Disorder in CAID-2

Damiano Piovesan

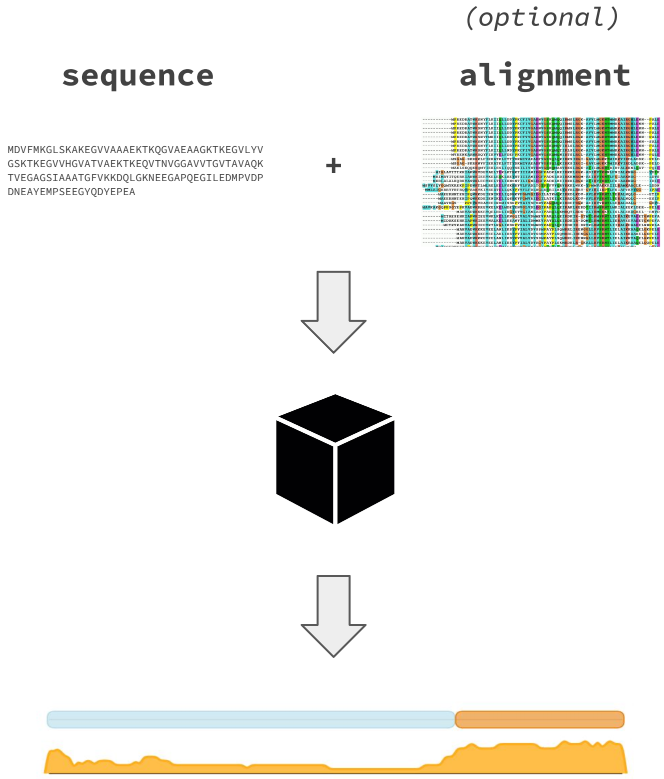
University of Padova - Italy

CASP15, 10-13 December 2022, Antalya

The challenge

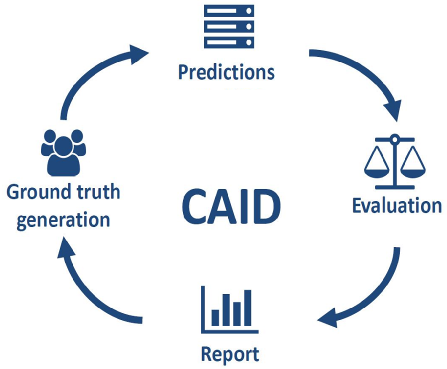
Prediction categories

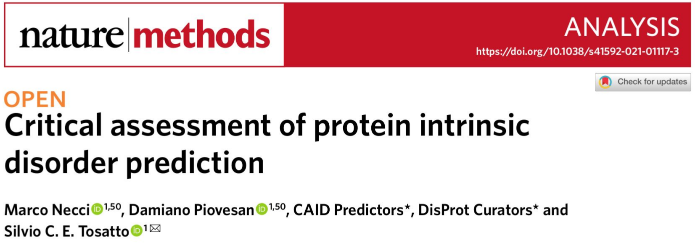
- **Disorder** - Disordered regions
- **Binding** - Binding residues inside disordered regions
- **Nucleic Acid binding** - Residues inside disordered regions that bind DNA/RNA molecules
- **Linker** - Entropic chains



```
MDVFMKGLSKAKEGVVAAAEKTKQGVAEAAGKTKEGVLYV
GSKTKEGVVHGVATVAEKTKEQVTNVGGAVVTGVTAVAQK
TVEGAGSIAAATGFVKKDQLGKNEEGAPQEGILEDMPVDP
DNEAYEMPSEEGYQDYEPEA
```

The CAID cycle




nature methods

ANALYSIS

https://doi.org/10.1038/s41592-021-01117-3

Check for updates

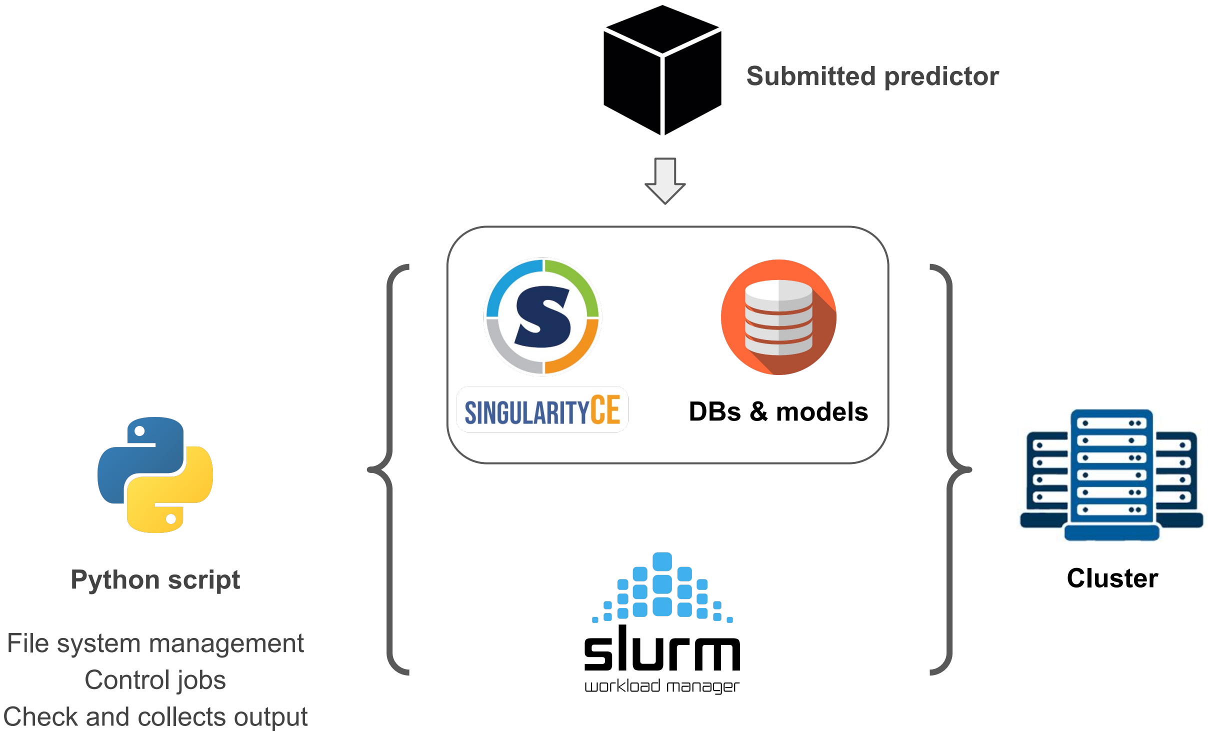
OPEN

Critical assessment of protein intrinsic disorder prediction

Marco Necci[1,50], Damiano Piovesan[1,50], CAID Predictors*, DisProt Curators* and Silvio C. E. Tosatto[1]✉

- **Ground truth generation**
  - Literature curation (DisProt)
- **Prediction**
  - Execution of stand-alone software (containers)
- **Assessment**
  - Accuracy & Technical evaluation
- **Report**
  - CAID & CASP conferences

Execution pipeline



Submitted predictor

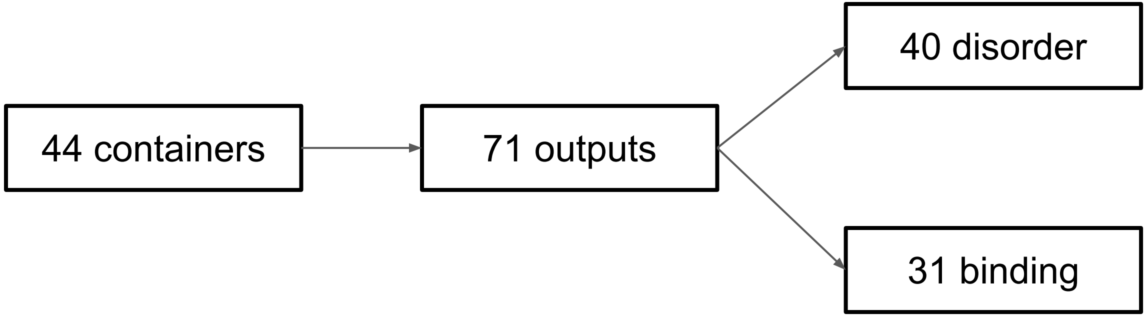
DBs & models

Python script

File system management
Control jobs
Check and collects output

Cluster

Methods



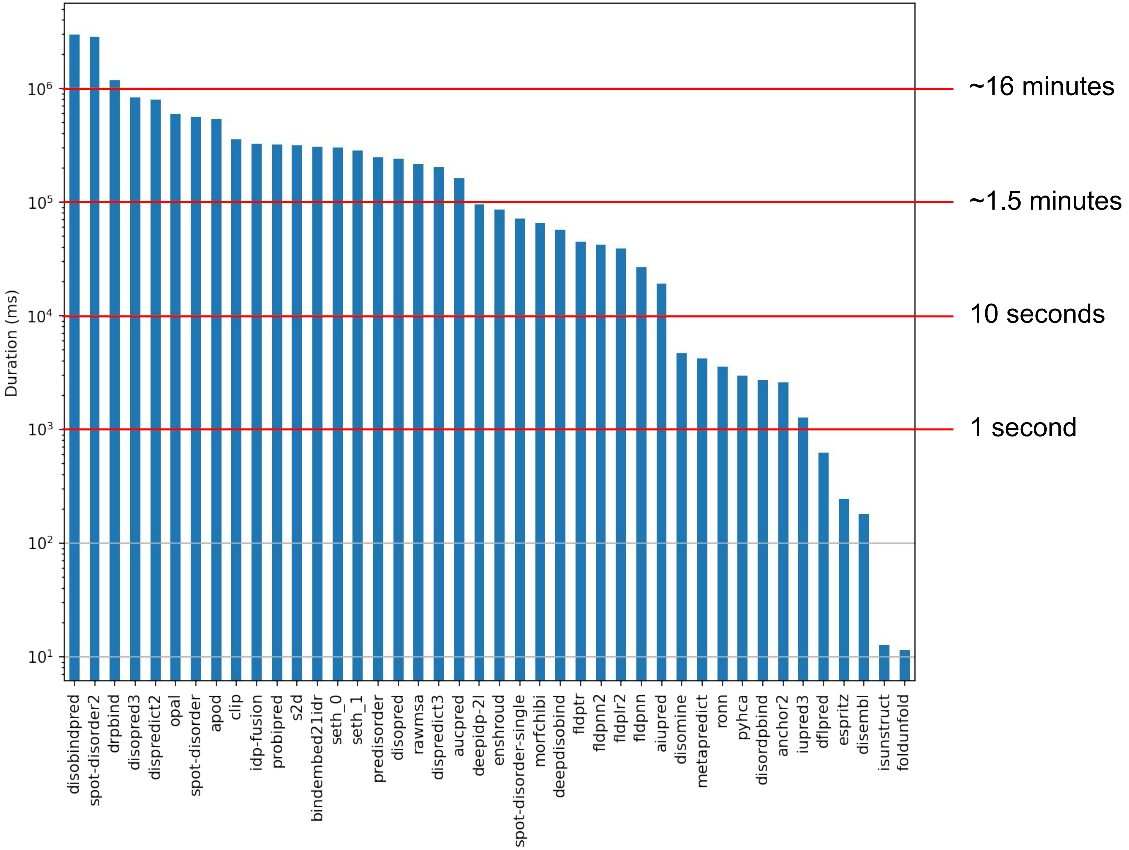
Container size (Gb)	# containers
< 0.5	14
> 0.5	9
> 1	13
> 2	5
> 3	3

- **22** packages are **new**, i.e. not tested in CAID 1
- **18** packages also available as **web servers**

Average execution time per protein (containers)

PRELIMINARY

- Additional execution time is included
  - BLAST
  - …
- Image instantiation time not included
- Some containers include both fast and slow predictors
  - MorfChibi / MorfChibi light
  - SPOT-Disorder / SPOT-Disorder-single
  - …



CAID-2

Ground truth

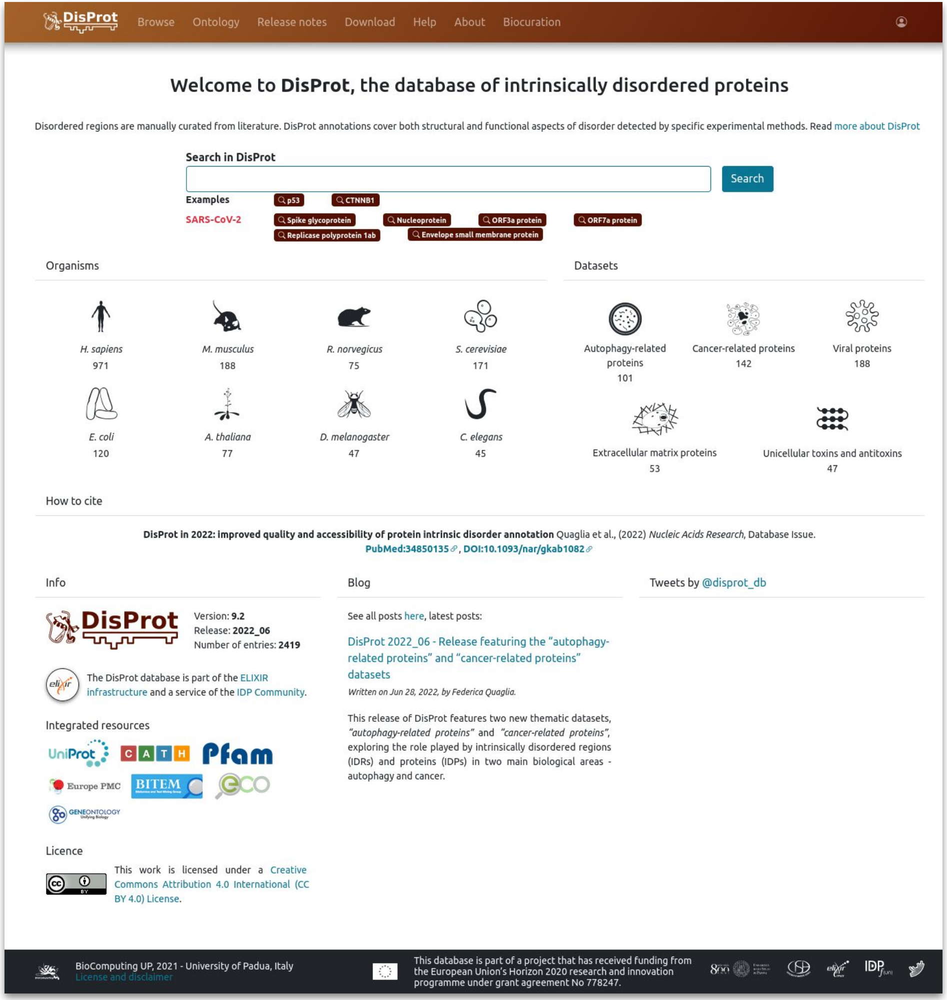
DisProt

Manually curated annotations of from the literature

Community

- 40 curators
- 30 laboratories
- 20 countries
- 2 reviewers

disprot.org



Welcome to **DisProt**, the database of intrinsically disordered proteins

Disordered regions are manually curated from literature. DisProt annotations cover both structural and functional aspects of disorder detected by specific experimental methods. Read more about DisProt

Search in DisProt

Examples

SARS-CoV-2

Organisms

H. sapiens 971

M. musculus 188

R. norvegicus 75

S. cerevisiae 171

E. coli 120

A. thaliana 77

D. melanogaster 47

C. elegans 45

Datasets

Autophagy-related proteins 101

Cancer-related proteins 142

Viral proteins 188

Extracellular matrix proteins 53

Unicellular toxins and antitoxins 47

How to cite

DisProt in 2022: improved quality and accessibility of protein intrinsic disorder annotation Quaglia et al., (2022) *Nucleic Acids Research*, Database Issue.
PubMed:34850135, **DOI:10.1093/nar/gkab1082**

Info

Version: **9.2**
Release: **2022_06**
Number of entries: **2419**

The DisProt database is part of the ELIXIR infrastructure and a service of the IDP Community.

Integrated resources

Licence

This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) License.

Blog

See all posts here, latest posts:

DisProt 2022_06 - Release featuring the “autophagy-related proteins” and “cancer-related proteins” datasets

Written on Jun 28, 2022, by Federica Quaglia.

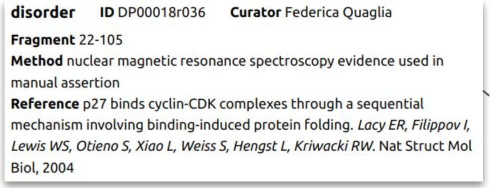
This release of DisProt features two new thematic datasets, *“autophagy-related proteins”* and *“cancer-related proteins”*, exploring the role played by intrinsically disordered regions (IDRs) and proteins (IDPs) in two main biological areas - autophagy and cancer.

Tweets by @disprot_db

BioComputing UP, 2021 - University of Padua, Italy
License and disclaimer

This database is part of a project that has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 778247.

Disorder annotations in DisProt

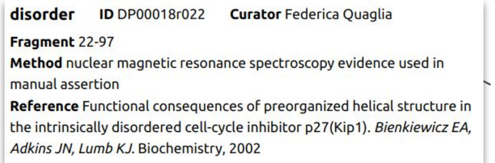


disorder **ID** DP00018r036 **Curator** Federica Quaglia

Fragment 22-105

Method nuclear magnetic resonance spectroscopy evidence used in manual assertion

Reference p27 binds cyclin-CDK complexes through a sequential mechanism involving binding-induced protein folding. *Lacy ER, Filippov I, Lewis WS, Otieno S, Xiao L, Weiss S, Hengst L, Kriwacki RW.* Nat Struct Mol Biol, 2004

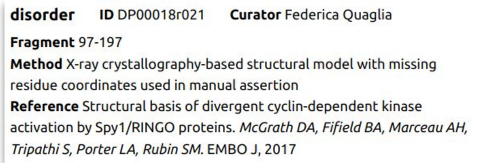


disorder **ID** DP00018r022 **Curator** Federica Quaglia

Fragment 22-97

Method nuclear magnetic resonance spectroscopy evidence used in manual assertion

Reference Functional consequences of preorganized helical structure in the intrinsically disordered cell-cycle inhibitor p27(Kip1). *Bienkiewicz EA, Adkins JN, Lumb KJ.* Biochemistry, 2002

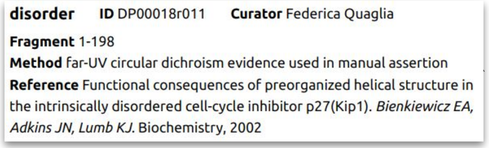


disorder **ID** DP00018r021 **Curator** Federica Quaglia

Fragment 97-197

Method X-ray crystallography-based structural model with missing residue coordinates used in manual assertion

Reference Structural basis of divergent cyclin-dependent kinase activation by Spy1/RINGO proteins. *McGrath DA, Fifield BA, Marceau AH, Tripathi S, Porter LA, Rubin SM.* EMBO J, 2017

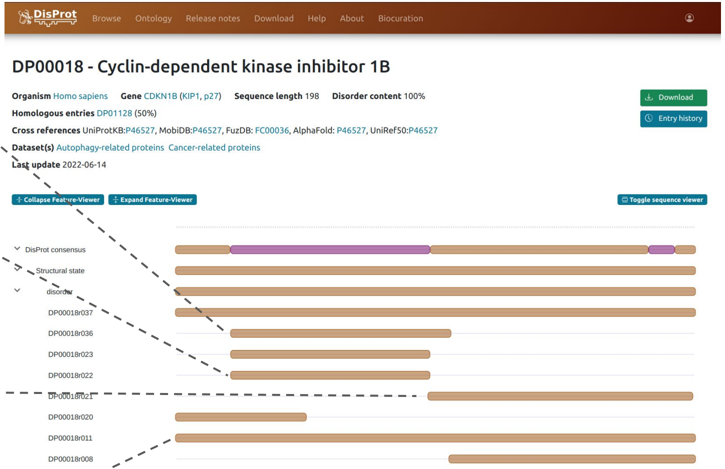


disorder **ID** DP00018r011 **Curator** Federica Quaglia

Fragment 1-198

Method far-UV circular dichroism evidence used in manual assertion

Reference Functional consequences of preorganized helical structure in the intrinsically disordered cell-cycle inhibitor p27(Kip1). *Bienkiewicz EA, Adkins JN, Lumb KJ.* Biochemistry, 2002



DP00018 - Cyclin-dependent kinase inhibitor 1B

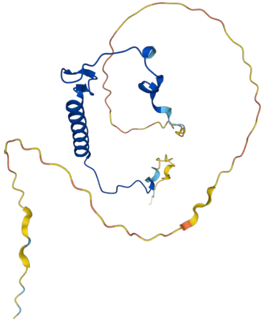
Organism Homo sapiens **Gene** CDKN1B (KIP1, p27) **Sequence length** 198 **Disorder content** 100%

Homologous entries DP01128 (50%)

Cross references UniProtKB:P46527, MobiDB:P46527, FuzDB: FC00036, AlphaFold: P46527, UniRef50:P46527

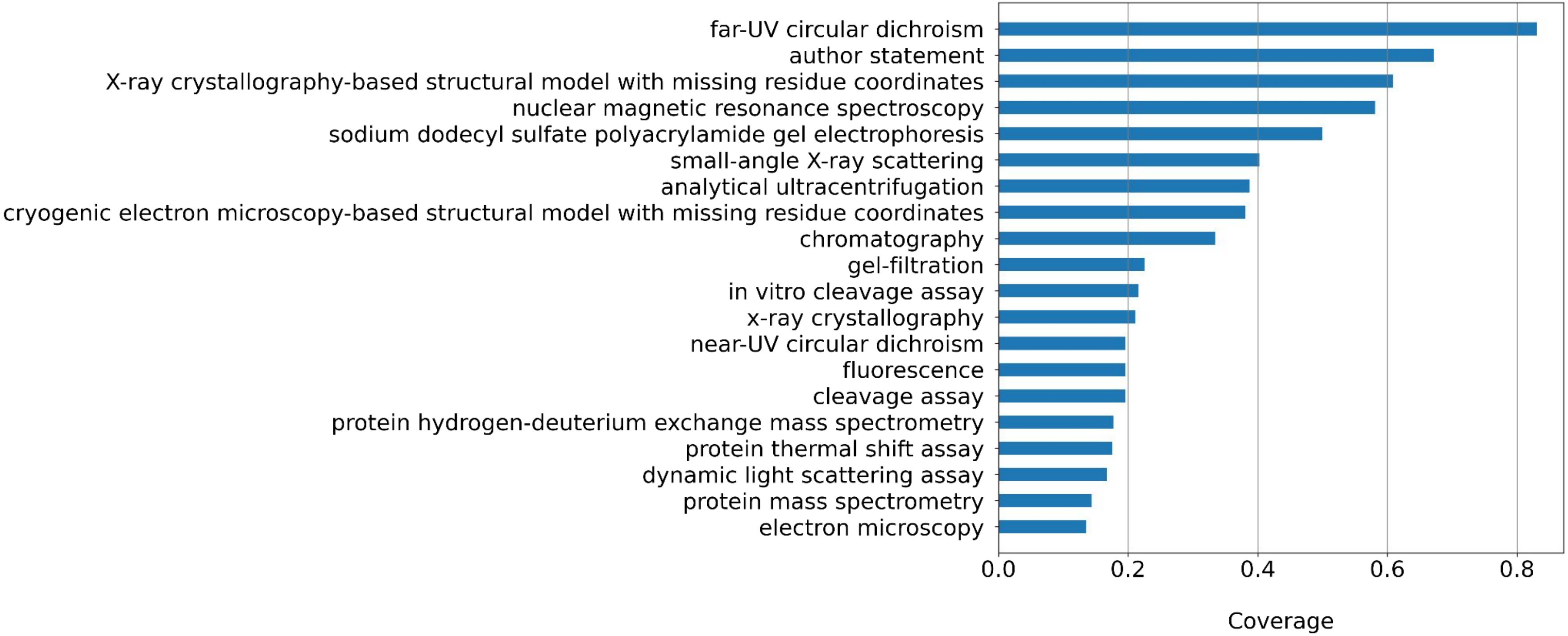
Dataset(s) Autophagy-related proteins Cancer-related proteins

Last update 2022-06-14

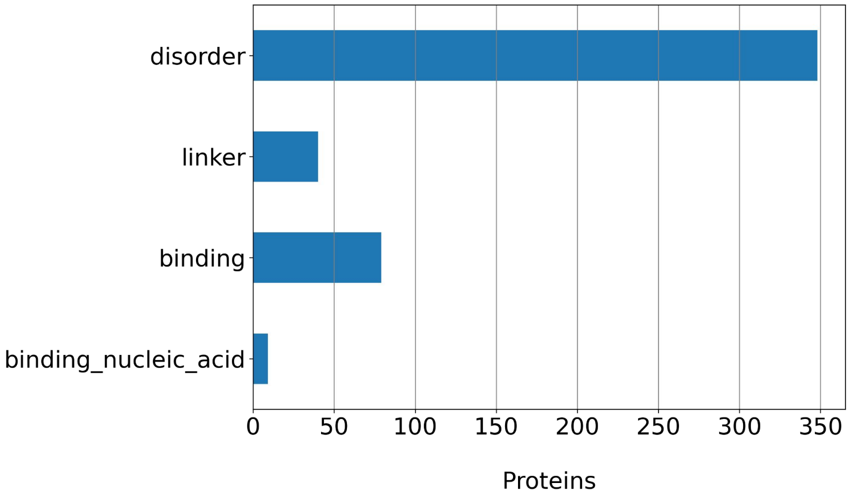


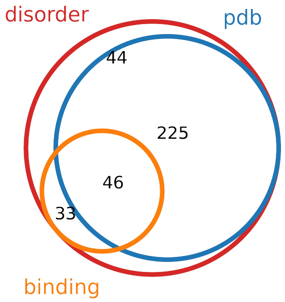
P46527

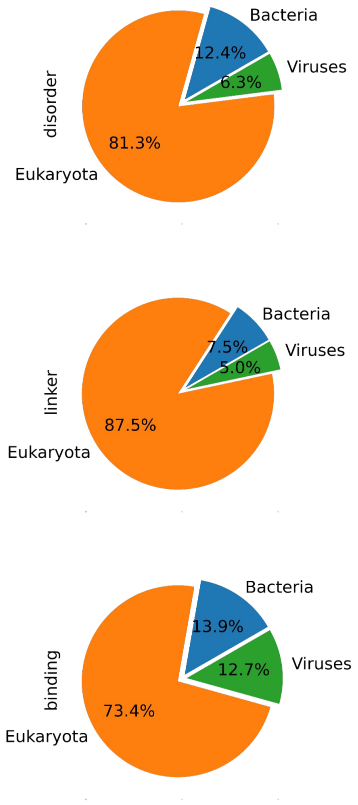
Experimental methods in DisProt



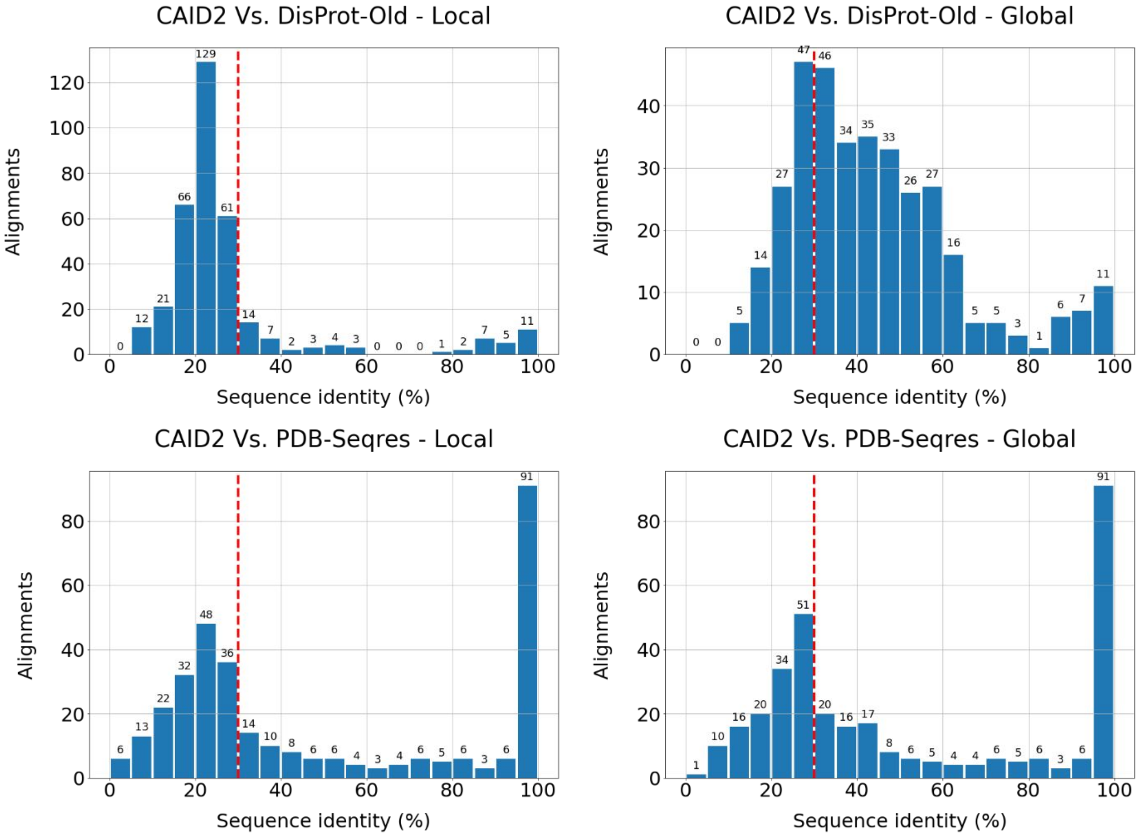
CAID-2 benchmark proteins



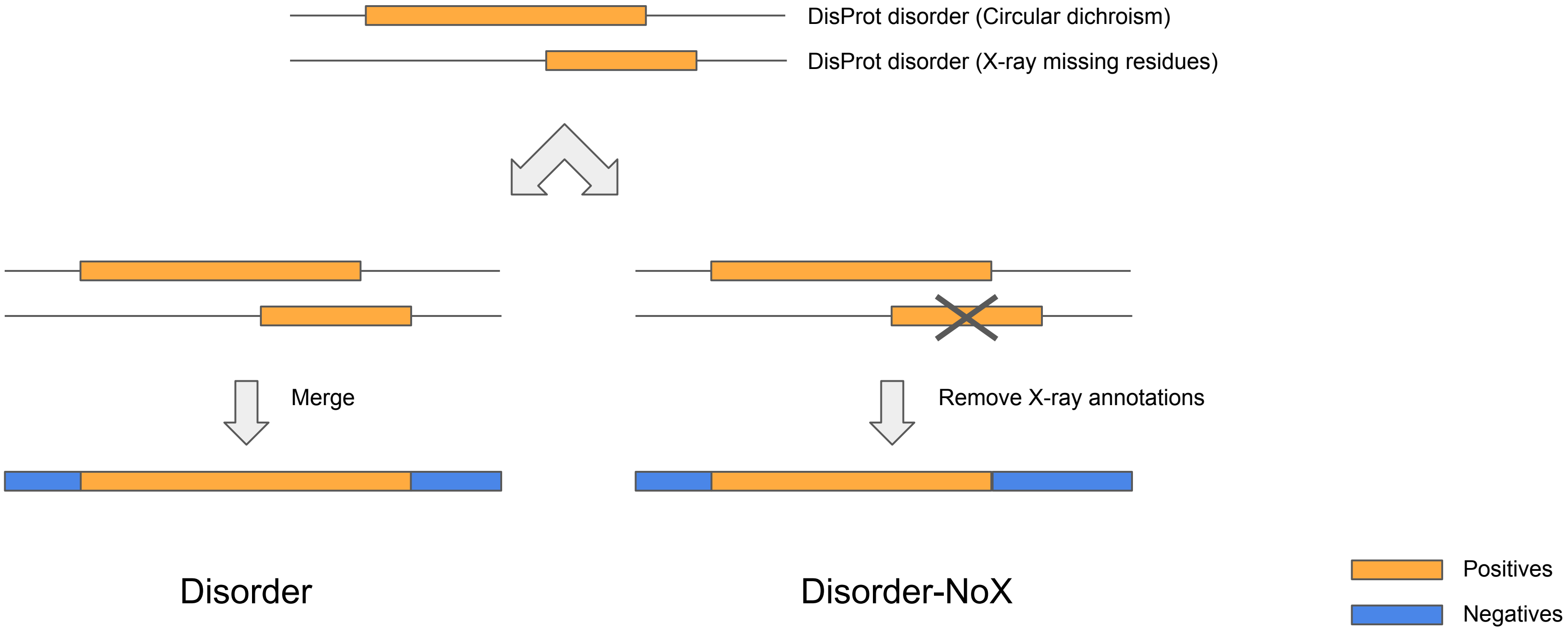




CAID-2 benchmark sequence identity



Reference definition - Disorder



DisProt disorder (Circular dichroism)

DisProt disorder (X-ray missing residues)

Merge

Remove X-ray annotations

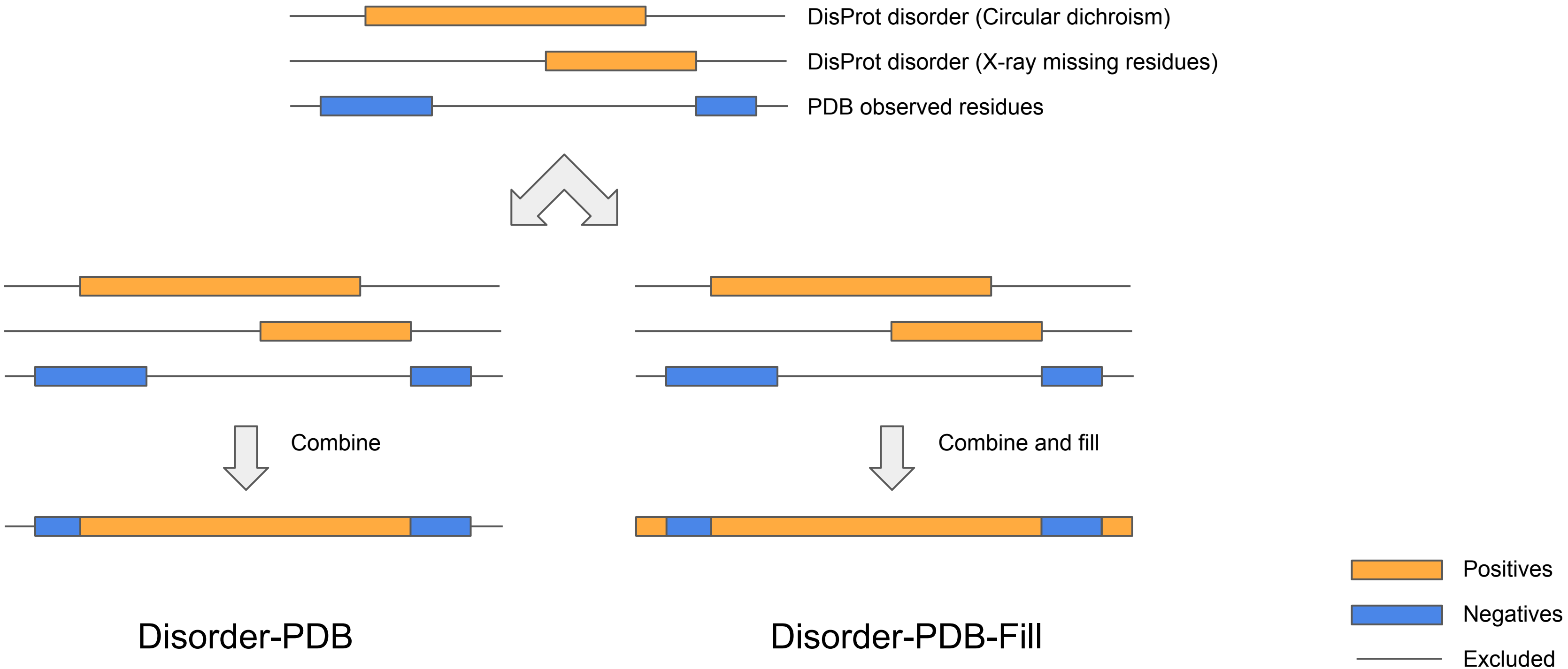
Disorder

Disorder-NoX

🟧 Positives

🟦 Negatives

Reference definition - Disorder & PDB



DisProt disorder (Circular dichroism)

DisProt disorder (X-ray missing residues)

PDB observed residues

Combine

Combine and fill

Disorder-PDB

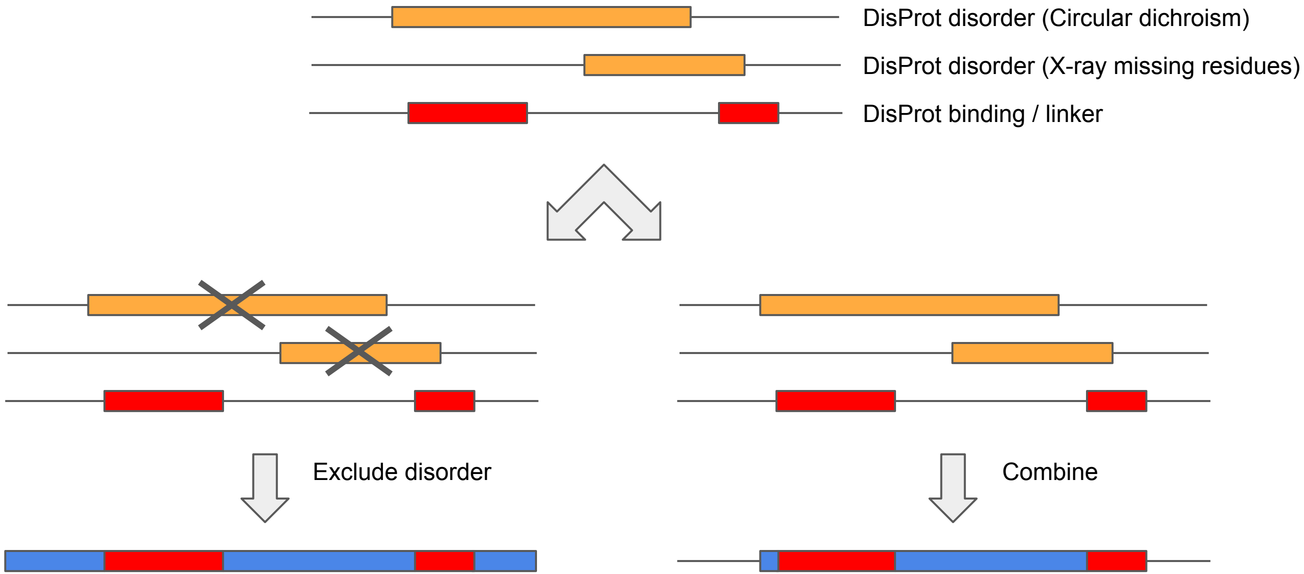
Disorder-PDB-Fill

🟧 Positives

🟦 Negatives

— Excluded

Reference definition - Binding / Linker


DisProt disorder (Circular dichroism)
DisProt disorder (X-ray missing residues)
DisProt binding / linker
Exclude disorder
Combine

Binding / Linker /
Binding-NucleicAcid

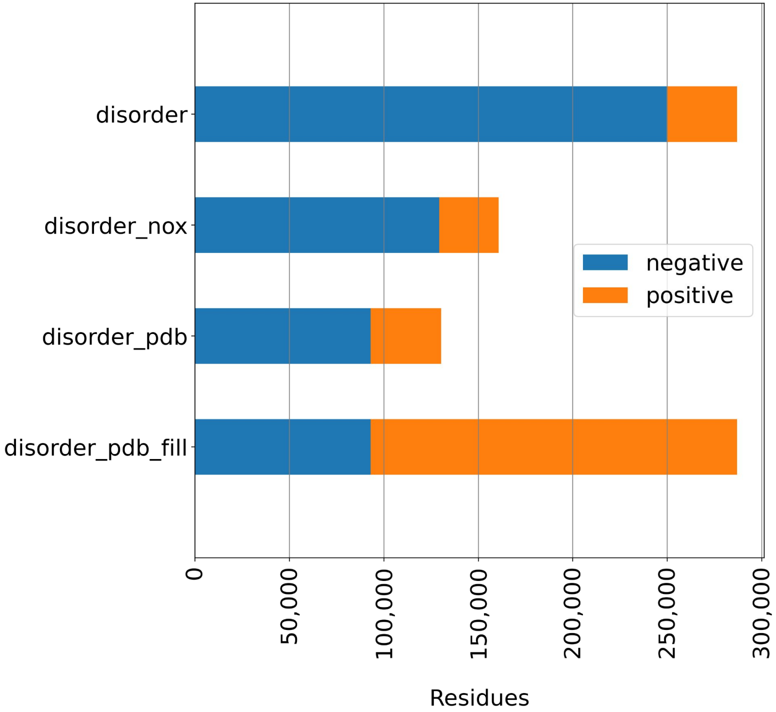
Binding-Disorder /
Linker-Disorder

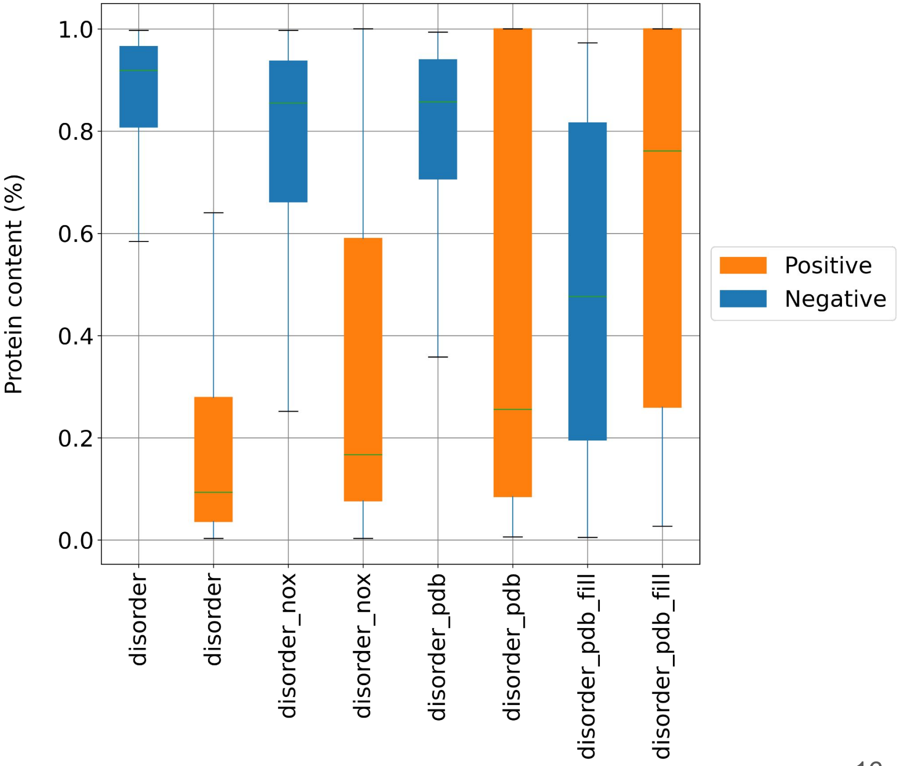
🟥 Positives

🟦 Negatives

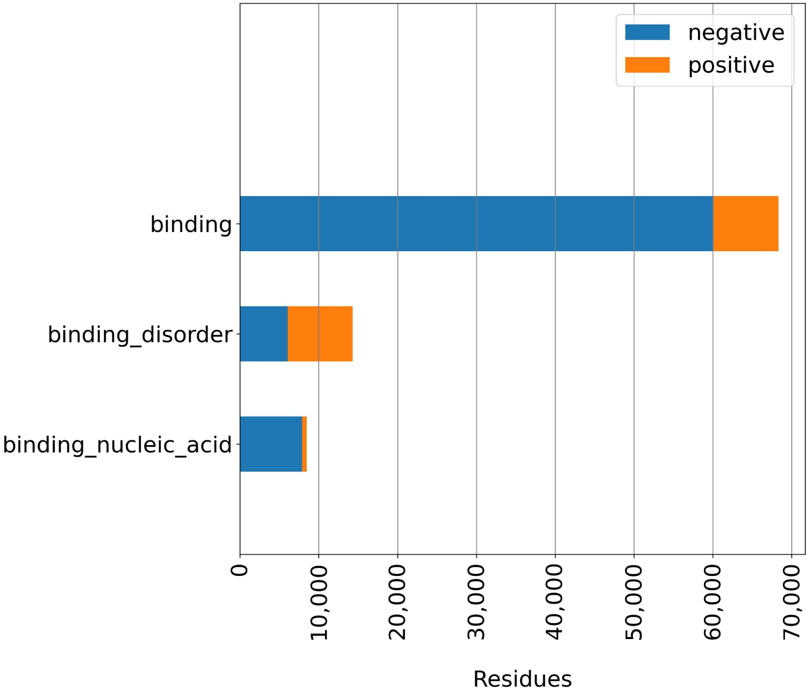
— Excluded

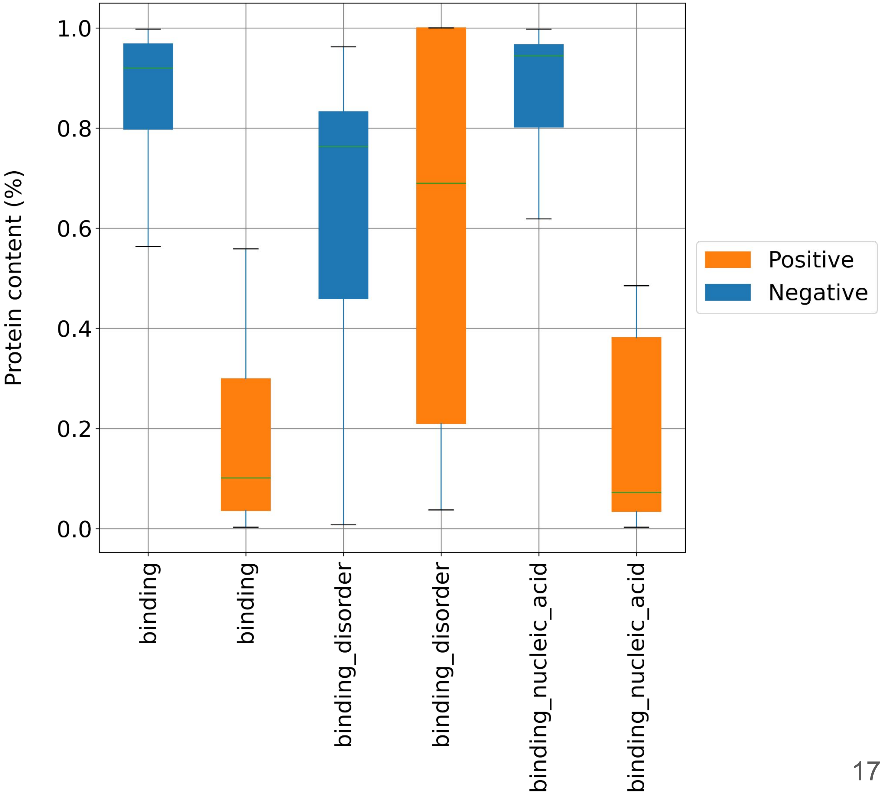
Class distribution - Disorder



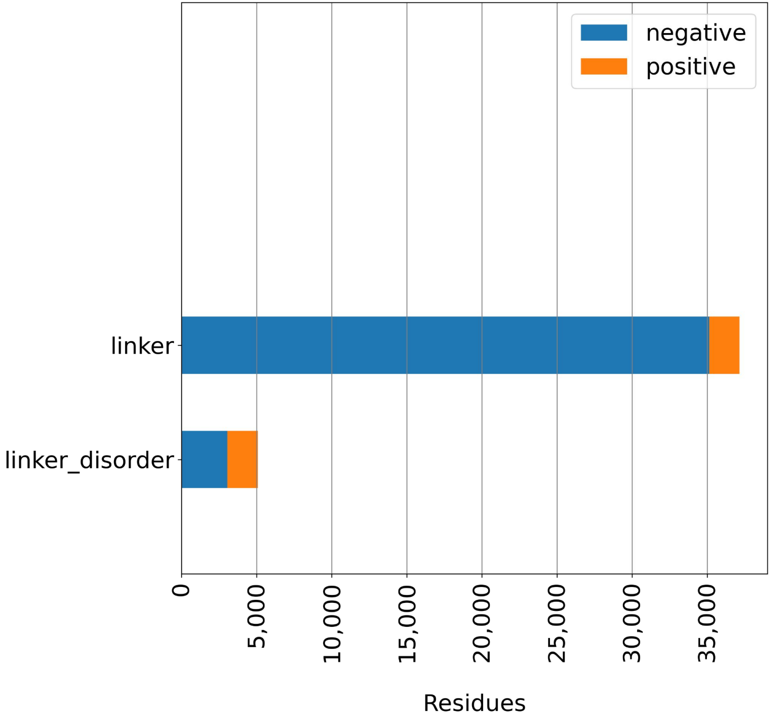


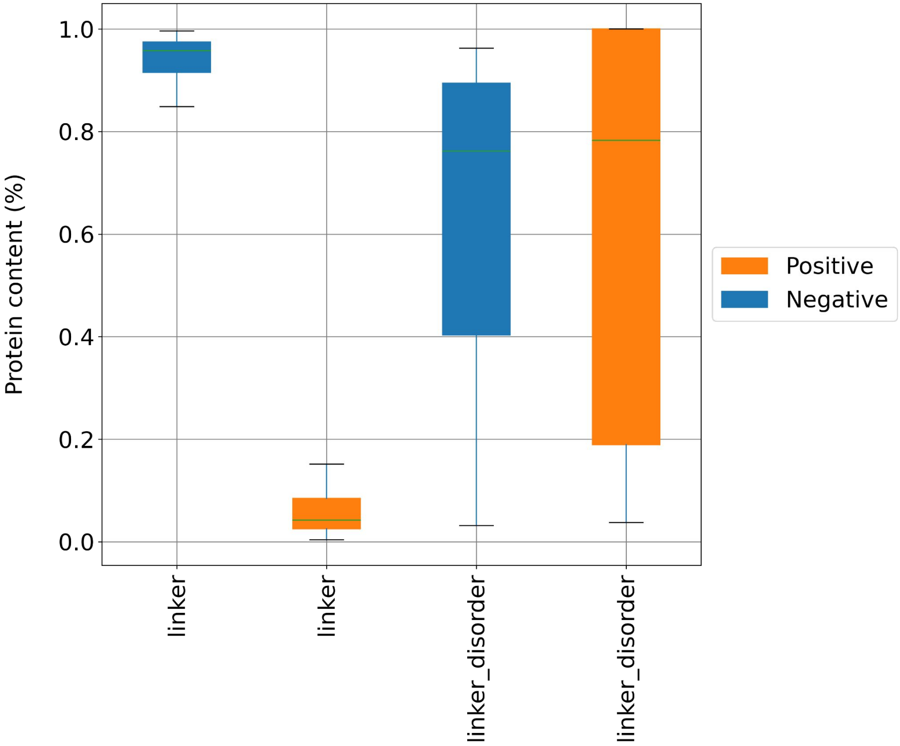
Class distribution - Binding





Class distribution - Linker





CAID-2

Assessment

1222·2022 800 ANNI

Università
degli Studi
di Padova

Dipartimento di Scienze Biomediche

IDP f(un)

elixir ITALY

BioComputing

Assessment

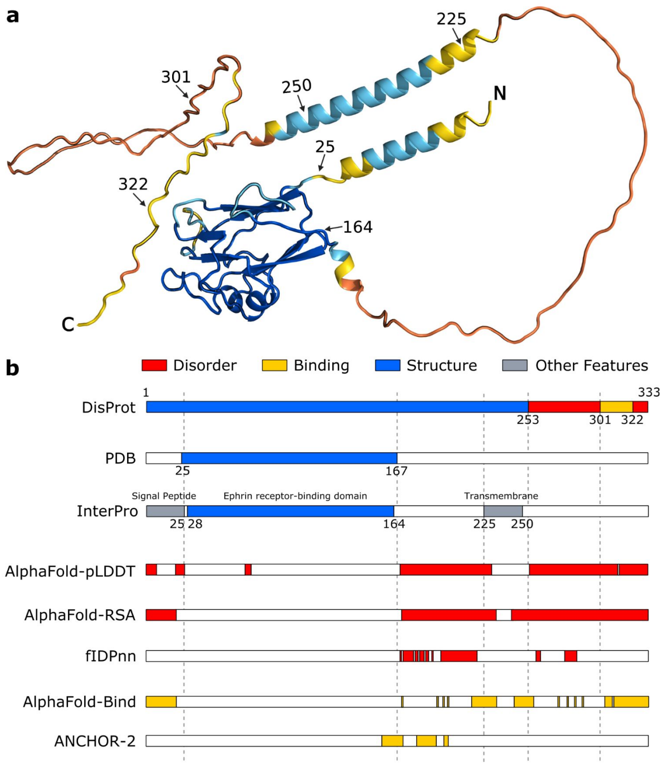
- Predictors **thresholds** are selected in order to optimize the **F1-score** in the considered benchmark
- Statistics are provided both at the **dataset** level or averaged over **targets**
- **Baseline**
  - **Random**
  - **Shuffled dataset** →Class imbalance at the dataset level is preserved
- Assessment **code**
  - **CodeOcean** capsule - https://codeocean.com/capsule/2223745/tree/v1
  - **GitHub** (“v2” branch → CAID2) - https://github.com/BioComputingUP/CAID

AlphaFold & disorder

- AlphaFold-**disorder**
  - 1 - pLDDT
- AlphaFold-**RSA**
  - DSSP relative solvent accessibility)
- AlphaFold-**Binding**
  - $$\begin{cases} AlphaFold\_RSA, & AlphaFold\_RSA \le T \\ T + pLDDT(1-T), & AlphaFold\_RSA > T \end{cases}$$

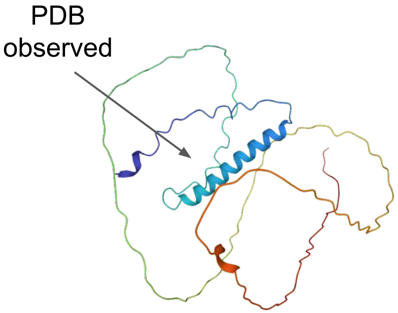
Piovesan D, Monzon AM, Tosatto SCE. *Intrinsic protein disorder and conditional folding in AlphaFoldDB.* Protein Sci. 2022. 31(11):e4466

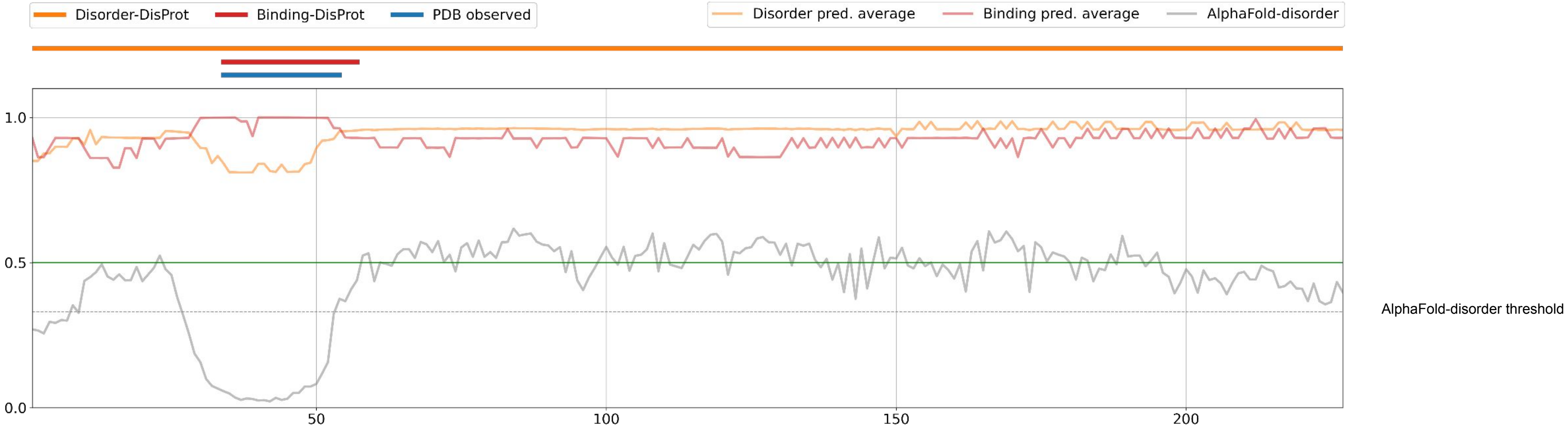
DP01588 (P52799)



DP02342 - P06837

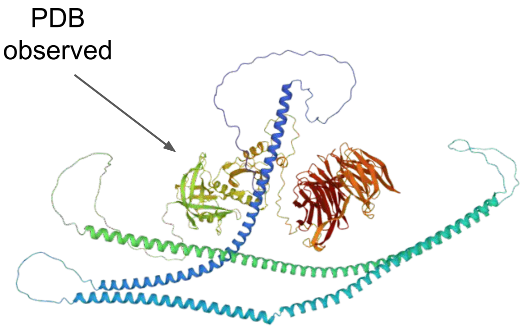
- *Neuromodulin*
- Fully disordered protein
- Average F_{max} ~ 0.96

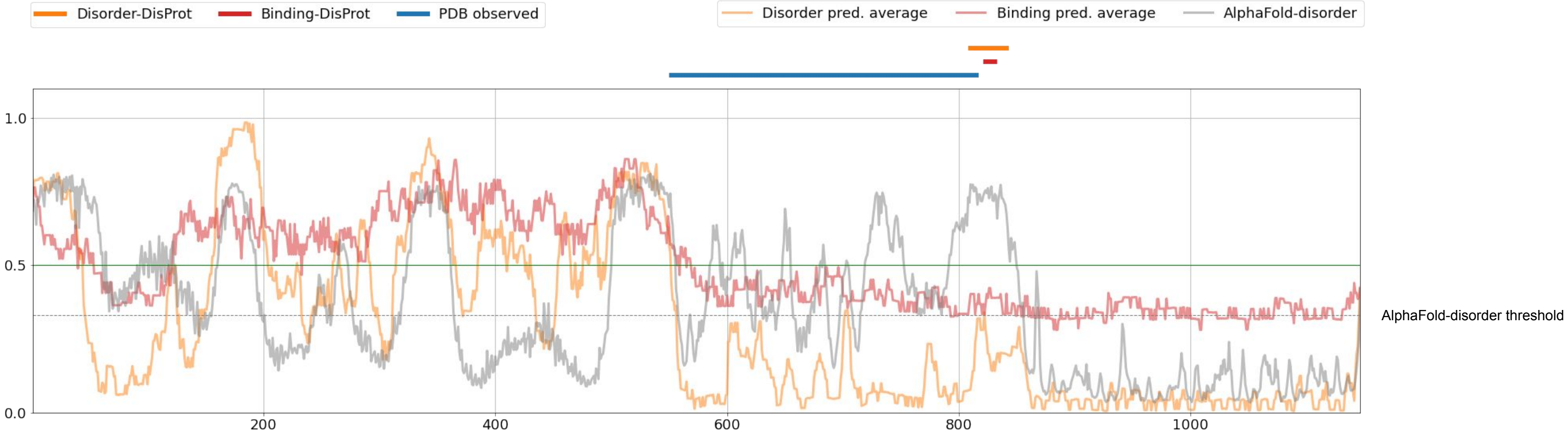




DP02959 - P42527

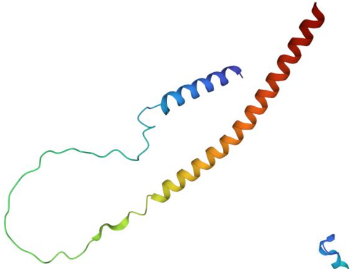
- *Myosin heavy chain kinase A*
- X-ray missing residues evidence
- Average disorder F_{max} ~ 0.1

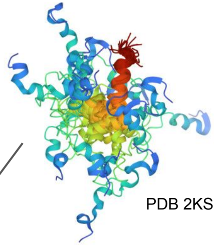

PDB observed



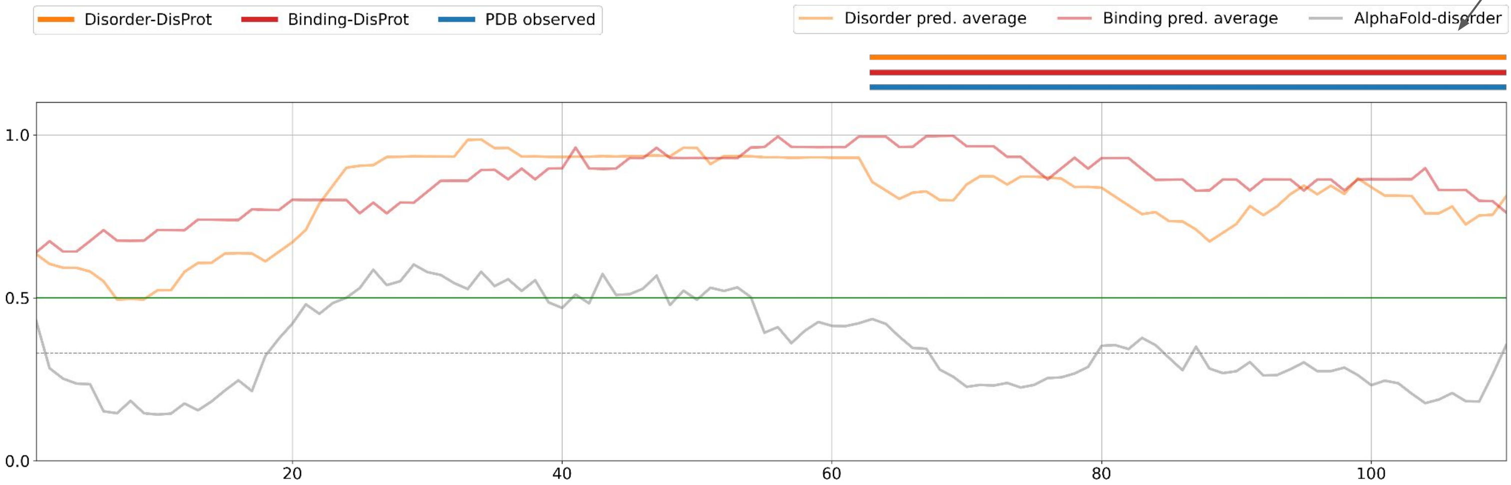
DP03635 - P81605

- *Dermcidin*
- Average F_{max} ~ 0.74 (Disorder-PDB), ~ 0.45 (Disorder-NoX)
- Circular dichroism and NMR evidence



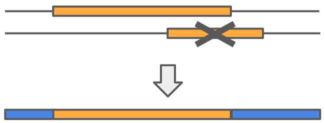


PDB 2KSG

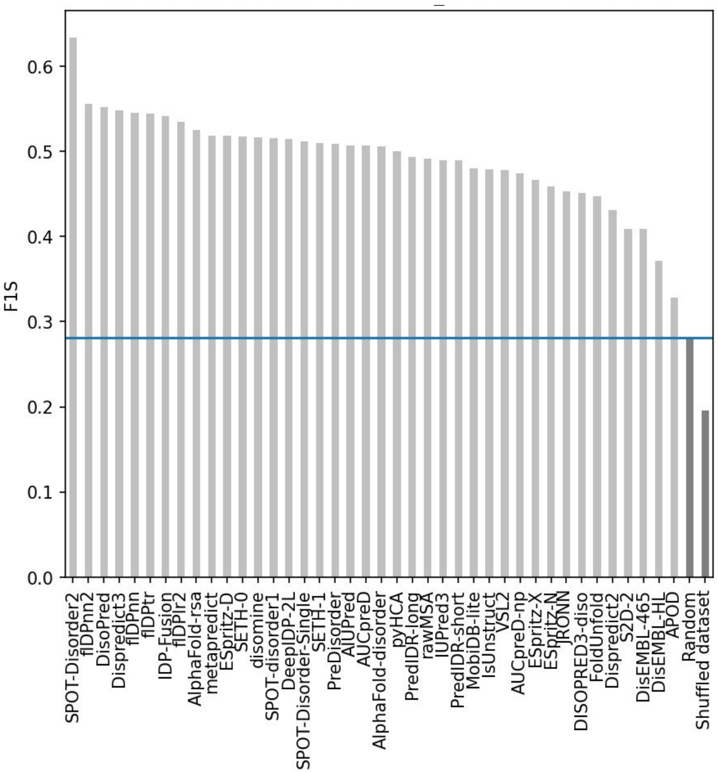


AlphaFold-disorder threshold

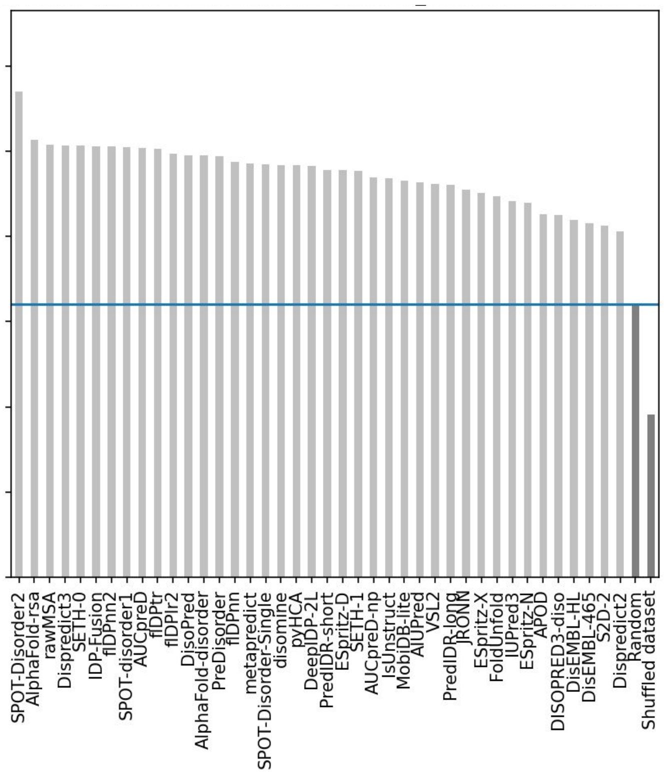
Disorder-NoX (no X-ray)



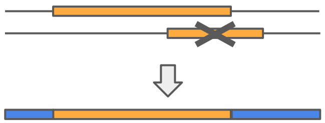
Dataset evaluation

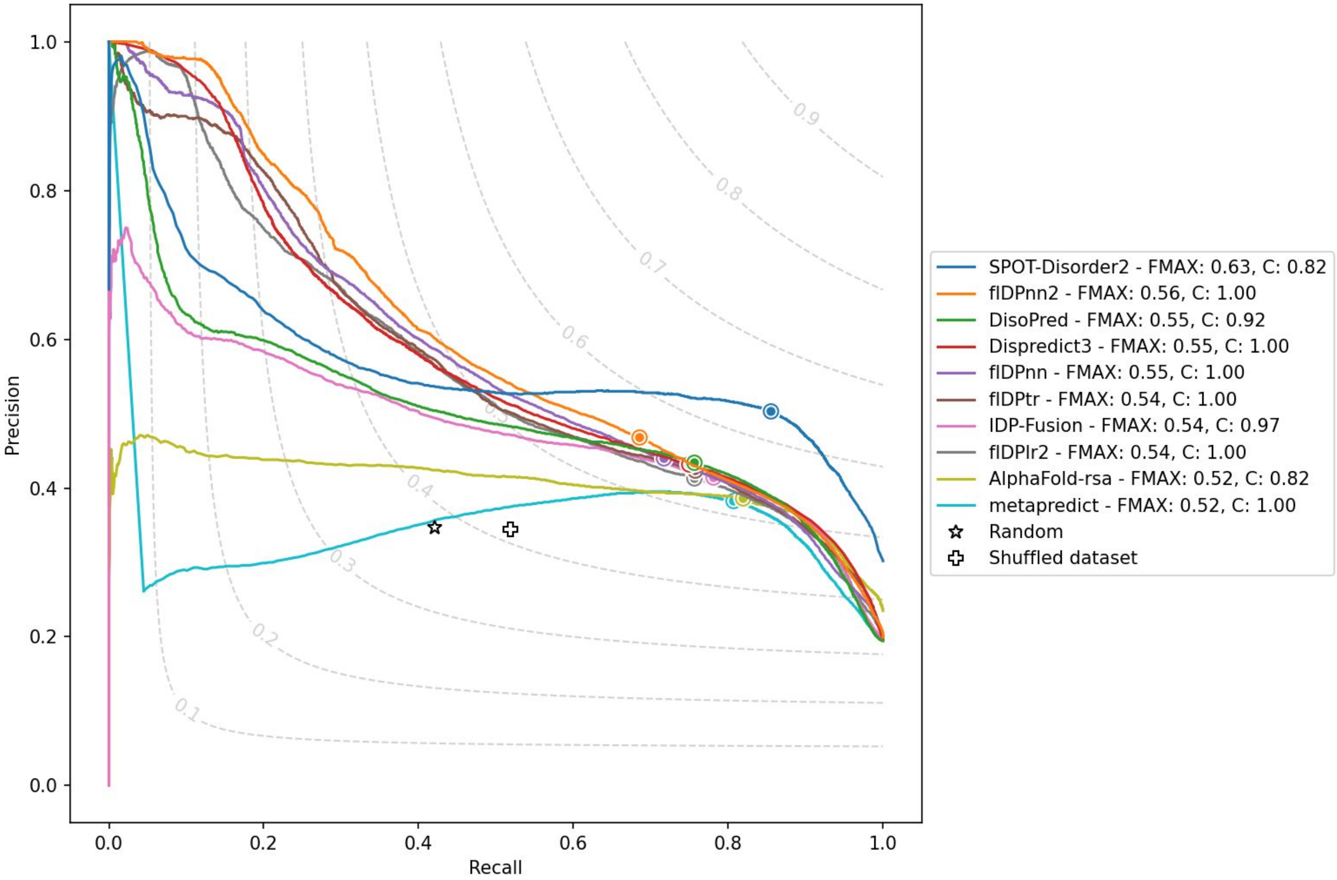


Target average



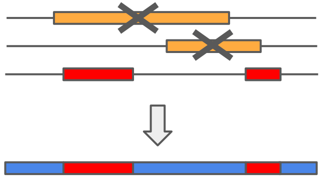
Disorder-NoX (no X-ray)

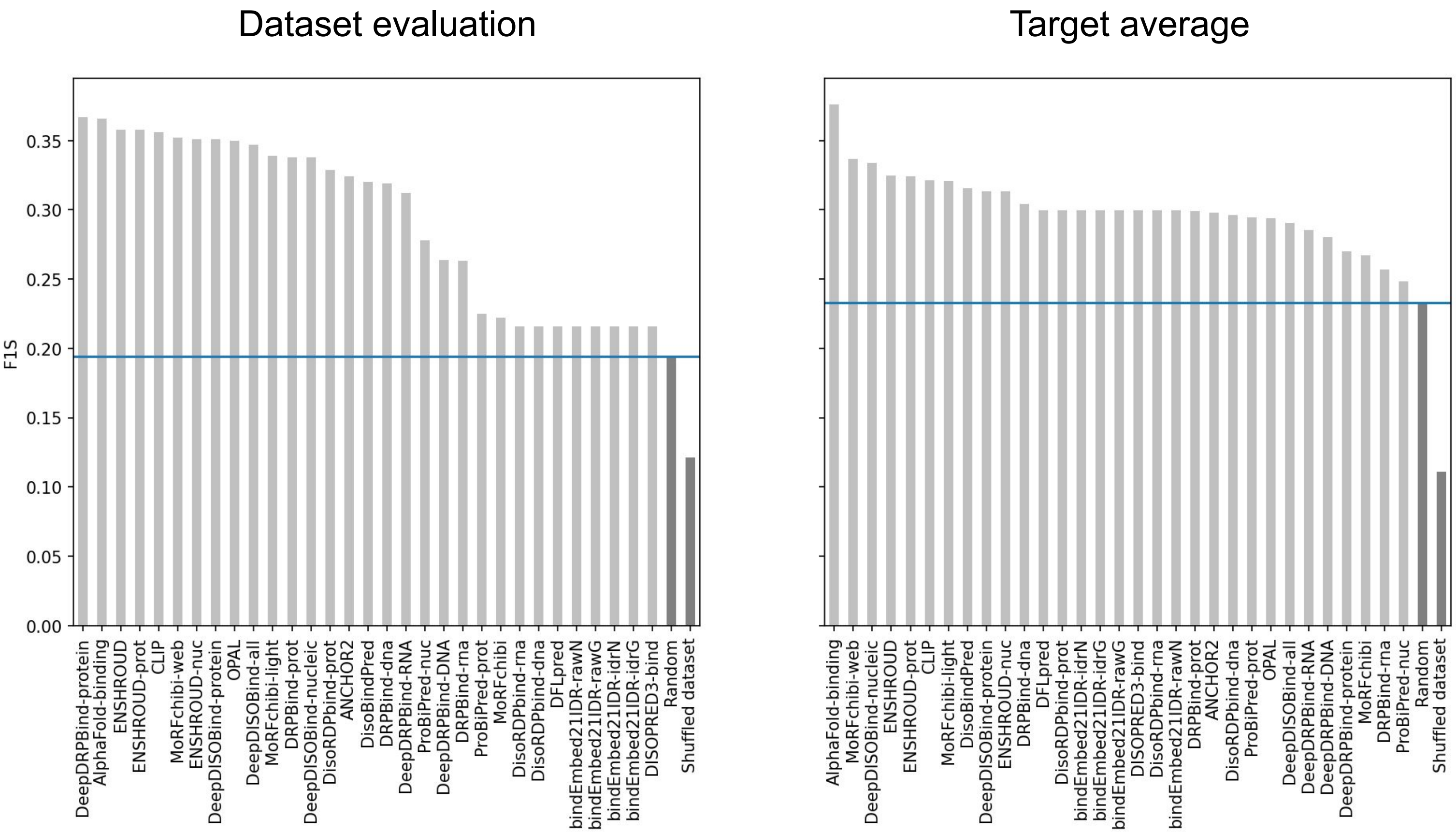




Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
SPOT-Disorder2 - FMAX: 0.63, C: 0.82
fIDPnn2 - FMAX: 0.56, C: 1.00
DisoPred - FMAX: 0.55, C: 0.92
Dispredict3 - FMAX: 0.55, C: 1.00
fIDPnn - FMAX: 0.55, C: 1.00
fIDPtr - FMAX: 0.54, C: 1.00
IDP-Fusion - FMAX: 0.54, C: 0.97
fIDPlr2 - FMAX: 0.54, C: 1.00
AlphaFold-rsa - FMAX: 0.52, C: 0.82
metapredict - FMAX: 0.52, C: 1.00
☆ Random
✚ Shuffled dataset

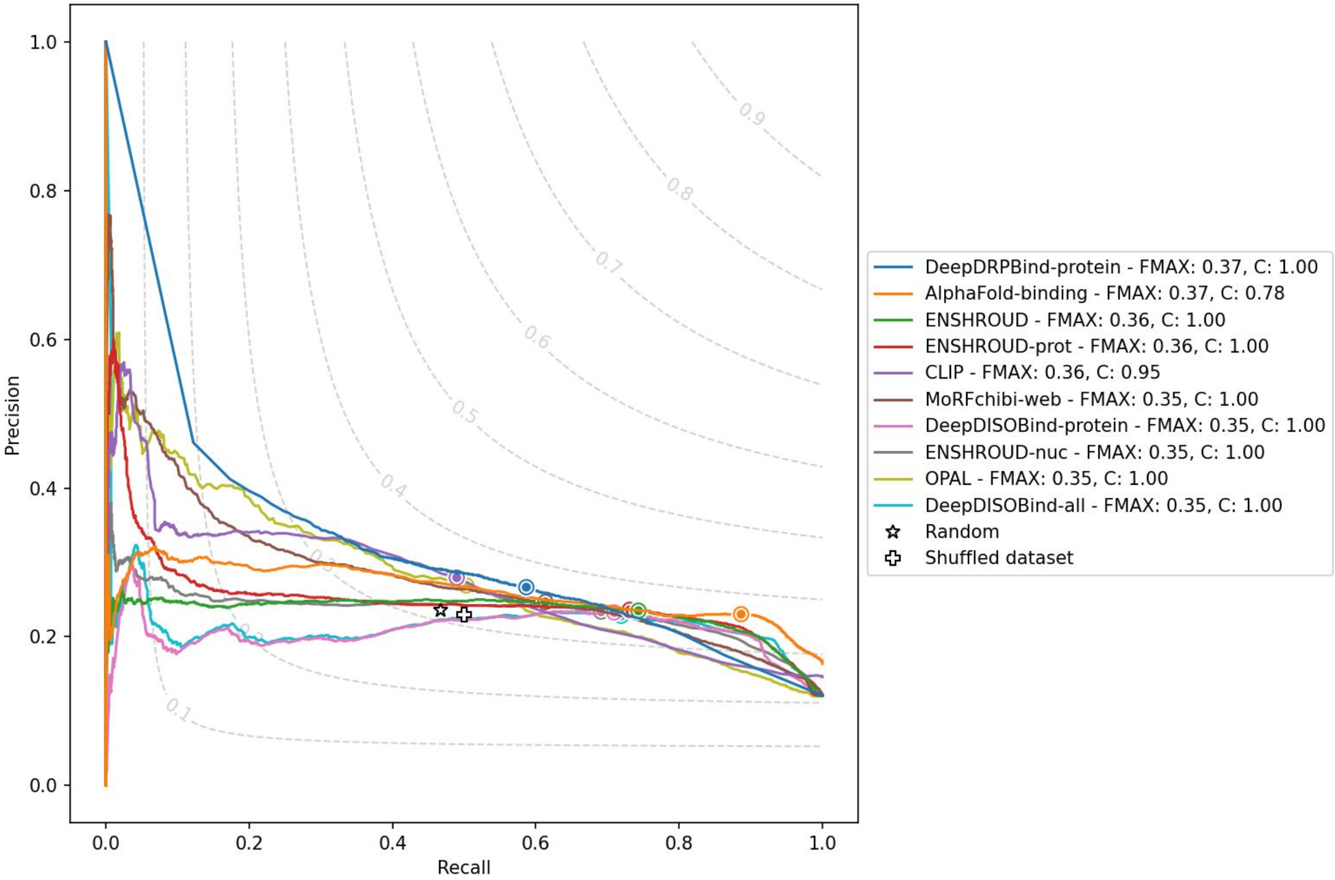
Binding

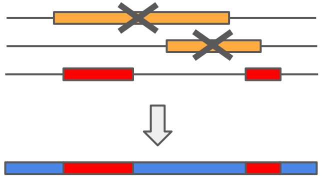




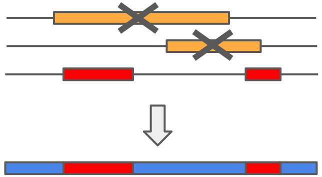
Dataset evaluation
Target average
F1S
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
DeepDRPBind-protein
AlphaFold-binding
ENSHROUD
ENSHROUD-prot
CLIP
MoRFchibi-web
ENSHROUD-nuc
DeepDISOBind-protein
OPAL
DeepDISOBind-all
MoRFchibi-light
DRPBind-prot
DeepDISOBind-nucleic
DisoRDPbind-prot
ANCHOR2
DisoBindPred
DRPBind-dna
DeepDRPBind-RNA
ProBiPred-nuc
DeepDRPBind-DNA
DRPBind-rna
ProBiPred-prot
MoRFchibi
DisoRDPbind-rna
DisoRDPbind-dna
DFLpred
bindEmbed21IDR-rawN
bindEmbed21IDR-rawG
bindEmbed21IDR-idrN
bindEmbed21IDR-idrG
DISOPRED3-bind
Random
Shuffled dataset
AlphaFold-binding
MoRFchibi-web
DeepDISOBind-nucleic
ENSHROUD
ENSHROUD-prot
CLIP
MoRFchibi-light
DisoBindPred
DeepDISOBind-protein
ENSHROUD-nuc
DRPBind-dna
DFLpred
DisoRDPbind-prot
bindEmbed21IDR-idrN
bindEmbed21IDR-idrG
bindEmbed21IDR-rawG
DISOPRED3-bind
DisoRDPbind-rna
bindEmbed21IDR-rawN
DRPBind-prot
ANCHOR2
DisoRDPbind-dna
ProBiPred-prot
OPAL
DeepDISOBind-all
DeepDRPBind-RNA
DeepDRPBind-DNA
DeepDRPBind-protein
MoRFchibi
DRPBind-rna
ProBiPred-nuc
Random
Shuffled dataset

Binding

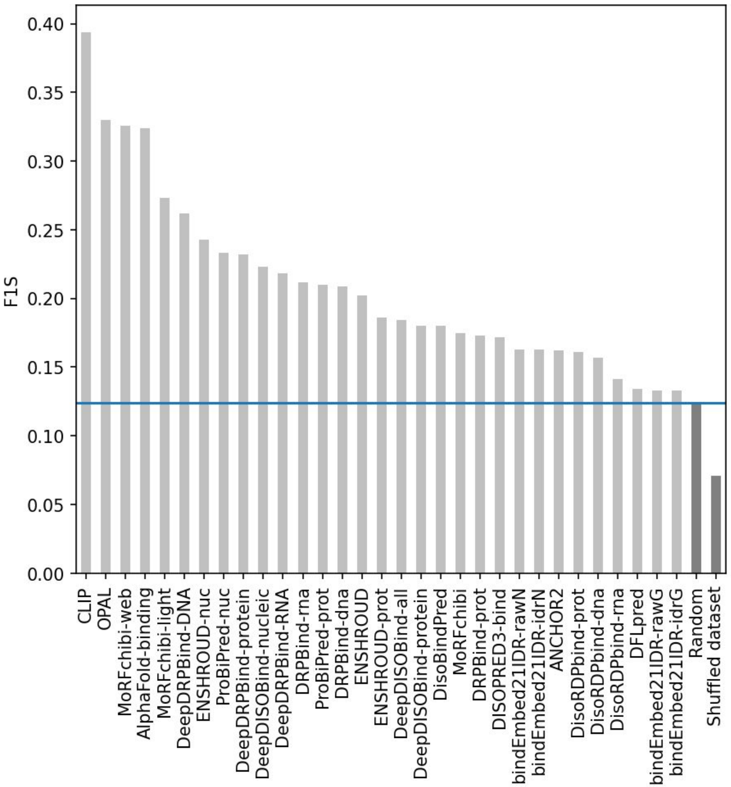

Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
DeepDRPBind-protein - FMAX: 0.37, C: 1.00
AlphaFold-binding - FMAX: 0.37, C: 0.78
ENSHROUD - FMAX: 0.36, C: 1.00
ENSHROUD-prot - FMAX: 0.36, C: 1.00
CLIP - FMAX: 0.36, C: 0.95
MoRFchibi-web - FMAX: 0.35, C: 1.00
DeepDISOBind-protein - FMAX: 0.35, C: 1.00
ENSHROUD-nuc - FMAX: 0.35, C: 1.00
OPAL - FMAX: 0.35, C: 1.00
DeepDISOBind-all - FMAX: 0.35, C: 1.00
☆ Random
✣ Shuffled dataset



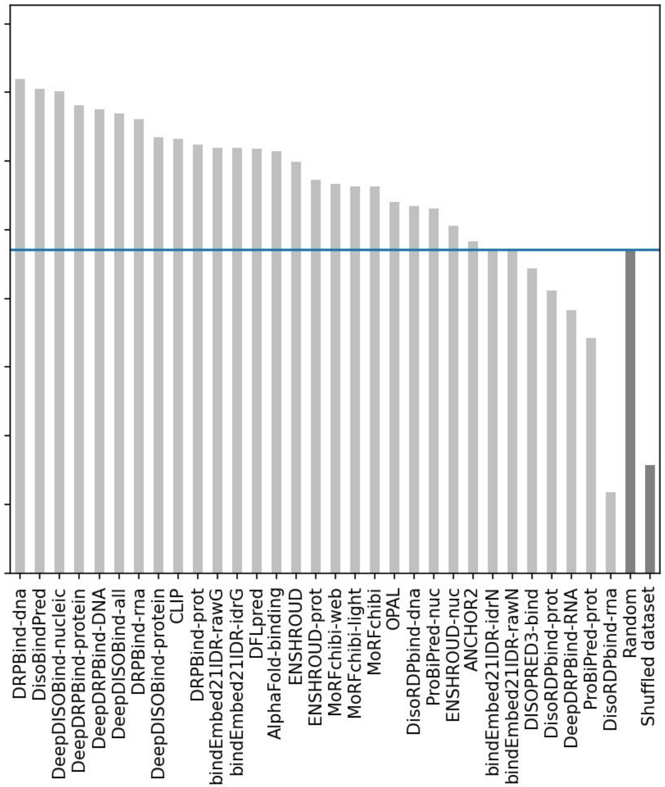
Binding-NucleicAcid



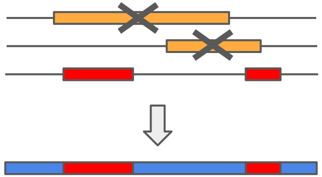
Dataset evaluation

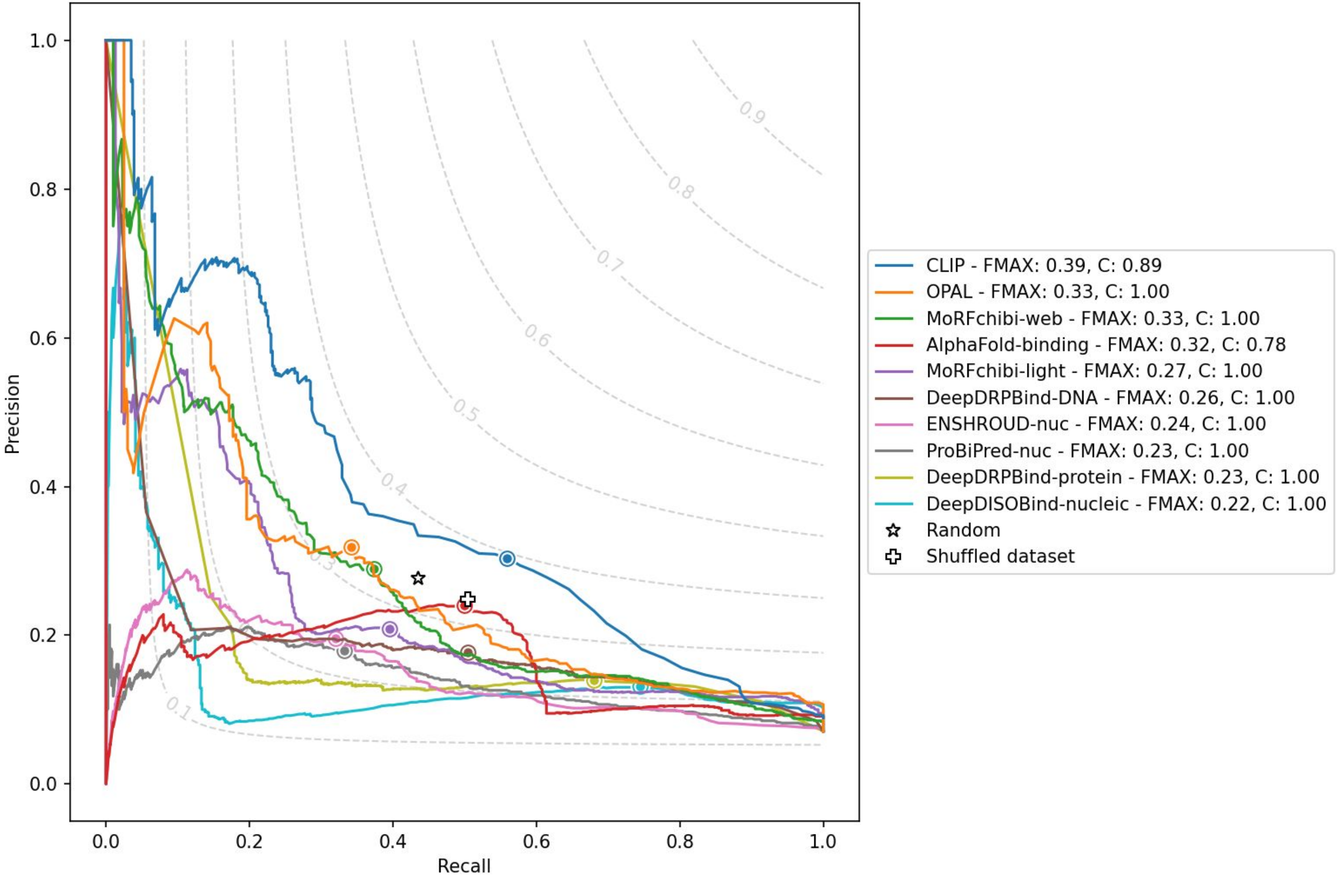


Target average



Binding-NucleicAcid

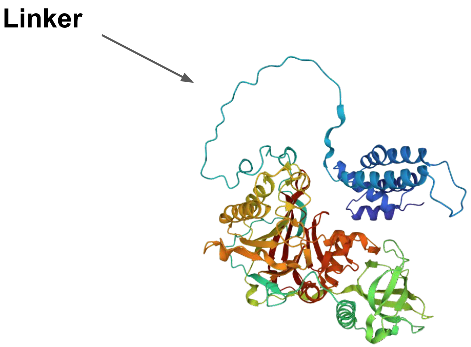


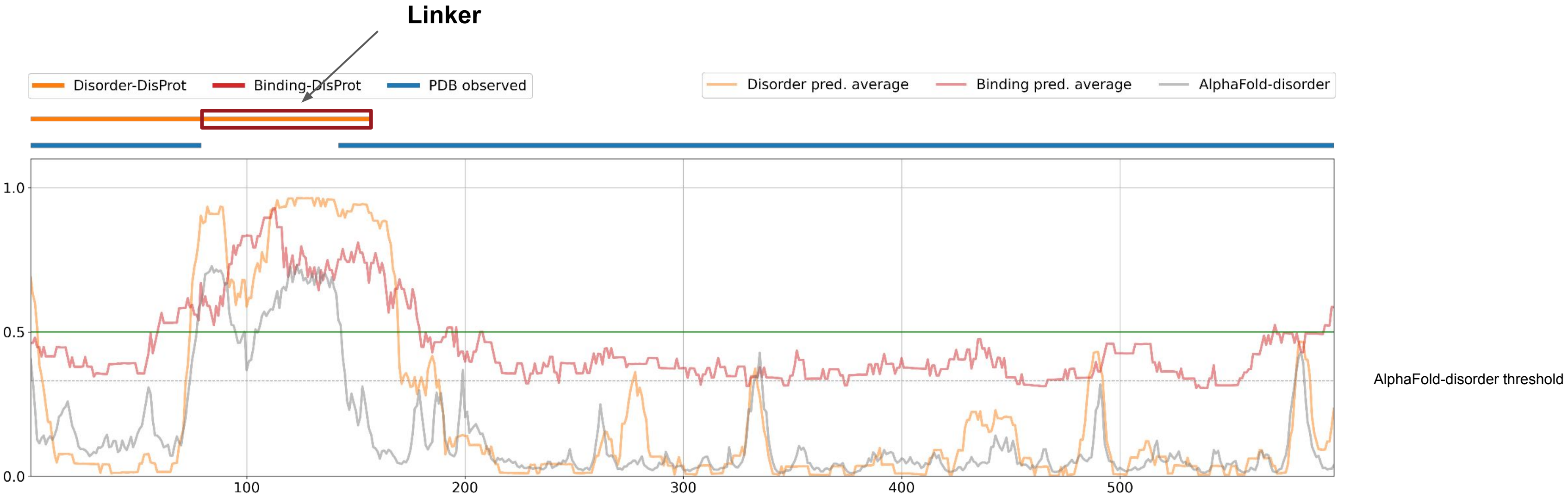


1.0
0.8
0.6
0.4
0.2
0.0
Precision
0.0
0.2
0.4
0.6
0.8
1.0
Recall
0.1
0.3
0.4
0.5
0.6
0.7
0.8
0.9
CLIP - FMAX: 0.39, C: 0.89
OPAL - FMAX: 0.33, C: 1.00
MoRFchibi-web - FMAX: 0.33, C: 1.00
AlphaFold-binding - FMAX: 0.32, C: 0.78
MoRFchibi-light - FMAX: 0.27, C: 1.00
DeepDRPBind-DNA - FMAX: 0.26, C: 1.00
ENSHROUD-nuc - FMAX: 0.24, C: 1.00
ProBiPred-nuc - FMAX: 0.23, C: 1.00
DeepDRPBind-protein - FMAX: 0.23, C: 1.00
DeepDISOBind-nucleic - FMAX: 0.22, C: 1.00
☆ Random
✣ Shuffled dataset

DP02759 - Q14181

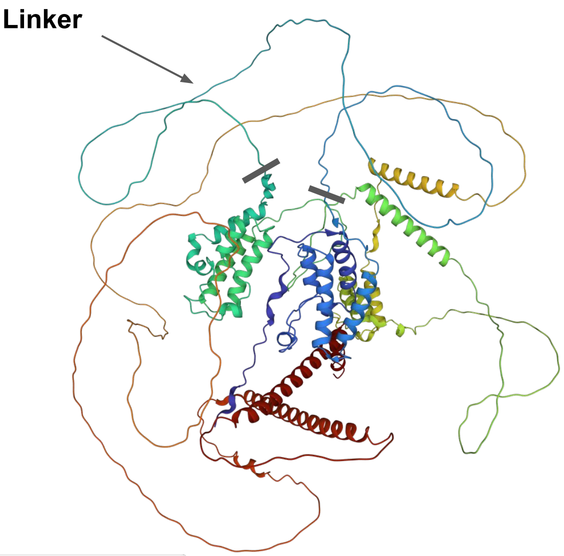
- *DNA Polymerase Alpha subunit B*
- X-ray missing residues evidence
- Average F_{max} ~ 0.65

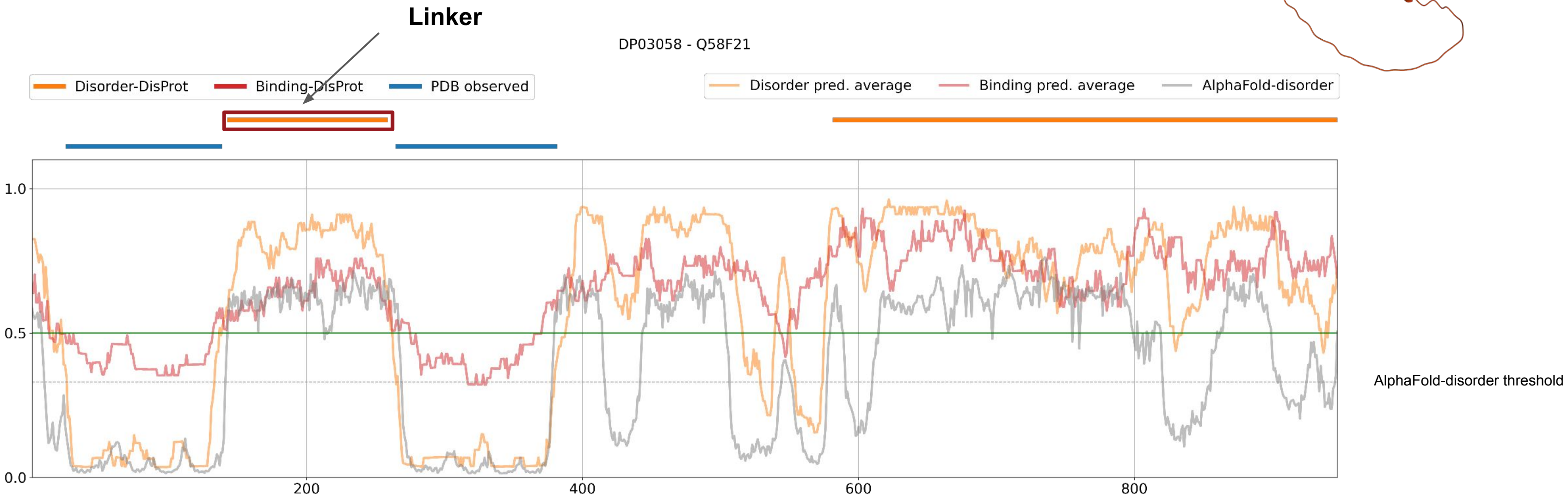




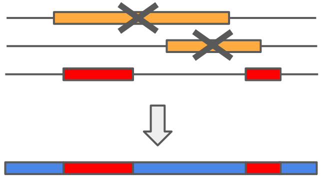
DP03058 - Q58F21

- *Bromodomain testis-specific protein*
- X-ray missing residues and NMR evidence
- Average F_{max} ~ 0.39

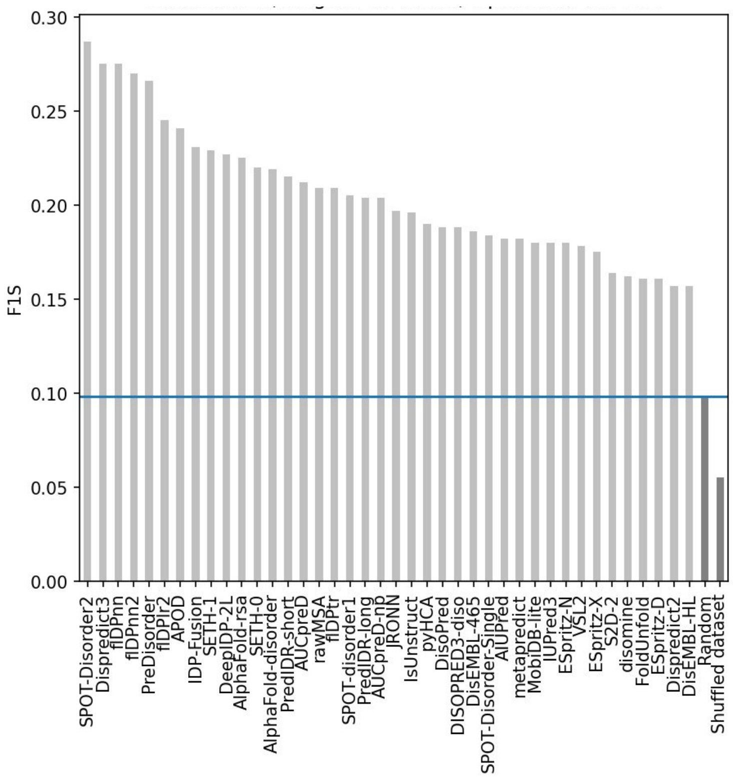




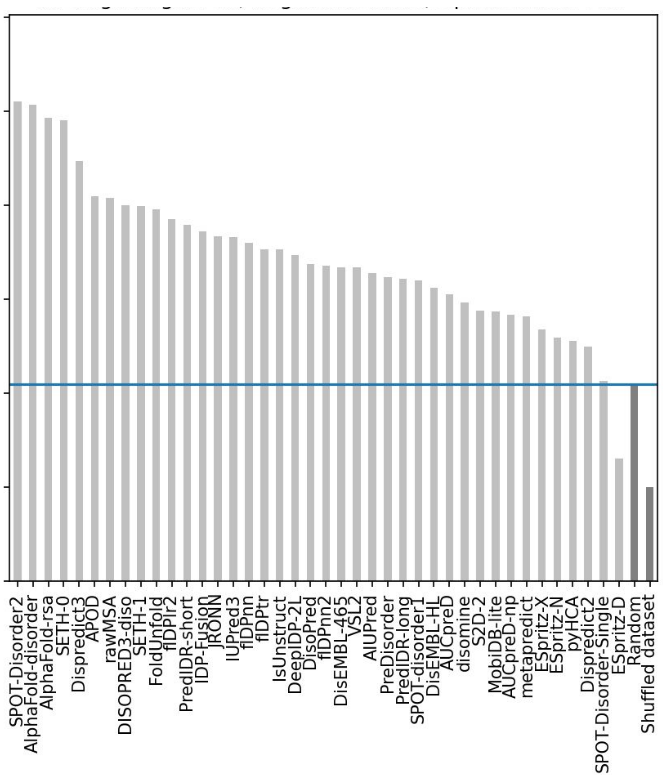
Linker



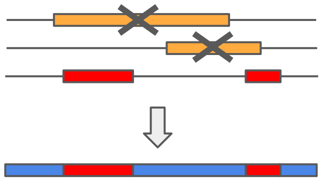
Dataset evaluation

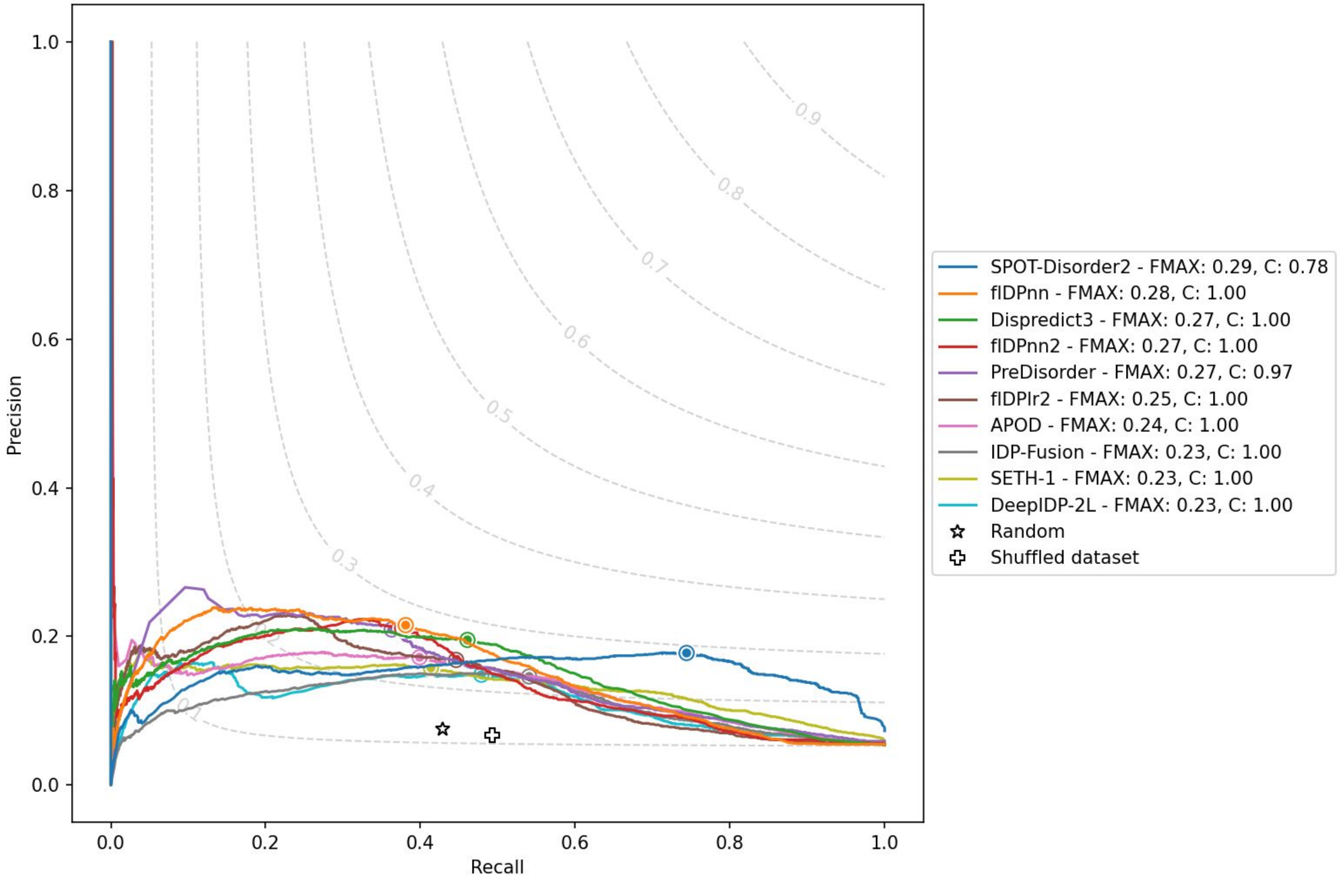


Target average



Linker




Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.1
SPOT-Disorder2 - FMAX: 0.29, C: 0.78
fIDPnn - FMAX: 0.28, C: 1.00
Dispredict3 - FMAX: 0.27, C: 1.00
fIDPnn2 - FMAX: 0.27, C: 1.00
PreDisorder - FMAX: 0.27, C: 0.97
fIDPlr2 - FMAX: 0.25, C: 1.00
APOD - FMAX: 0.24, C: 1.00
IDP-Fusion - FMAX: 0.23, C: 1.00
SETH-1 - FMAX: 0.23, C: 1.00
DeepIDP-2L - FMAX: 0.23, C: 1.00
☆ Random
✥ Shuffled dataset

Acknowledgements

Organizers

Silvio CE Tosatto

Damiano Piovesan

Alexander M Monzon

Curators

Federica Quaglia

Victoria Nugnes

Maria Cristina Aspromonte

Assessors

Alessio Del Conte

Adel Bouhraoua

BioComputingUP

University of Padova

Biomedical Sciences

https://biocomputingup.it

@BioComputingUP

IDP f(un)

elixir ITALY

800 ANNI 1222·2022

UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI SCIENZE BIOMEDICHE

BioComputing