	0.878	0.864	0.907	646	646
*RawMSA	11718	13702	96851	42886	0.831	0.771	0.656	0.758	0.785	0.876	0.841	0.902	646	646
AUCpreD-np	13015	13743	96810	41589	0.819	0.757	0.635	0.752	0.762	0.876	0.846	0.894	646	646
Predisorder	12354	14870	93952	42143	0.818	0.756	0.630	0.739	0.773	0.863	0.845	0.891	642	646
*DISOPRED-3.1	12924	14224	96329	41680	0.817	0.754	0.631	0.746	0.763	0.871	0.831	0.885	646	646
fIDPnn	10541	21544	89009	43977	0.806	0.733	0.588	0.671	0.807	0.805	0.806	0.883	645	646
IsUnstruct	12874	17178	93375	41730	0.804	0.735	0.598	0.708	0.764	0.845	0.815	0.879	646	646
VSL2B	13496	16237	94244	40797	0.802	0.733	0.597	0.715	0.751	0.853	0.816	0.874	644	646
IUPred-long	13989	15940	94541	40571	0.800	0.731	0.594	0.718	0.744	0.856	0.816	0.871	645	646
IUPred2A-short	14213	15846	94707	40391	0.798	0.729	0.592	0.718	0.740	0.857	0.797	0.868	646	646
IUPred2A-long	14491	15308	95245	40113	0.798	0.729	0.594	0.724	0.735	0.862	0.813	0.869	646	646
MobiDB-lite	13041	18458	92023	41519	0.797	0.725	0.581	0.692	0.761	0.833	0.805	0.852	645	646
IUPred-short	13330	17858	92623	41230	0.797	0.726	0.583	0.698	0.756	0.838	0.795	0.866	645	646
ESpritz-X	12969	18717	91764	41591	0.796	0.724	0.579	0.690	0.762	0.831	0.802	0.871	645	646
DisoMine	12237	21289	89264	42367	0.792	0.717	0.564	0.666	0.776	0.807	0.737	0.864	646	646
ESpritz-D	9833	25957	84524	44727	0.792	0.714	0.556	0.633	0.820	0.765	0.722	0.861	645	646
fIDPln	11366	23569	86984	43152	0.789	0.712	0.554	0.647	0.792	0.787	0.776	0.869	645	646
JRONN	15103	18073	92408	39457	0.780	0.704	0.552	0.686	0.723	0.836	0.789	0.852	645	646
ESpritz-N	16011	16662	93819	38549	0.778	0.702	0.554	0.698	0.707	0.849	0.787	0.847	645	646
DynaMine	14313	23650	86831	40247	0.762	0.680	0.506	0.630	0.738	0.786	0.759	0.837	645	646
PyHCA	16486	20504	90049	38118	0.756	0.673	0.504	0.650	0.698	0.815	0.737	0.828	646	646
*S2D-1	18284	19118	91363	36009	0.745	0.658	0.488	0.653	0.663	0.827	0.731	0.812	644	646
naive-gene3d	6767	42457	67477	47544	0.745	0.659	0.462	0.528	0.875	0.614	0.719	0.754	639	646
chunk	18988	18988	91564	35615	0.740	0.652	0.480	0.652	0.652	0.828	NaN	NaN	100	100
DisEMBL-465	19113	23210	87271	35180	0.719	0.624	0.430	0.603	0.648	0.790	0.689	0.789	644	646
*S2D-2	15958	30710	79771	38335	0.714	0.622	0.408	0.555	0.706	0.722	0.568	0.768	644	646
GlobPlot	24926	24636	85845	29634	0.660	0.545	0.321	0.546	0.543	0.777	0.594	0.714	645	646
DisEMBL-HL	13466	50302	60179	40827	0.648	0.561	0.280	0.448	0.752	0.545	0.560	0.707	644	646
*DisPredict-2	21938	36724	73829	32666	0.633	0.527	0.254	0.471	0.598	0.668	0.519	0.673	646	646
cons	21503	38714	71839	33101	0.628	0.524	0.243	0.461	0.606	0.650	0.477	0.666	646	646
DFLpred	52901	2288	108265	1703	0.505	0.058	0.032	0.427	0.031	0.979	0.290	0.379	646	646
chain	36555	36555	73997	18048	0.500	0.331	-0.000	0.331	0.331	0.669	NaN	NaN	100	100

Where table column names mean:

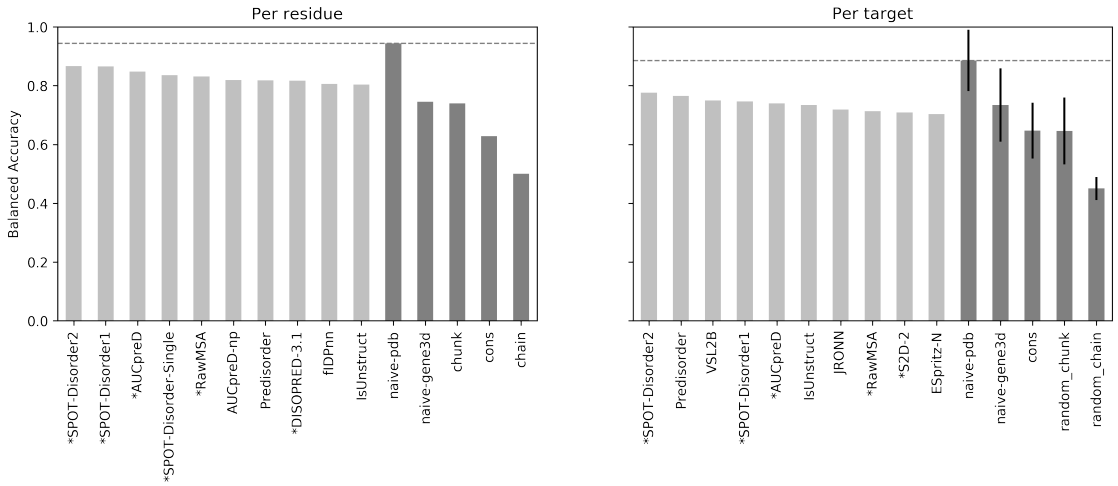
label	meaning
BAc	balanced accuracy
F1s	F1-score
MCC	Matthew's Correlation Coefficient
Pre	Precision/Selectivity
Rec	Recall/Sensitivity
Rec_n	Specificity
AUC_ROC	Area under the ROC curve
AUC_PRC	Area under the PR curve
npred	number of predicted targets
nref	number of targets in reference

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
```

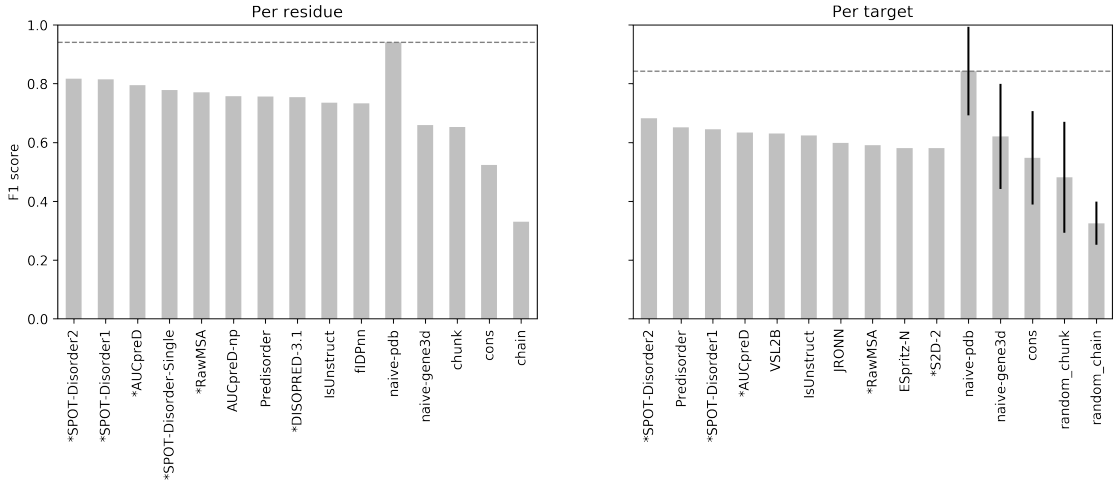


Overall (left panel) and average per-target (right panel) balanced accuracy. Light gray bars represent published prediction methods; dark gray bars represent baseline prediction methods. On the right panel standard deviation is shown as an error bar.

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
```

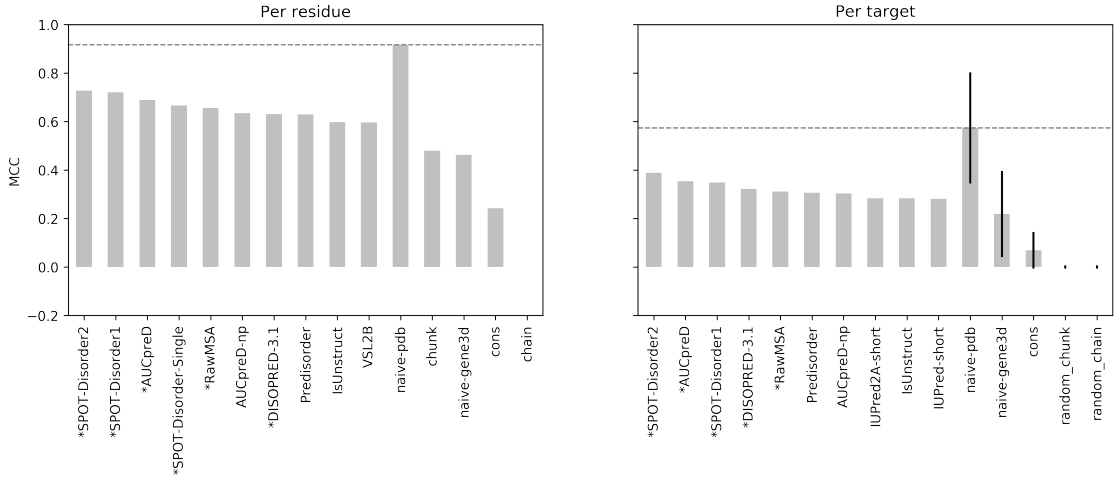


Overall (left panel) and average per-target (right panel) F1-score. Light gray bars represent published prediction methods; dark gray bars represent baseline prediction methods. On the right panel standard deviation is shown as an error bar.

1.2.3 MCC

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

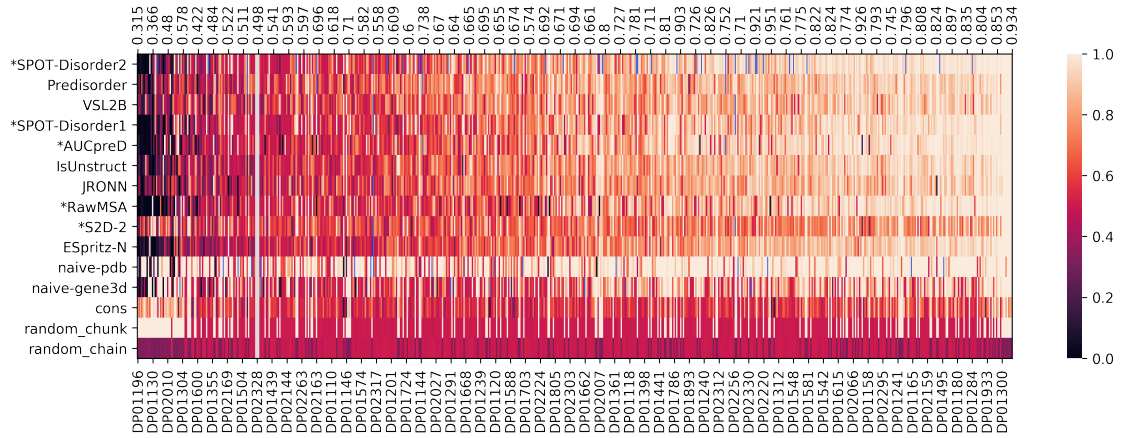
```
/home/marnec/.local/share/virtualenvs/caid-ICjYQIts/lib/python3.6/site-packages/ipykernel_launcher.py:2
Passing list-likes to .loc or [] with any missing label will raise
KeyError in the future, you can use .reindex() as an alternative.

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)

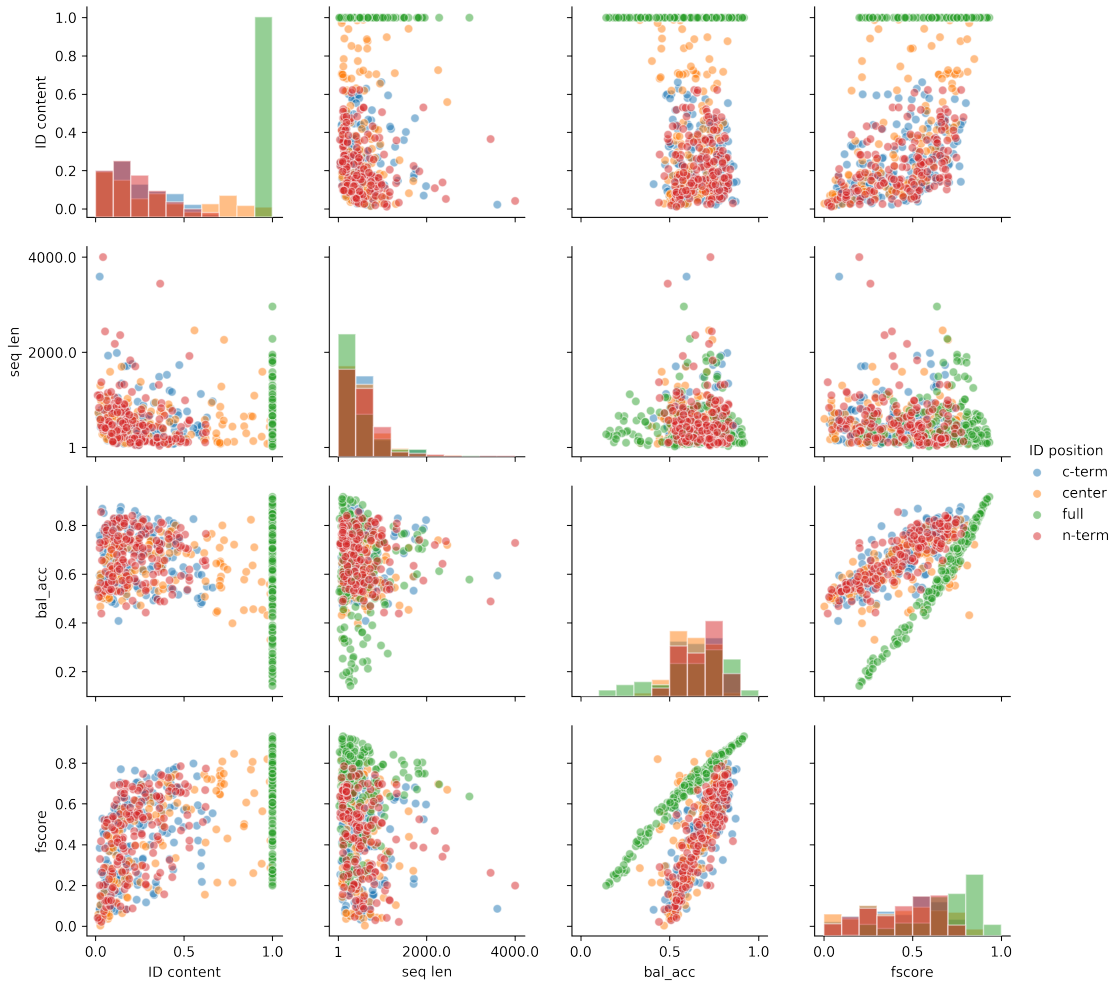


Heatmap of the predictors accuracy for each target. The higher the accuracy the lighter the color. Non-predicted targets are shown in blue. x and y axes are sorted by average accuracy over rows and columns respectively. A white semi-transparent vertical line marks the point where the average accuracy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

DP01196	DP01501	DP01456	DP01432	DP01971	DP02162
DP01500	DP01248	DP01949	DP01366	DP01437	DP01130
DP01339	DP01281	DP01436	DP01486	DP01168	DP02010
DP01181	DP01870	DP01498	DP01512	DP01278	DP01427
DP02296	DP01163	DP01600	DP01134	DP01898	DP01307
DP01285	DP01195	DP01584	DP01494	DP01355	DP02168
DP01612	DP01430	DP02234	DP01883	DP01505	DP01869
DP01967	DP01877	DP01762	DP02328	DP01187	DP02025
DP01141	DP01224				

1.2.5 Target correlation matrix

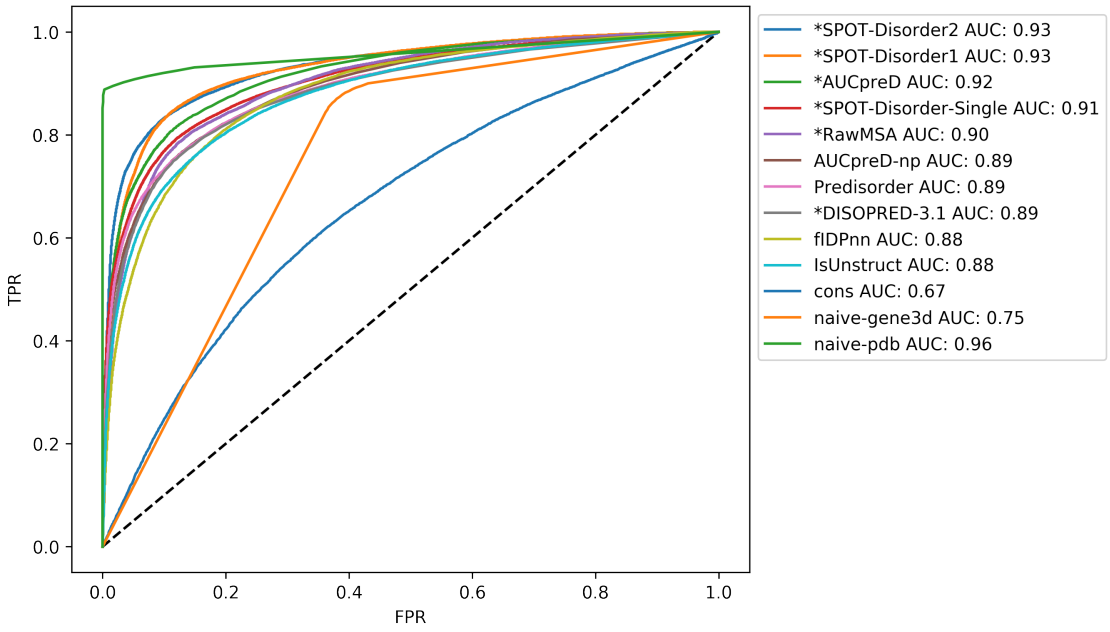
Commonly, experimental data has a bias for low disorder-content. DisProt targets have high disorder-content. Classifiers have been trained/engineered on low disorder-content. I expect difficult targets to have high disorder content. To verify is there is any correlation between target features I'm plotting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

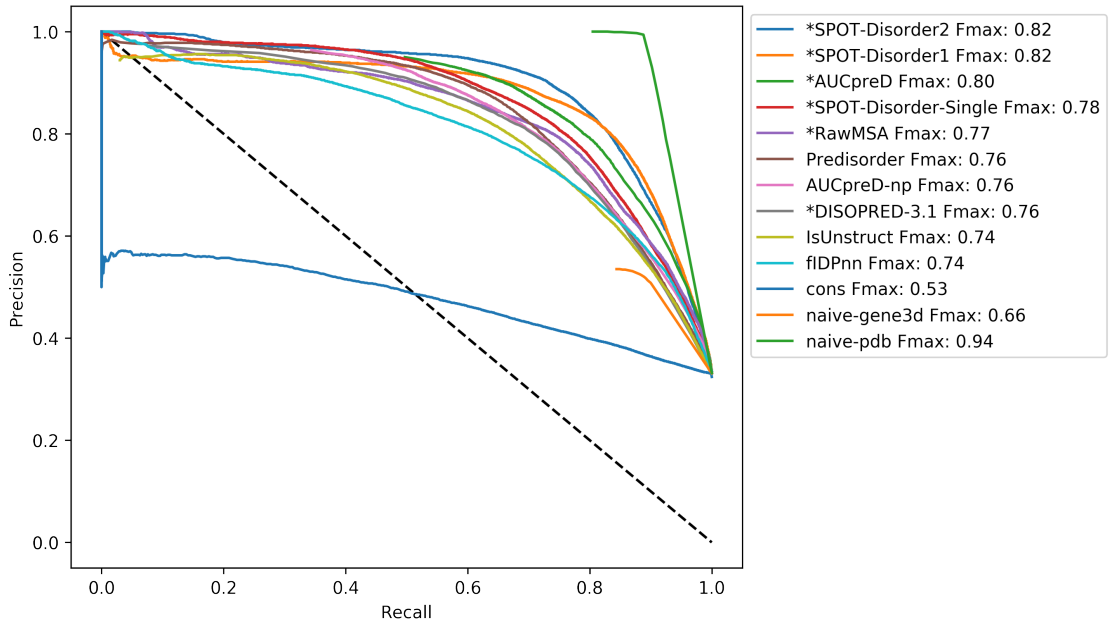


False Positive Rate (FPR) on x axis, True Positive Rate (TPR) on y axis. Methods are sorted by their Area Under the Curve (AUC). Only first ten methods plus baselines are shown.

1.2.7 PR curve

Precision Recall curve plot for predictors and baselines

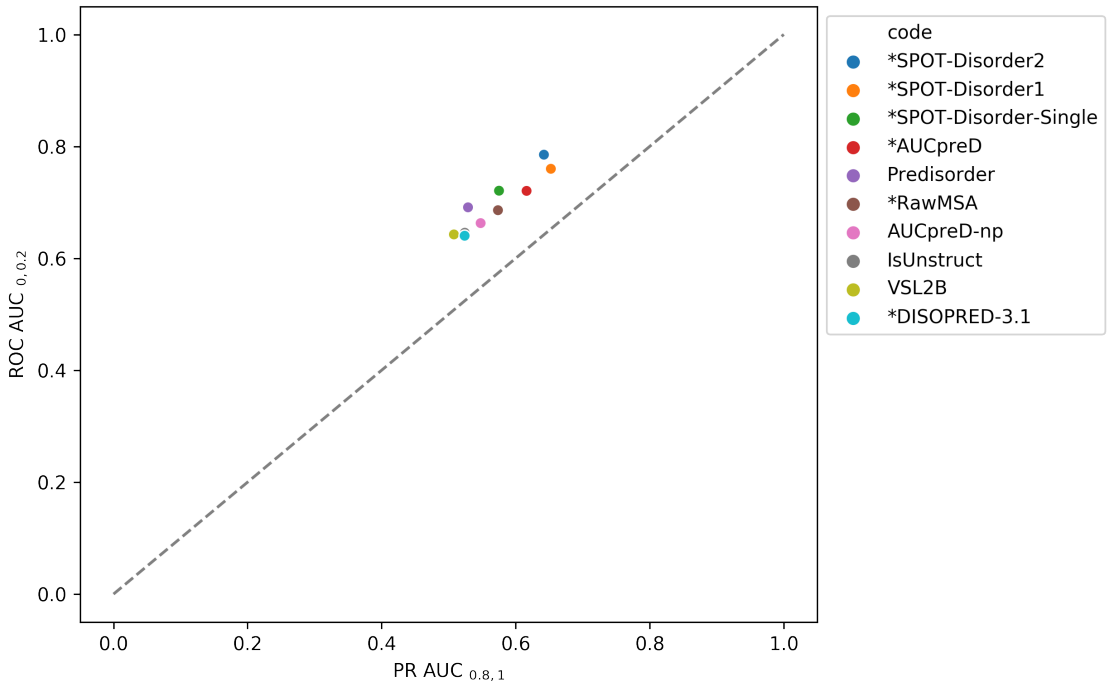
```
/home/marnec/.local/share/virtualenvs/caid-ICjYQIts/lib/python3.6/site-packages/ipykernel_launcher.py:7
  import sys
```



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

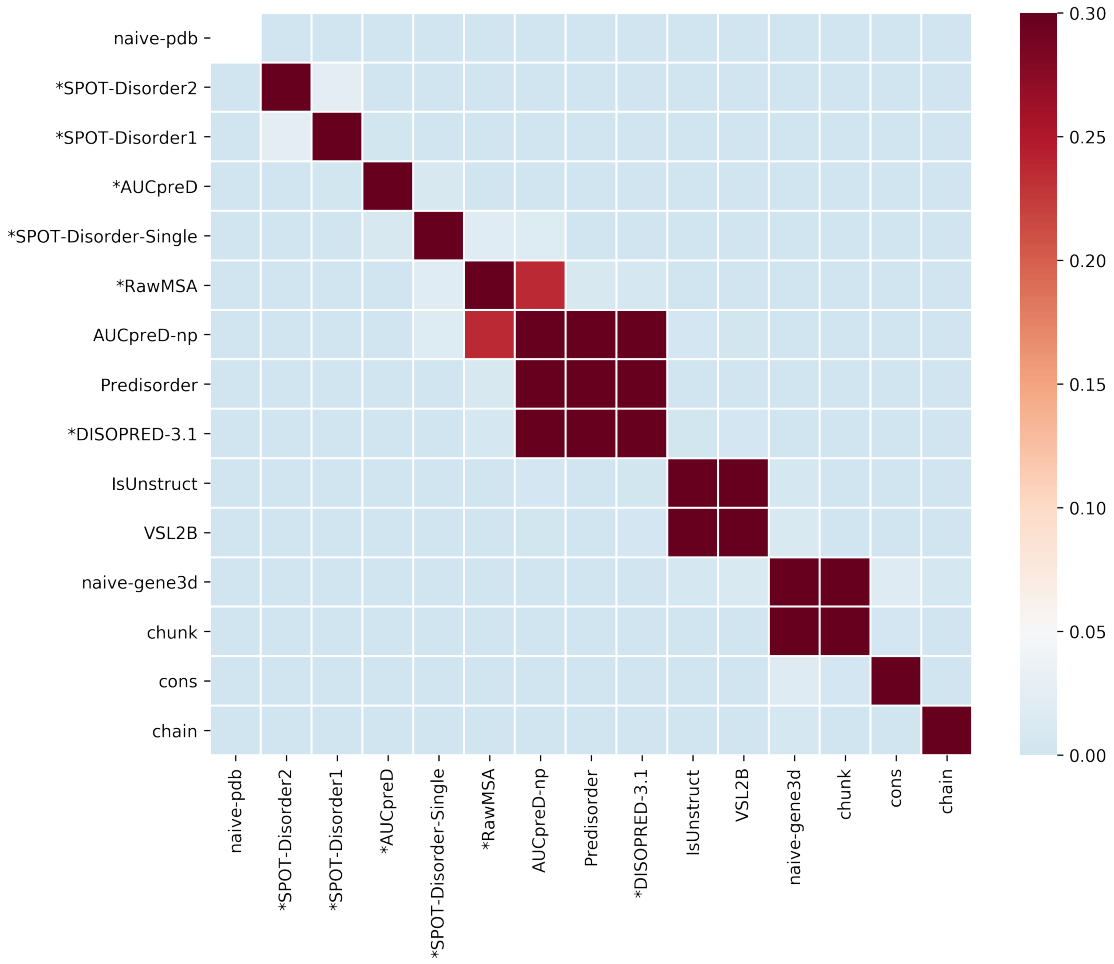
Plot of the AUCs from the ROC curve and PR curve



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

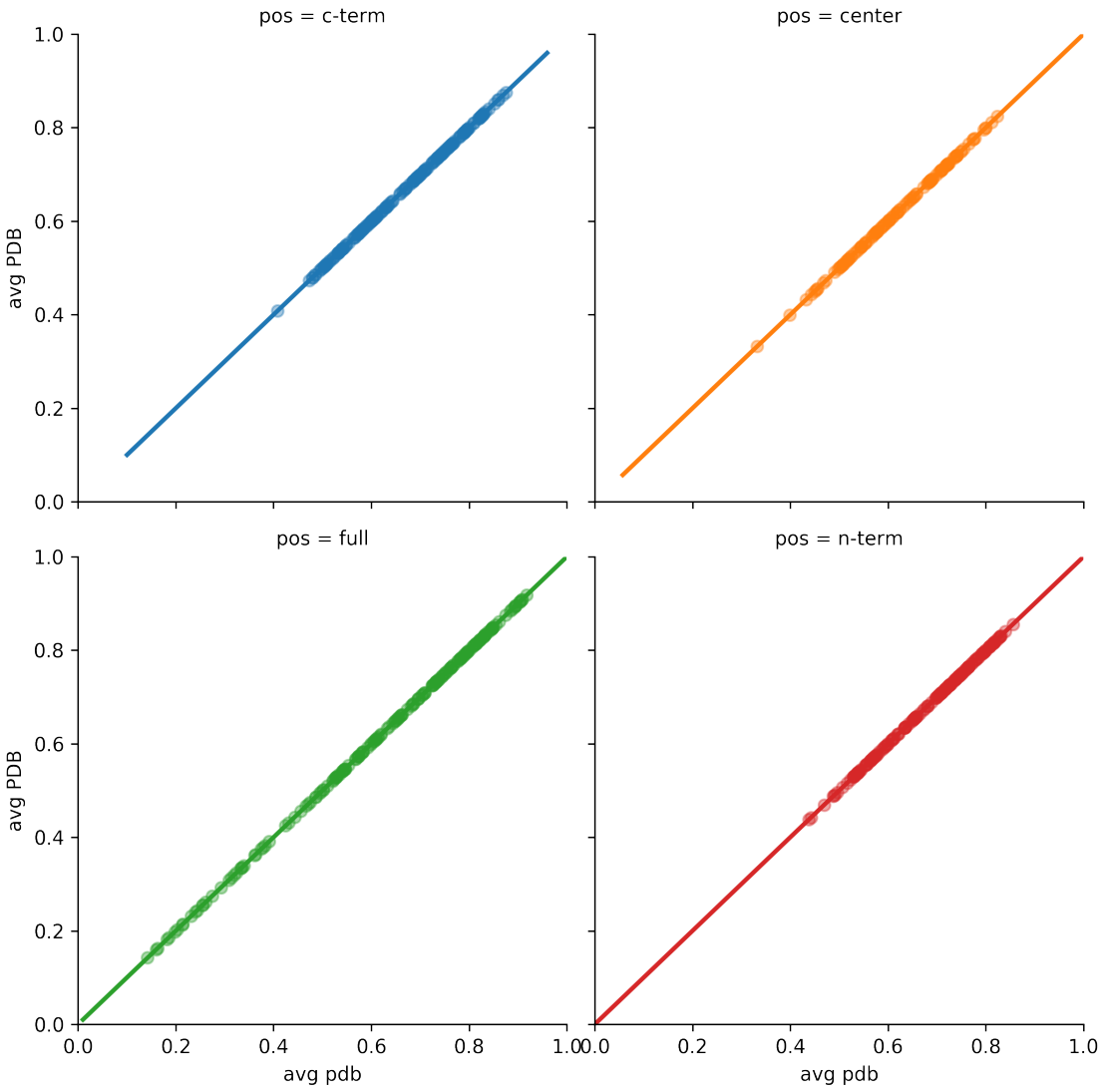
	Acc disorder	Acc order	Bias (%)
*SPOT-Disorder2	0.720	0.769	38.378
*SPOT-Disorder1	0.707	0.749	51.063
*AUCpreD	0.696	0.730	44.459
Predisorder	0.695	0.723	54.112
*RawMSA	0.708	0.689	39.400
IsUnstruct	0.692	0.704	50.708
VSL2B	0.680	0.702	59.809
IUPred-long	0.683	0.690	41.128
*DISOPRED-3.1	0.658	0.734	50.077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



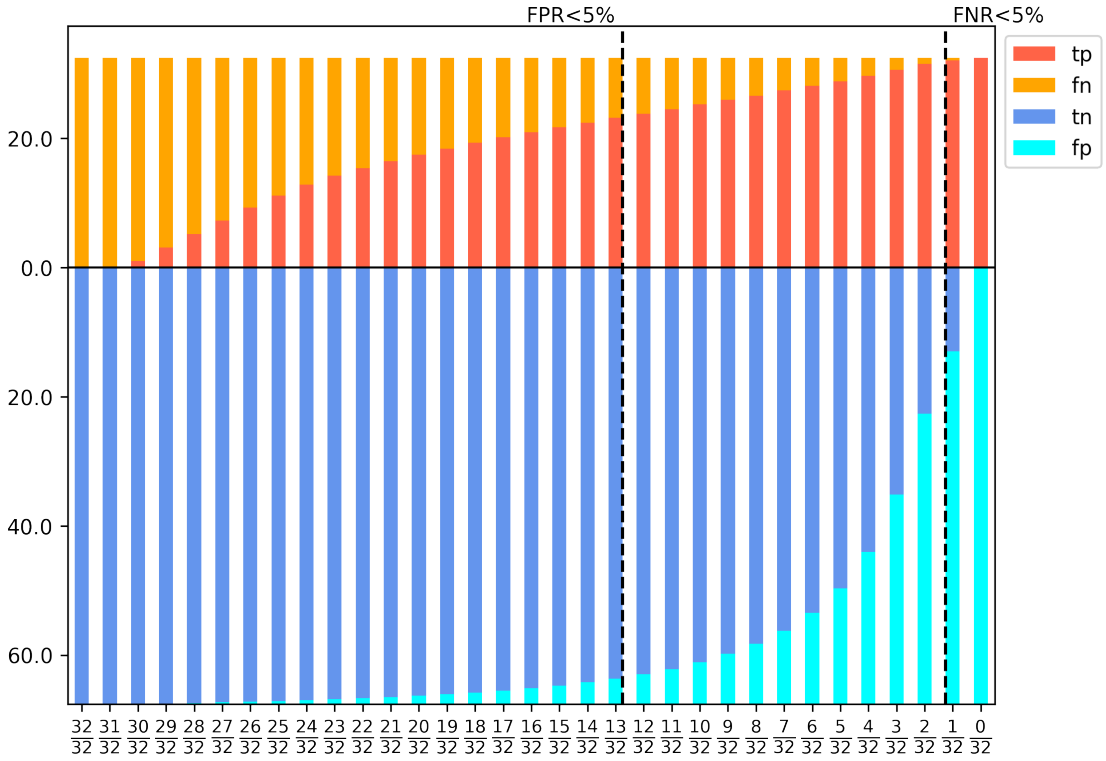
Average balanced accuracy for targets with reference negative defined by the *simple* rule on x axis. Average balanced accuracy for targets with reference negative defined by the *pdb* rule on y axis. Each panel includes only targets with a specific average disorder position (C-terminal, N-Terminal,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

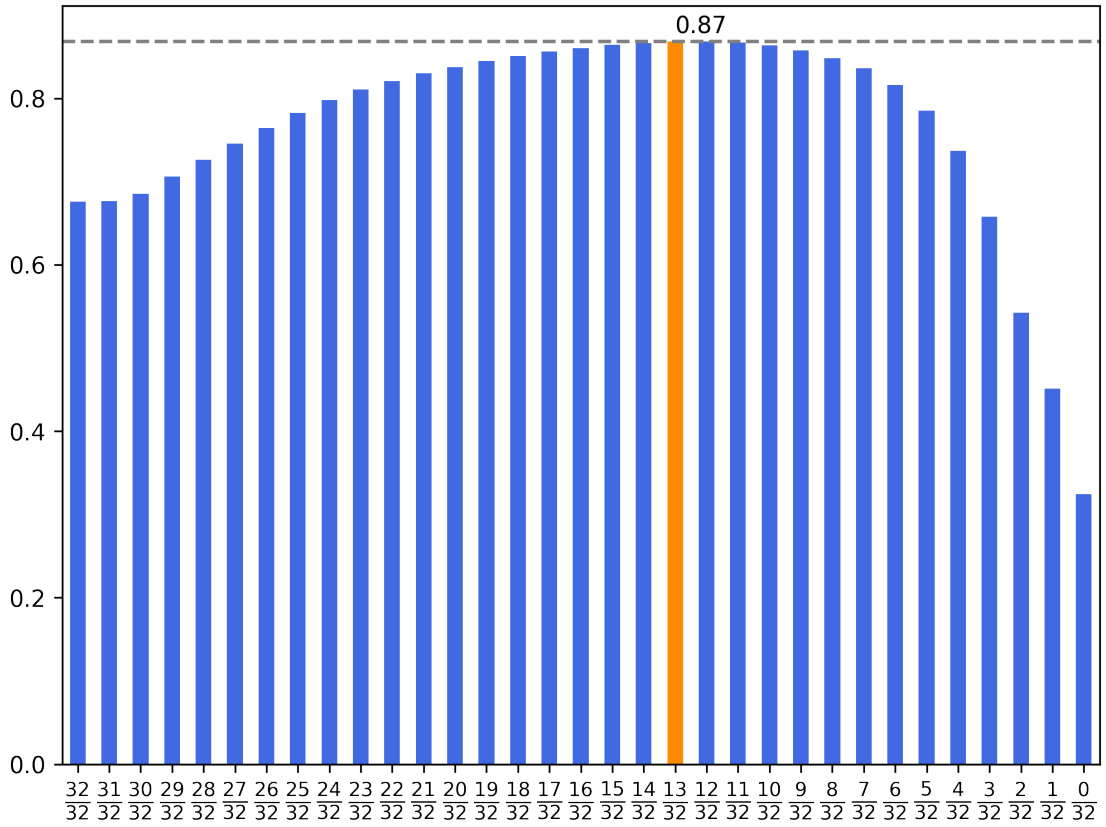
Predicted and actual positive and negatives for each threshold on the consensus score.



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.

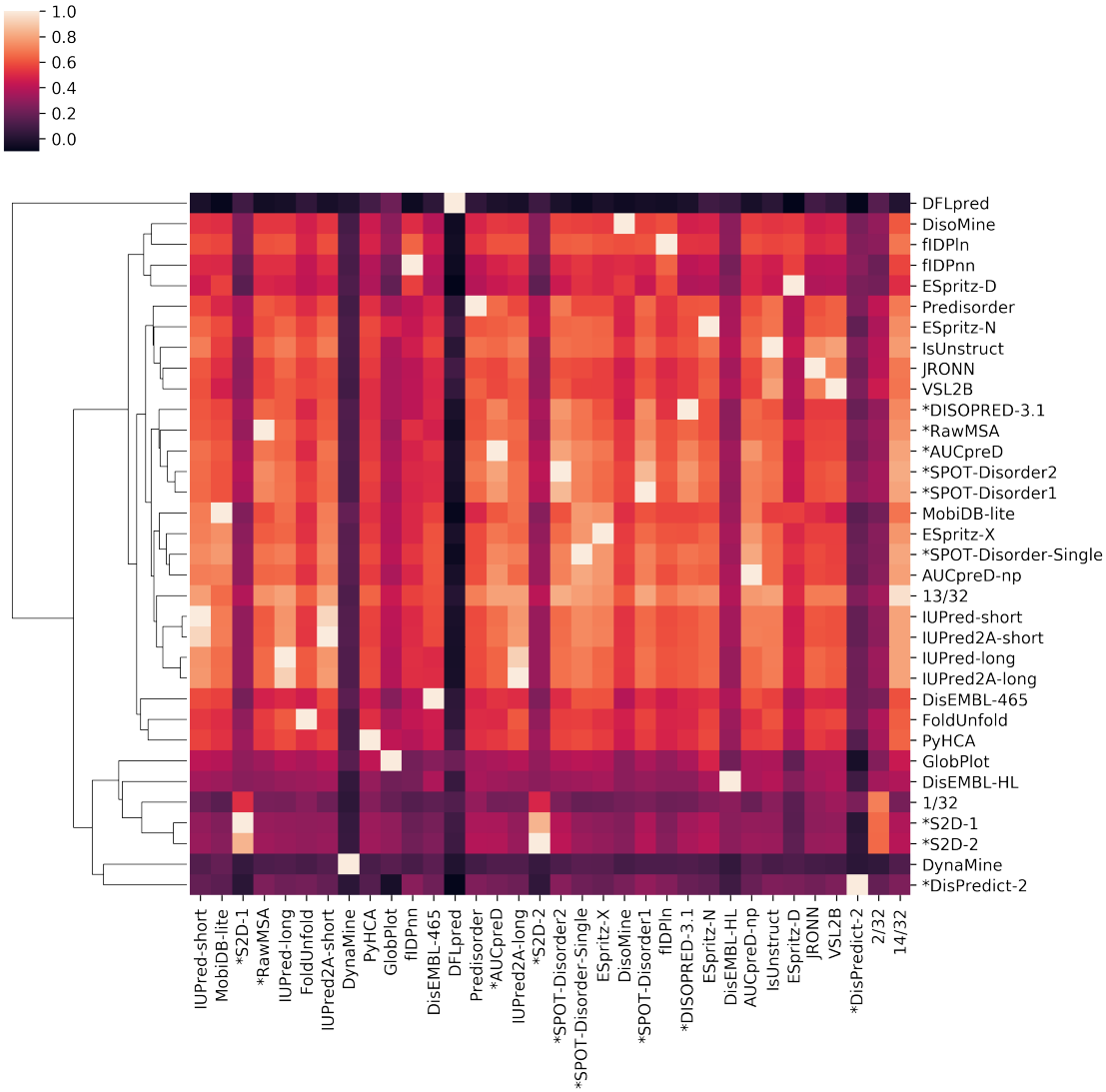
1.4.3 Percentage of correct/incorrect classifications

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.

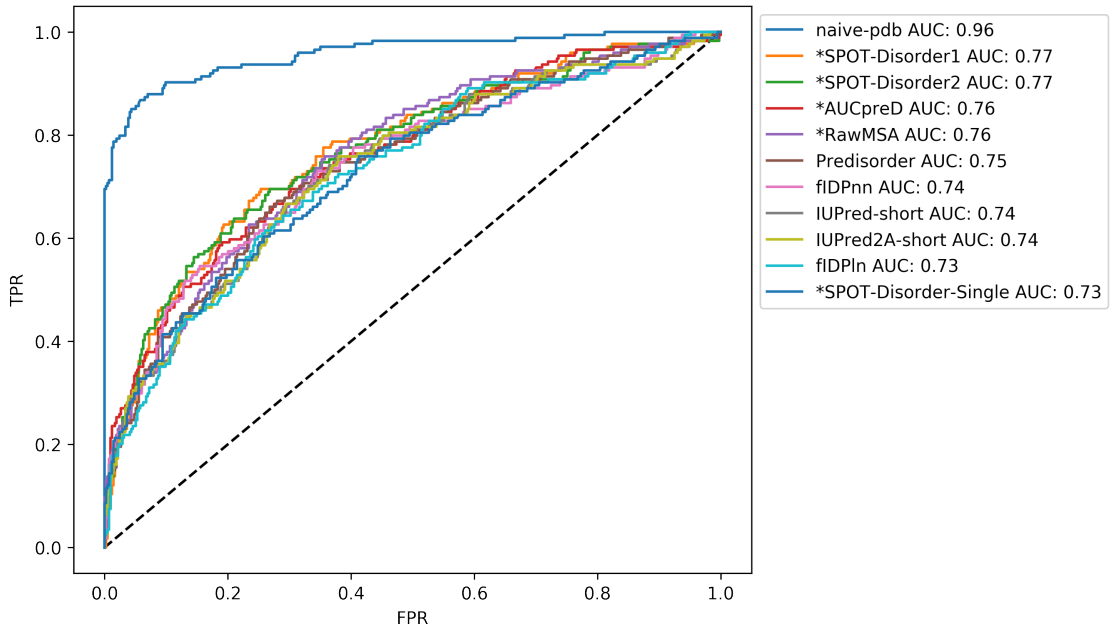
1.5.1 Correctly and incorrectly classified full IDPs

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

Actual	Positives		Negatives	
Predicted	TP	FN	FP	TN
random_chain	189	0	0	457
random_chunk	189	0	4	453
cons	189	0	116	341
*S2D-2	188	1	20	437
*S2D-1	188	1	32	425
VSL2B	187	2	25	432
Predisorder	186	3	9	448
PyHCA	185	4	5	452
IsUnstruct	185	4	13	444
ESpritz-N	183	6	7	450
JRONN	182	7	8	449
naive-pdb	181	8	0	457
*AUCpreD	181	8	12	445
naive-gene3d	181	8	124	333
GlobPlot	179	10	0	457
IUPred-short	179	10	3	454
IUPred2A-short	179	10	4	453
*SPOT-Disorder1	179	10	24	433
DisEMBL-HL	178	11	2	455
FoldUnfold	176	13	97	360
DisEMBL-465	175	14	0	457
*DISOPRED-3.1	174	15	10	447
AUCpreD-np	173	16	7	450
IUPred2A-long	173	16	8	449
ESpritz-X	172	17	5	452
IUPred-long	172	17	11	446
*SPOT-Disorder2	168	21	19	438
*SPOT-Disorder-Single	167	22	7	450
DisoMine	165	24	43	414
*RawMSA	164	25	23	434
*DisPredict-2	160	29	16	441
flDPln	159	30	33	424
flDPnn	138	51	10	447
MobiDB-lite	137	52	3	454
ESpritz-D	114	75	43	414
DFLpred	81	108	0	457
DynaMine	49	140	0	457

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