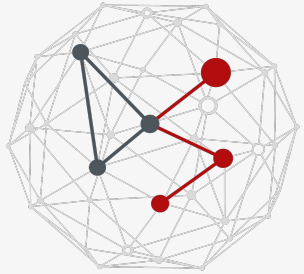

DIPARTIMENTO
MATEMATICA



DATA SCIENCE
UNIVERSITY OF PADOVA

# DISORDER PREDICTION FROM SEQUENCE

Master of Science in Data Science

Damiano Piovesan

# Low complexity - SEG

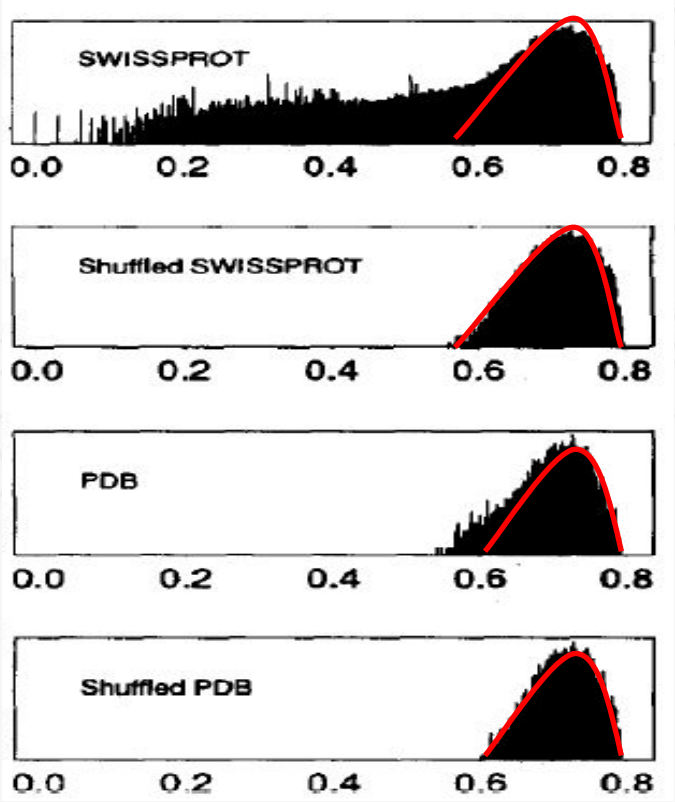
- **Globular** proteins have **high complexity**, while **atypical** regions have **low complexity** (ex. **QQQQQQ**)
- SEG is used by BLAST to mask low complexity regions to lower false positive rate (FPR)
- **SEG → Entropy** in a sequence window of size ***L****, where* $n_i$ is the amino acid count

$$k_2 = -\sum_{i=1}^{20} \frac{n_i}{L}\left(\log \frac{n_i}{L}\right).$$

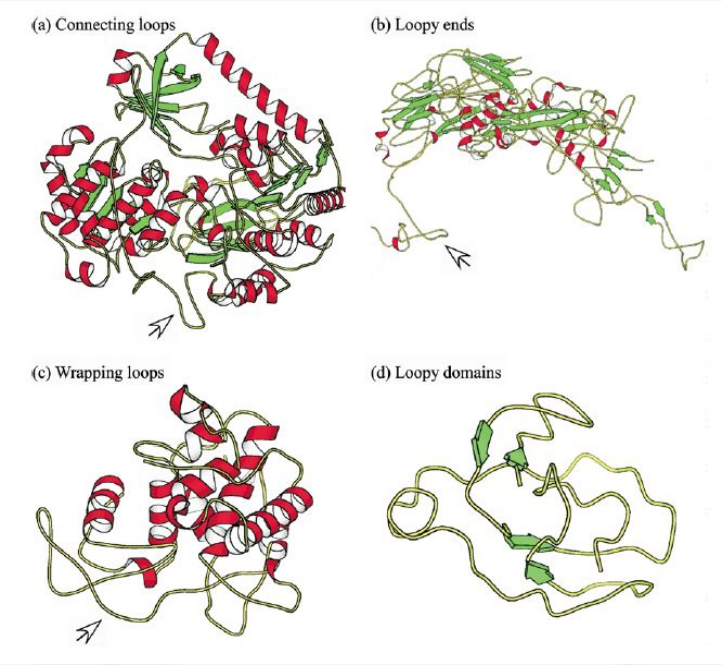
$$0 \geqslant n_i \geqslant L, \quad \sum_{i=1}^{N} n_i = L,$$

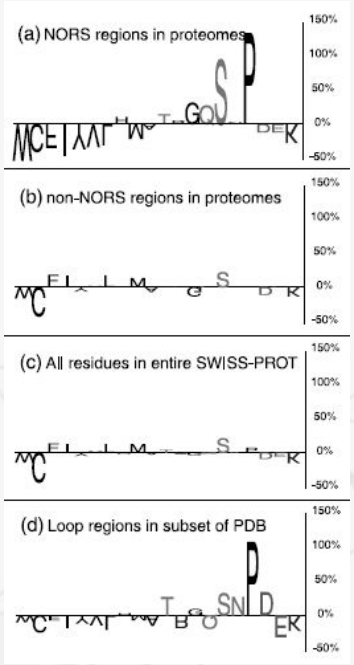
(Wootton 1994)

Complexity (window L = 40)



# Proteins with no regular secondary structure (NORS)




Loopy proteins appear conserved in evolution
Liu, Tan & Rost, JMB 2002

**Flexible residues** are abundant in **NORS** as well as in **PDB loops**

But **NORS** are slightly different compared to **loops**

**NORS** → >70 residues, <12% SS

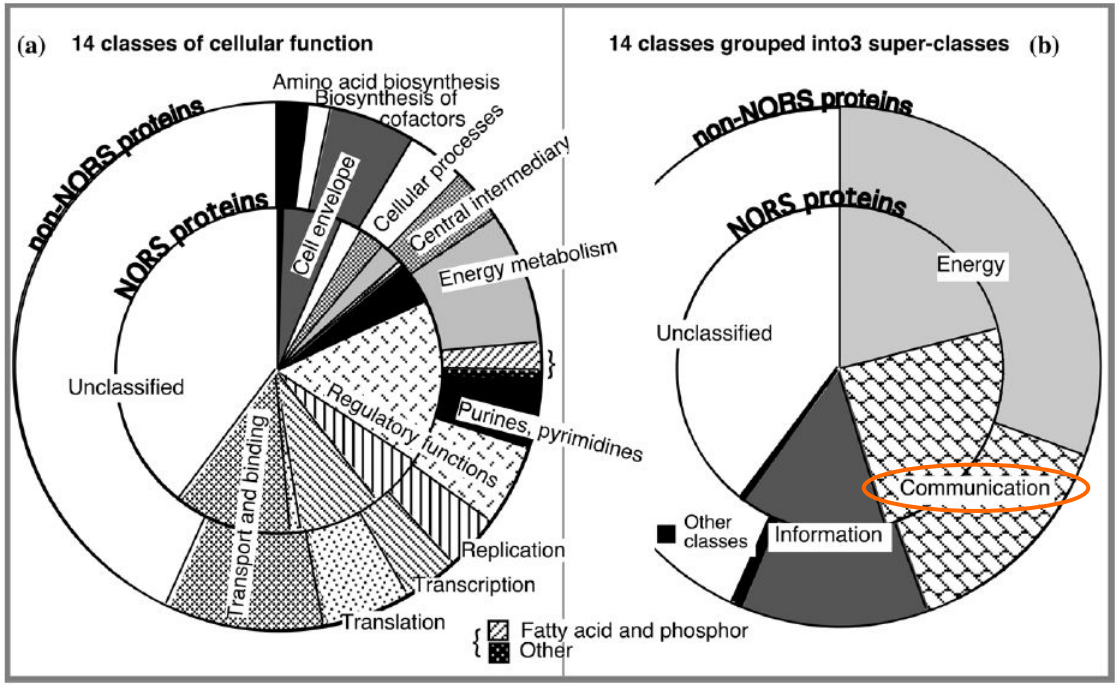
*AA enrichment analysis*

- *Amino acid sorted by flexibility (rigid → flexible)*
- *Background is the PDB*

# Function and distribution of genomes

DIPARTIMENTO MATEMATICA

| Kingdom organism | Number of sequences | Disorder frequency | Length >30 | Length >50 |
|---|---|---|---|---|
| Archaea | 11,742 | 3.8 | 2.0 | 0.7 |
| Bacteria | 35,389 | 5.7 | 4.2 | 1.6 |
| Eukaryota | 88,531 | 18.9 | 33.0 | 19.6 |
| PDB (non-redundant at 95% sequence identity) | 7169 | 3.2 | 0.5 | 0.1 |

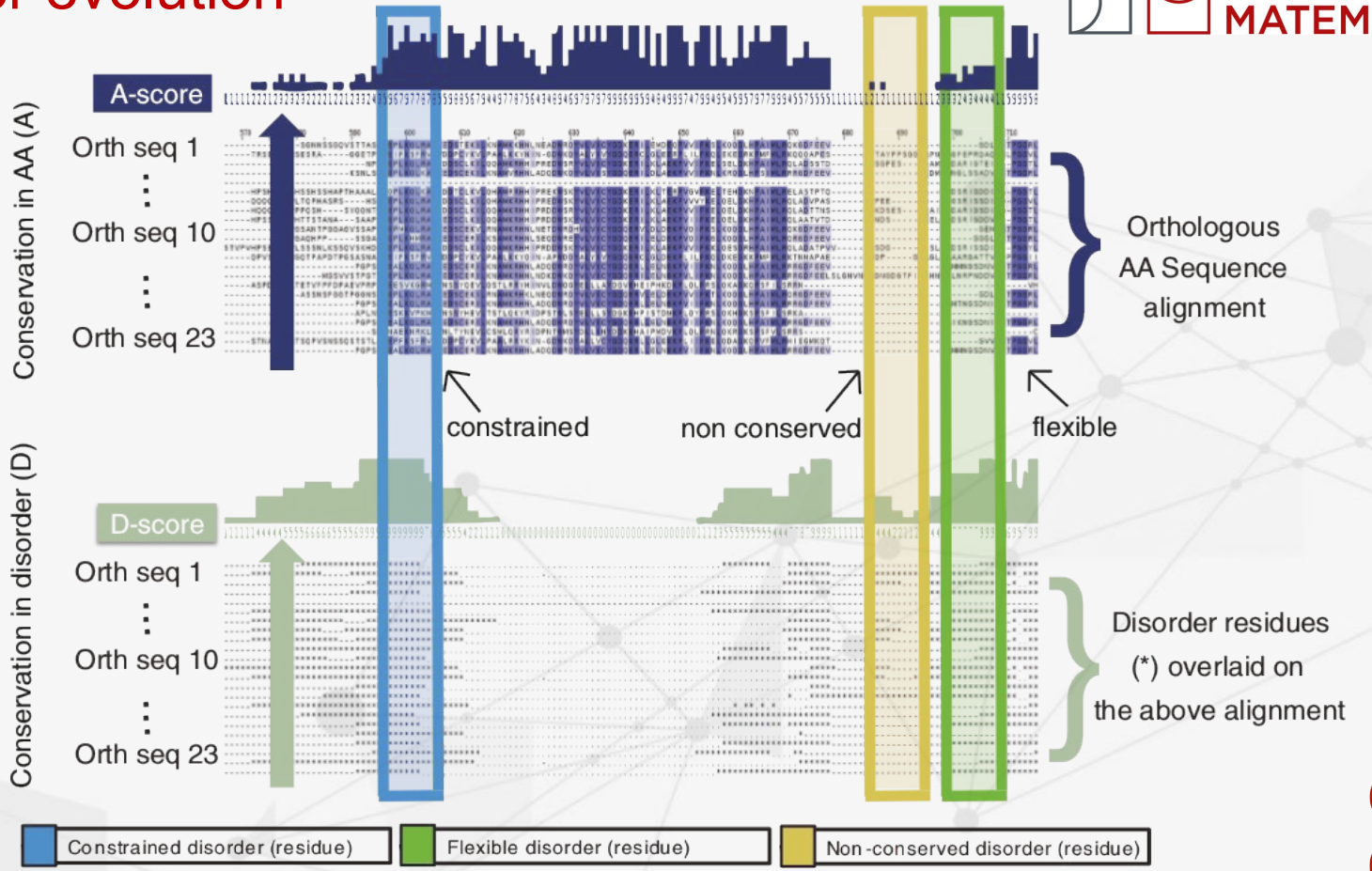



- **20%** of **eukaryotes** proteomes
- **4%** of **prokaryotes**
- **PDB** is **depleted** in long NORS

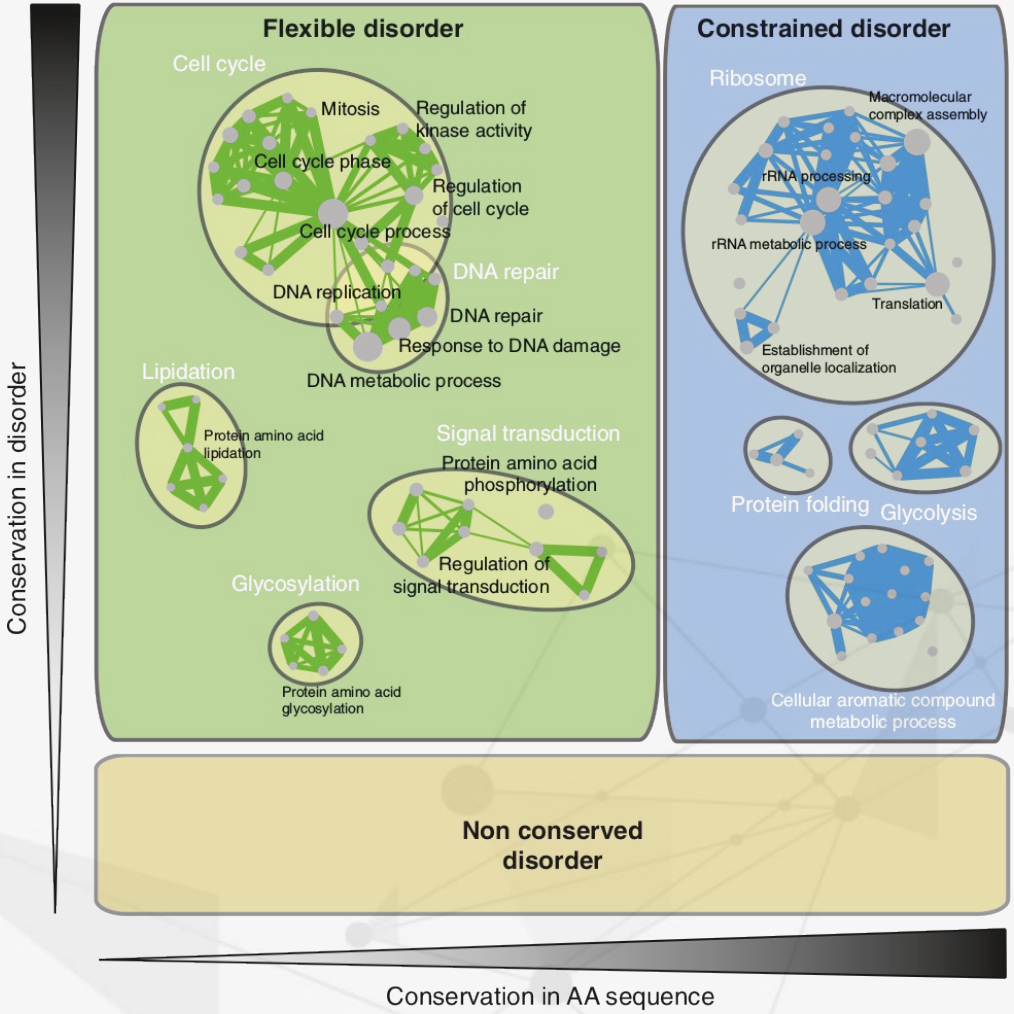
NORS

- More regulatory and transcription-related functions
- Less biosynthesis and energy metabolism functions

# Disorder evolution



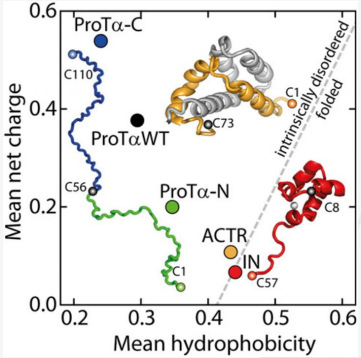

*(Bellay et al., Genome Biol. 2011)*



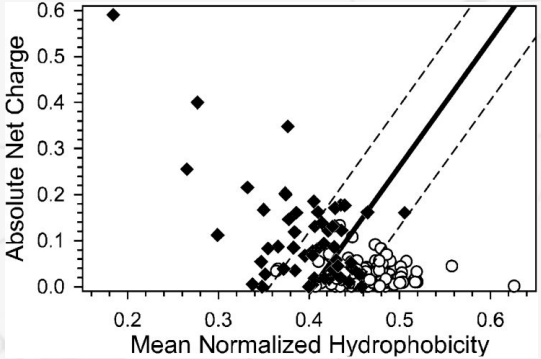

*(Bellay et al., Genome Biol. 2011)*

# Compositional bias and subtypes

- Low hydrophobicity
- Small and hydrophilic amino acids
- Prolines (secondary structure “structure breakers”)



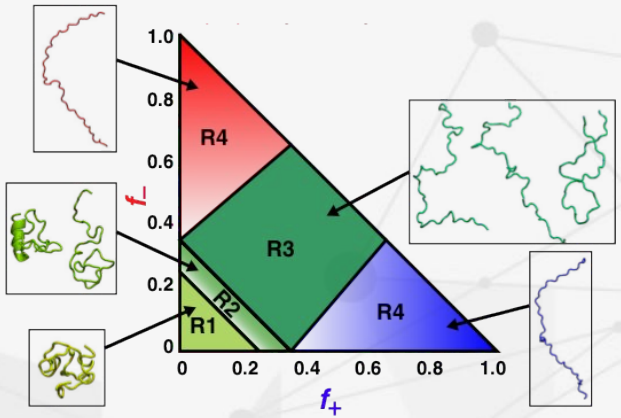

*(Soranno et al, PNAS, 2014)*



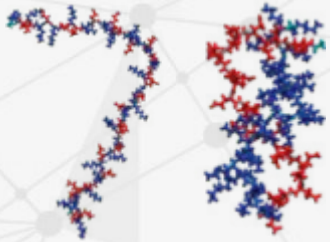

*(Oldfield et al, Biochemistry, 2005)*

# Compositional bias and subtypes

- Classification / conformations
  - Net charge
  - Charge patterning

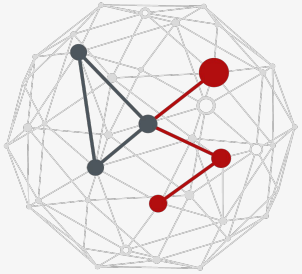



KEKEKEKEKE KKKKKEEEEE



*(Das & Pappu, PNAS, 2013)*


DIPARTIMENTO
MATEMATICA



DATA SCIENCE
UNIVERSITY OF PADOVA

# CAID

## Critical Assessment of protein Intrinsic Disorder prediction

Master of Science in Data Science

Damiano Piovesan

# The challenge

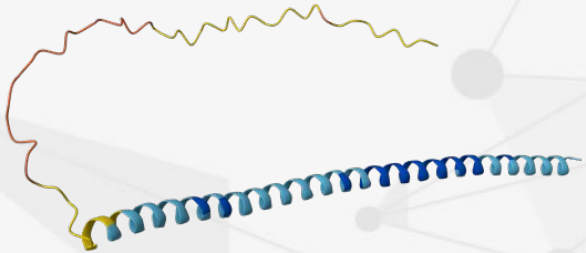
DIPARTIMENTO MATEMATICA

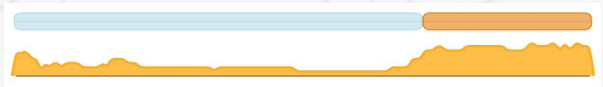
**>P37840 Alpha-synuclein**

MDVFMKGLSKAKEGVVAAAEKTKQGVAEAAGKTKEGVLYVGSKTKEG
VVHGVATVAEKTKEQVTNVGGAVVTGVTAVAQKTVEGAGSIAAATGF
VKKDQLGKNEEGAPQEGILEDMPVDPDNEAYEMPSEEGYQDYEPEA

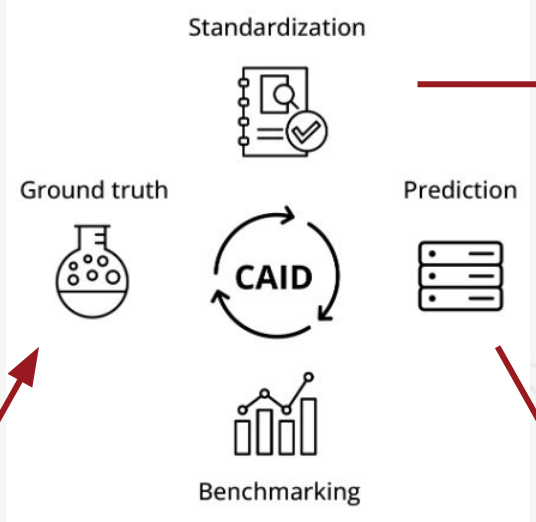
**+**

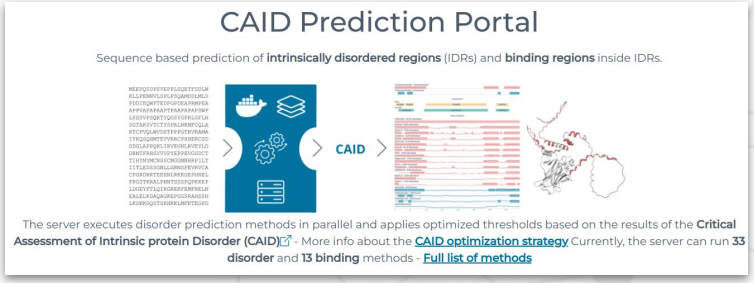
**[PSSM, MSA, …] (optional)**





# The CAID cycle



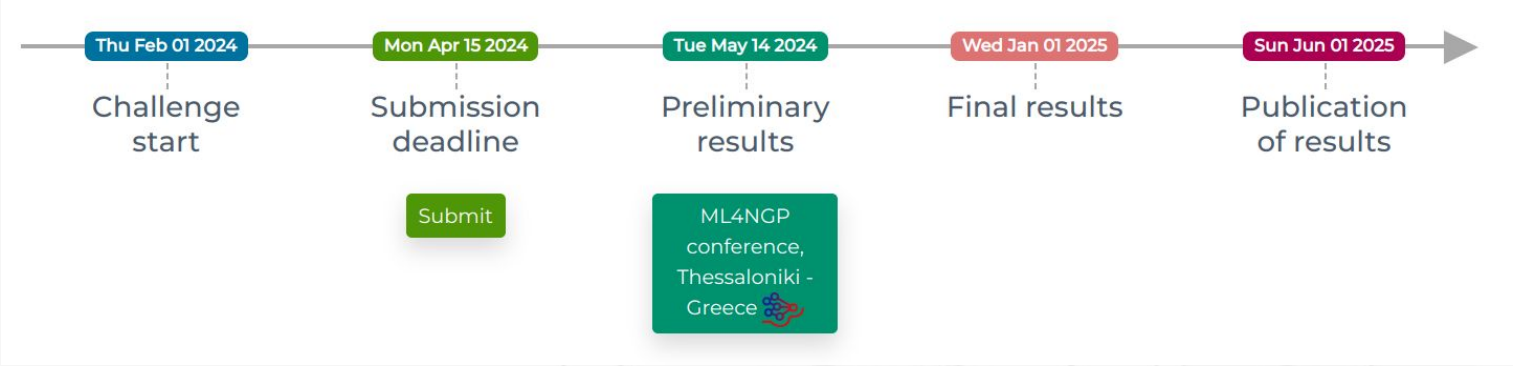

CAID Prediction Portal: A Comprehensive Service for Predicting Intrinsic Disorder and Binding Regions in Proteins
*Del Conte et al.* ***2023 NAR***

DisProt

MobiDB

https://caid.idpcentral.org

# CAID3 timeline




nature methods
ANALYSIS
https://doi.org/10.1038/s41592-021-01117-3
Check for updates

OPEN

## Critical assessment of protein intrinsic disorder prediction

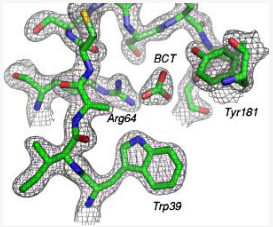
Marco Necci[1,50], Damiano Piovesan[1,50], CAID Predictors*, DisProt Curators* and Silvio C. E. Tosatto[1]

Received: 21 April 2023 | Revised: 22 June 2023 | Accepted: 8 August 2023
DOI: 10.1002/prot.26582
RESEARCH ARTICLE
PROTEINS WILEY

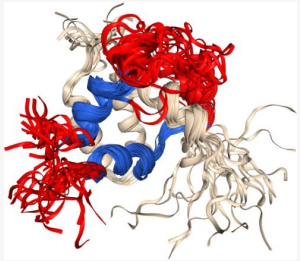
## Critical assessment of protein intrinsic disorder prediction (CAID) - Results of round 2

Alessio Del Conte[1] | Mahta Mehdiabadi[1] | Adel Bouhraoua[1] | Alexander Miguel Monzon[2] | Silvio C. E. Tosatto[1] | Damiano Piovesan[1]

# Ground truth




Missing residues



Mobile residues

Functional disorder
“Secondary methods”

**CASP5 - CASP10 (2002 - 2012)**

**CAID**

# Predictors - CAID3

- **70 predictors → 117 output (“flavours”)**
  - 46 already in CAID2
  - **24 new predictors**
    - 15 submitted as Docker containers (DockerHub)
    - Only 2 predictors had problems with installation (solved)

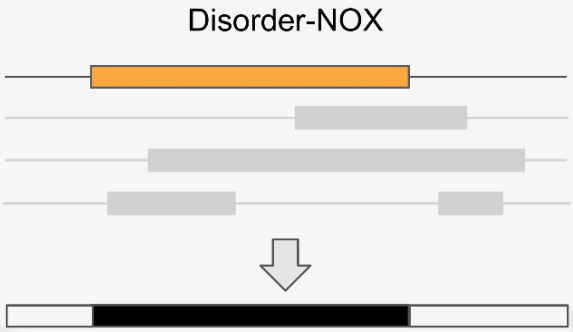
# Additional inputs

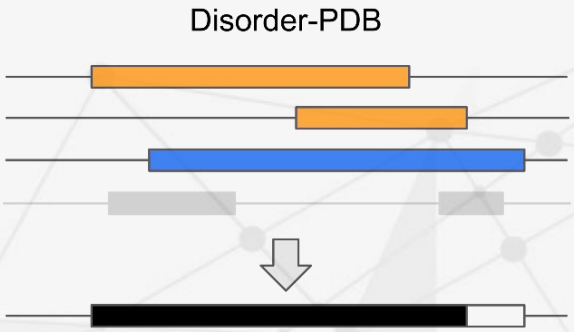
- **PSI-Blast** on Uniref90 → PSSM
- **MMseqs2** on UniRef30 → MSA
- **HHblits** on Uniclust30_2017_10 → HMM
- **SPIDER2** → SS, ASA, torsion angles
- **Embeddings**

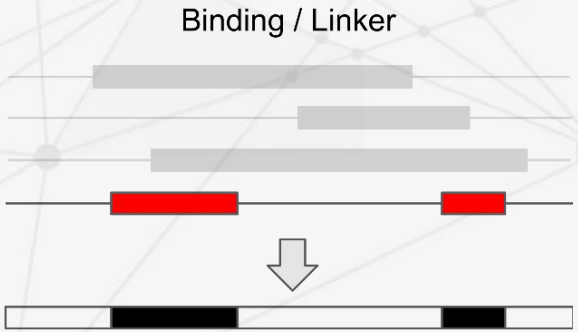
# Embeddings

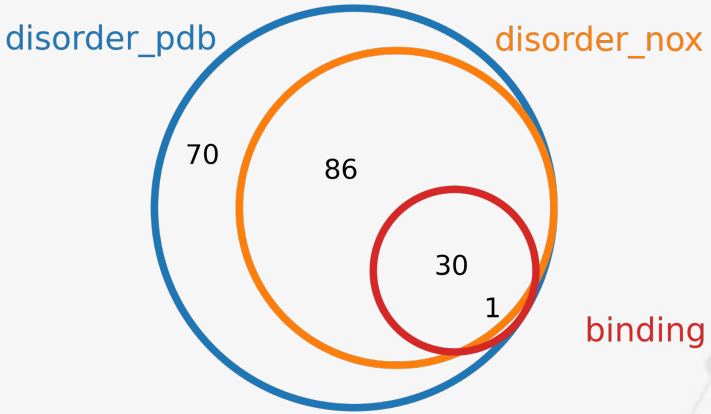
- Rost lab embeddings
  - **Prot-t5-xl-uniref50** used by 8 methods
  - **ProstT5** used by 1 method
- Facebook embeddings
  - **esm1b_t33_650M_UR50S** used by 1 method
  - **esm2_t33_650M_UR50D** used by 2 methods
  - **esm_msa1b_t12_100M_UR50S** used by 3 methods (embeddings from MSA)

# Ground truth

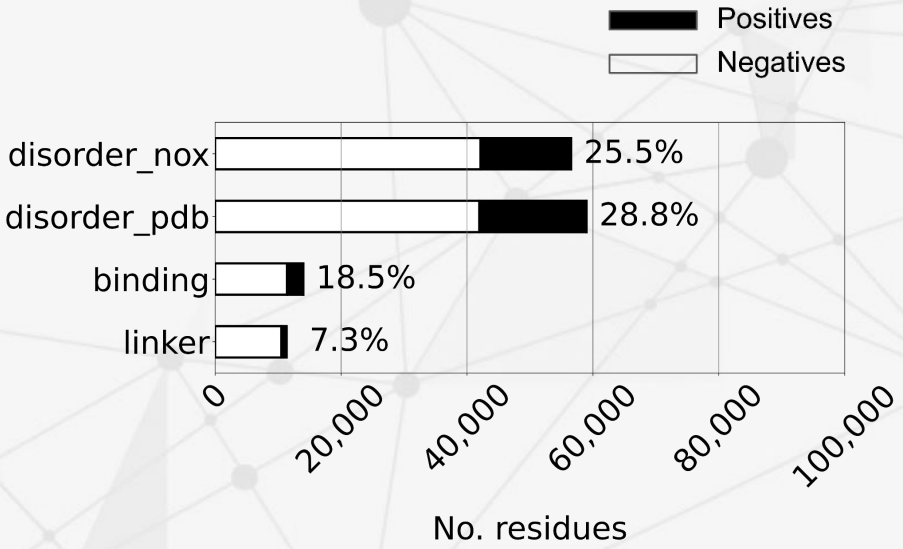




| | |
|---|---|
| Disorder-PDB | 187 |
| Disorder-NOX | 116 |
| Binding | 31 |
| Linkers | 17 |

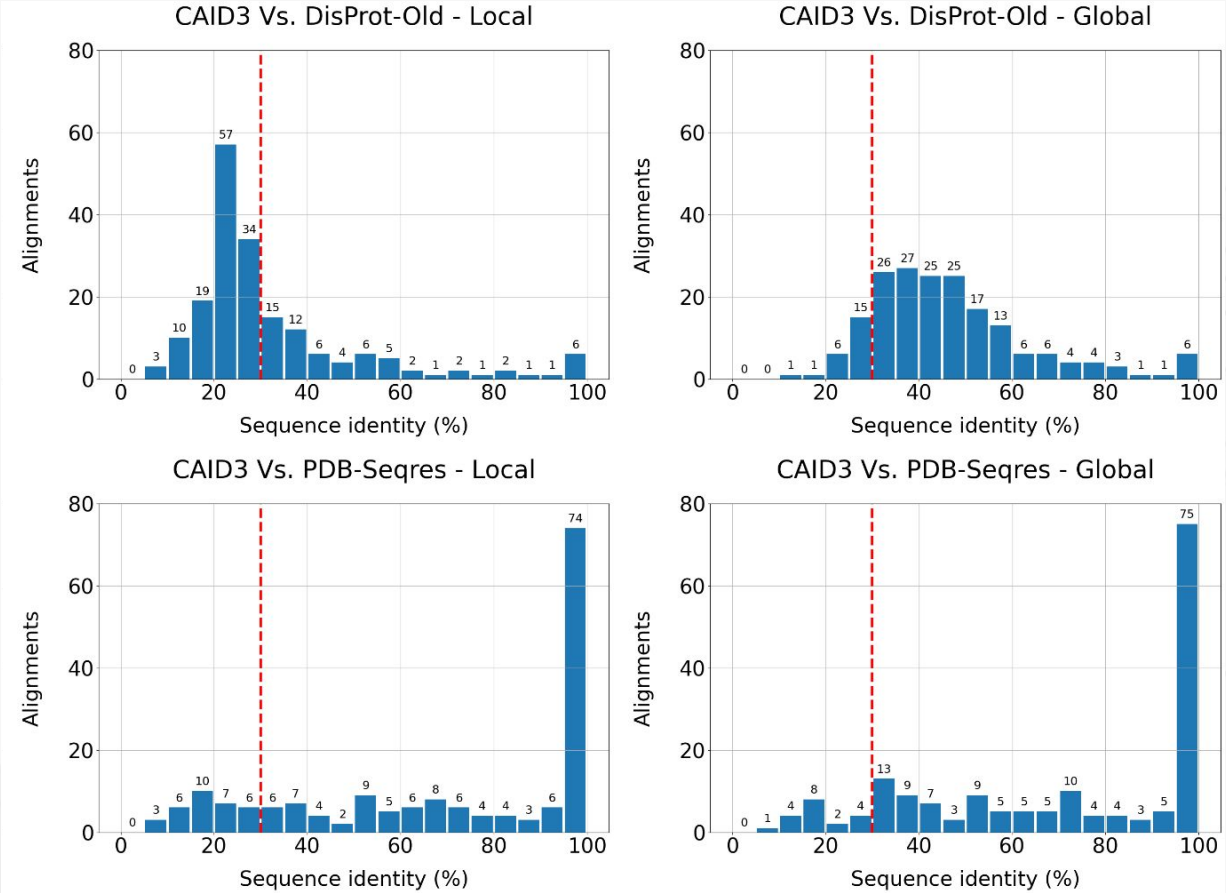
25 isoforms!



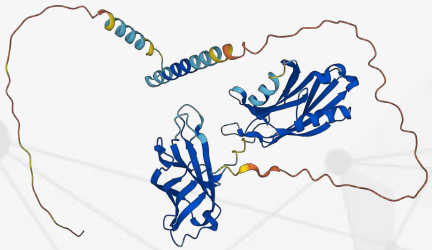
Smith–Waterman

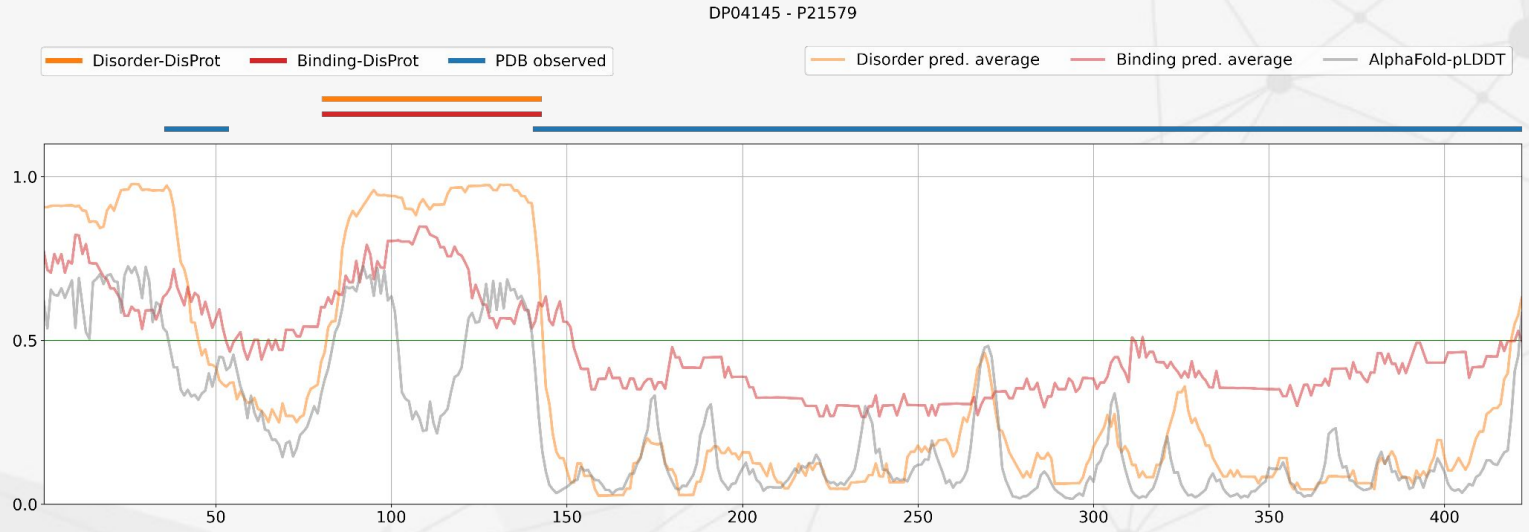
Needleman–Wunsch

DIPARTIMENTO MATEMATICA



# Examples






Synaptotagmin-1

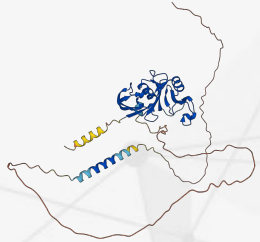
# Examples

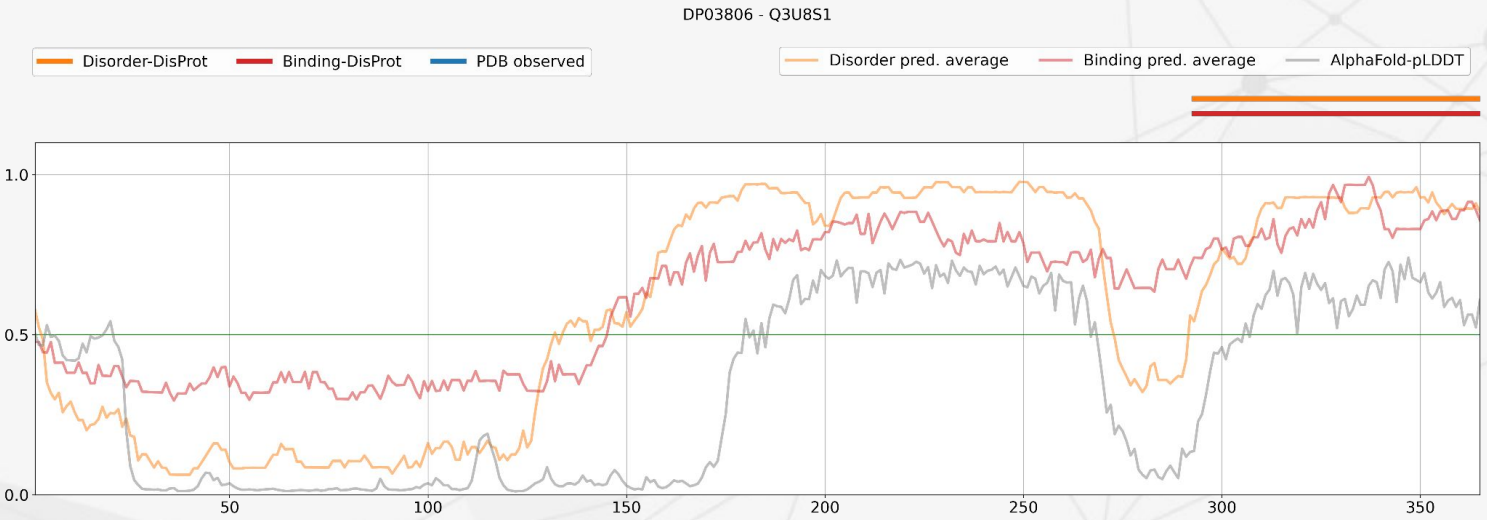
UniProtKB

**Function**[i]

**Caution**

Lacks conserved residue(s) required for the propagation of feature annotation. Automatic Annotation



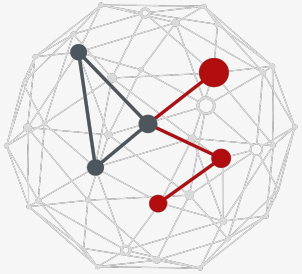



CD44 antigen

DIPARTIMENTO MATEMATICA

# Results

https://caid.idpcentral.org/



DATA SCIENCE
UNIVERSITY OF PADOVA

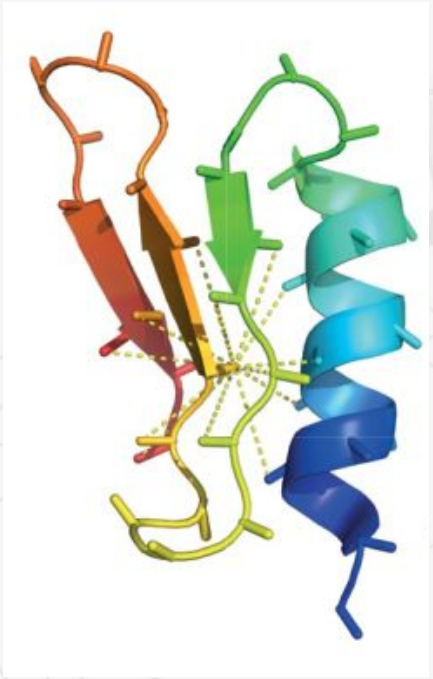
# IUPred

Master of Science in Data Science

Damiano Piovesan

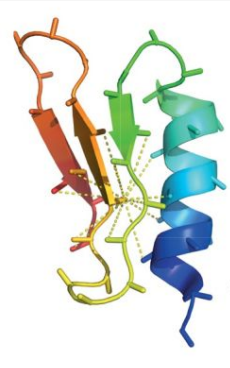
# Conformation energy

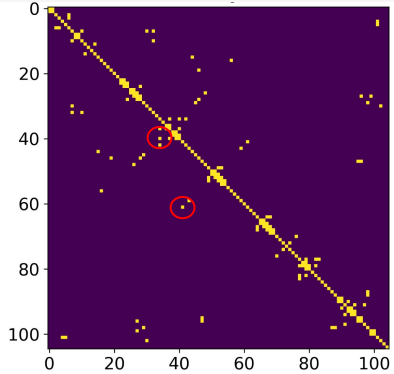
- **Native conformation** is at the **global energy minimum** in the conformational space
- **Interactions** are **simultaneously favored** in the native structure
- Energy is a function of the **conformation** as well as of the **amino acids sequence**
- **Total energy** → sum of **contacts contributions** weighted by the corresponding interaction energy



*The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins.* Dosztányi Z, Csizmók V, Tompa P, Simon I. **2005**. J Mol Biol.

# Real Vs Estimated energy

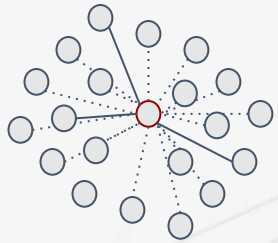


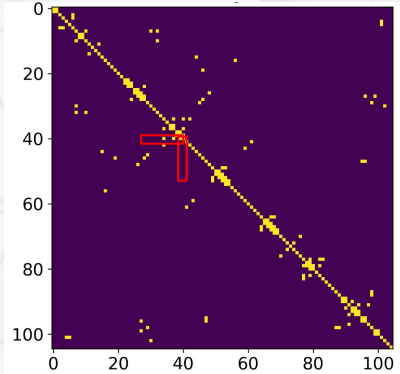



...MDVFMKGLSKAKE**G**VV**A**AA**E**KTKQGV**A**EAAGKTKEGVL**Y**VGSK...

Real contacts of position 40 (E, Glu)

- E - G
- E - A (x 2)
- E - Y






...MDVFMKGLSKAKE**G**VV**A**AA**E**KTKQGV**A**EAAGKTKEGVLYVGSK...

Virtual contacts of position 40 (E, Glu) in a window of 21 residues

- E - K (4 / 20)
- E - A (7 / 20)
- E - E (2 / 20)
- E - T (1 / 20)
- E - G (2 / 20)
- E - V (3 / 20)
- E - Q (1 / 20)

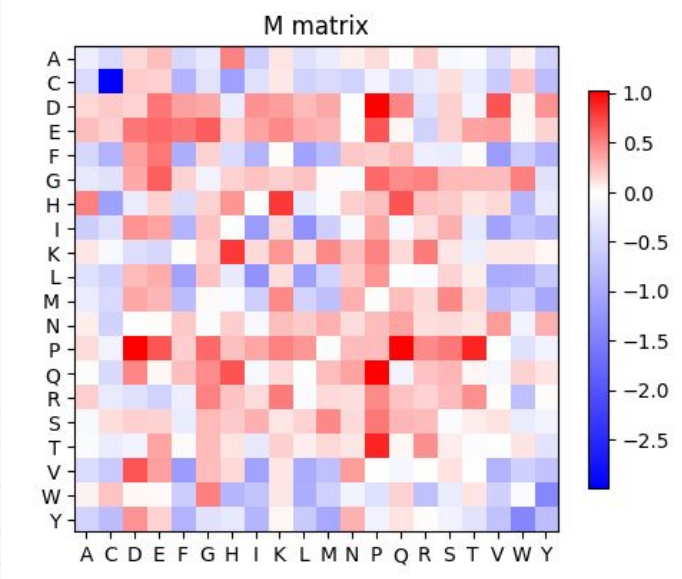
# IUPred - Contact energy

Propensity of amino acids pairs to make a favorable contact

- Energy depends only on amino acid types
- Contacts are assumed to be independent to each other (additive energy)
- Parameters are calculated from structural data → PDB

$$E = \sum_{ij=1}^{20} M_{ij} C_{ij}$$

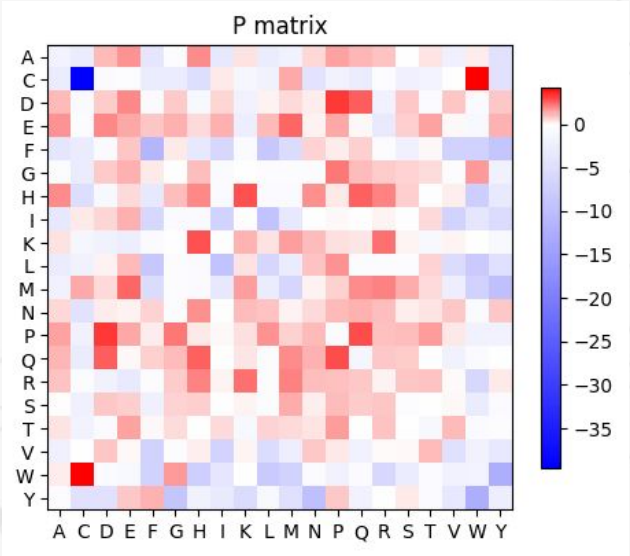
*i,j* → amino acids types
*M* → interaction energy matrix, 20 x 20
*C* → number of interactions of amino acid type i and j



# IUPred - Estimated energy

- The energy contribution of a residue depends on its potential partners in the sequence
- As if it would make a contact with all other neighbouring residues

$$e_i^k(\text{estimated}) = N_i^k \sum_{j=1}^{20} P_{ij} n_j^k$$




$N_i \rightarrow$ number of amino acid residues of type *i* in the sequence;
$n_i \rightarrow$ amino acid frequency $(N_i / L)$
$P \rightarrow$ energy predictor matrix

# IUPred - Training

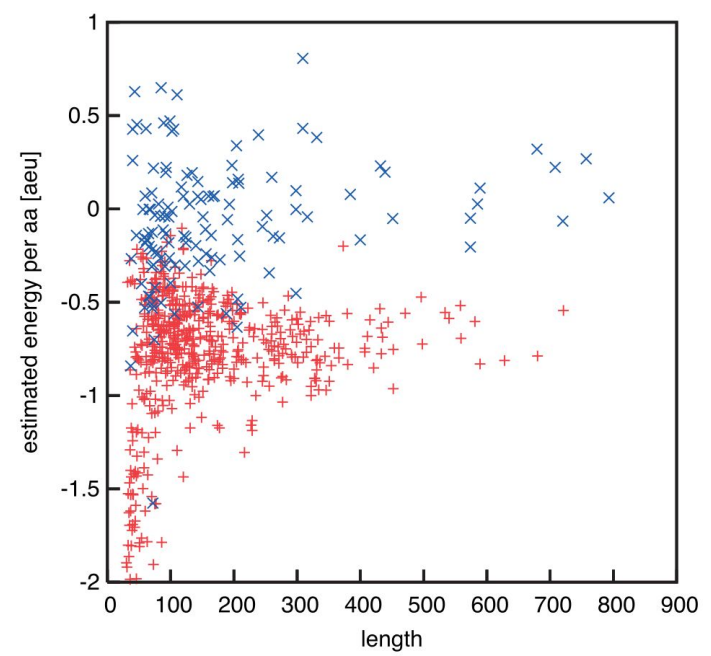
- The **energy predictor matrix (P)** is obtained minimizing the formula for each amino acid type (matrix row) with least-square fitting
- *i* → type of amino acid, *k* →protein target

$$Z_i = \sum_k (e_i^k - N_i^k \sum_{j=1}^{20} P_{ij} n_j^k)^2$$

$$e_i^k(\text{calculated}) = \sum_{j=1}^{20} M_{ij} C_{ij}^k$$

$$e_i^k(\text{estimated}) = N_i^k \sum_{j=1}^{20} P_{ij} n_j^k$$

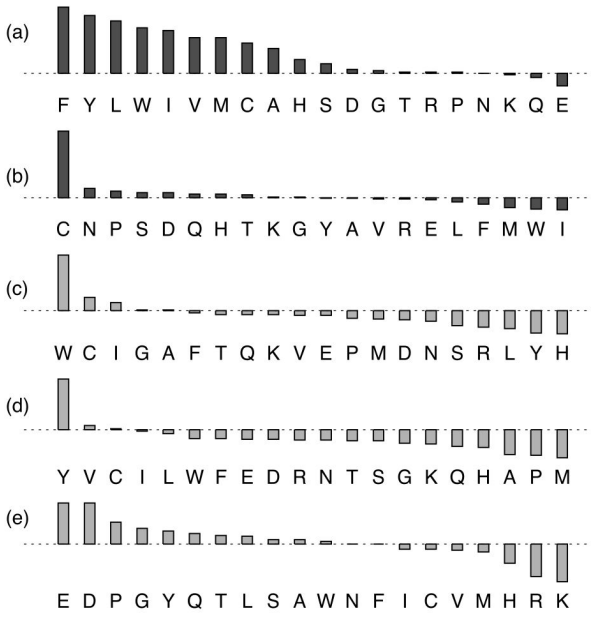
# IUPred - Validation




- Blue → IUPs (av. energy -0.07)
- Red → Globular (av. energy -0.81)

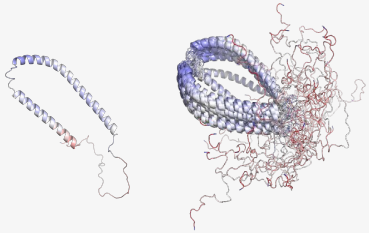
# IUPred - Contribution to the estimated energy

Eigenvectors of the P matrix

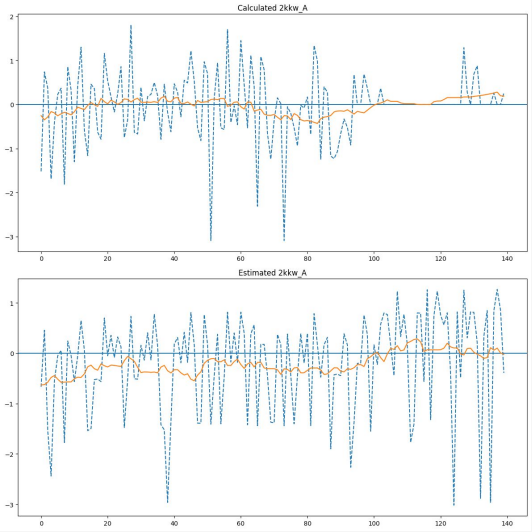



- Negative eigenvectors, **stabilizing**
  - (a) hydrophobicity
  - (b) cysteine content
- Positive eigenvectors, **destabilizing**
  - (c) ???
  - (d) Structure-breaking amino acids
  - (e) High net charge

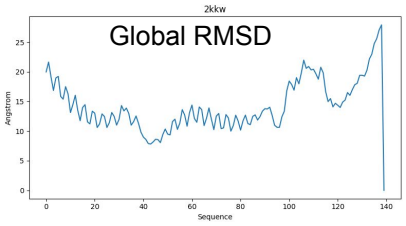
# 2KKW

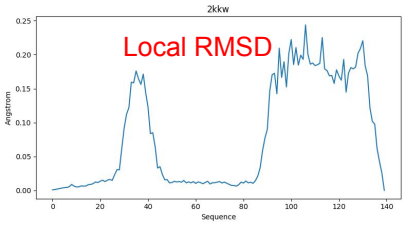


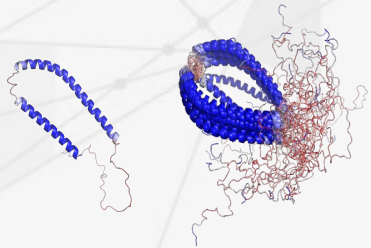
Contacts energy



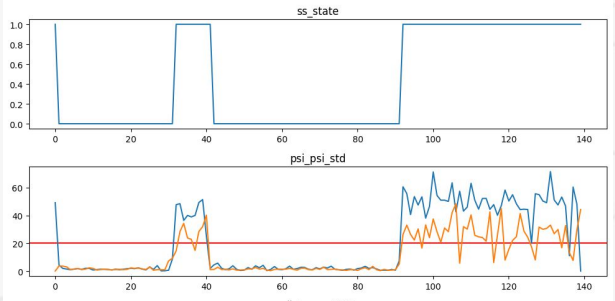

Estimated energy (IUPRED)



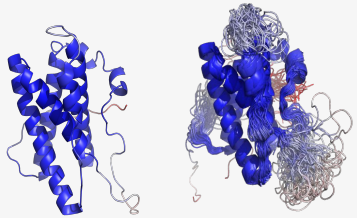
SS conservation



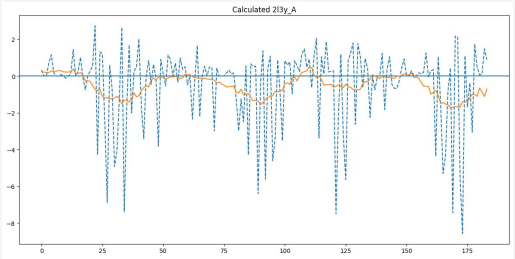

PHI / PSI

# 2L3Y

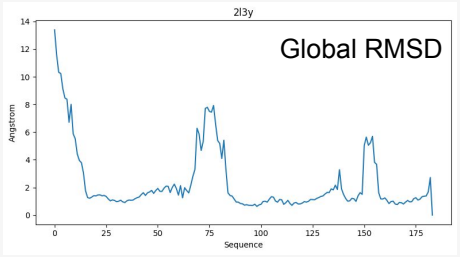
DIPARTIMENTO MATEMATICA

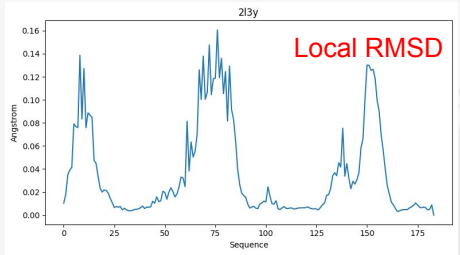


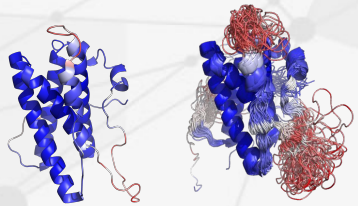
Contacts energy



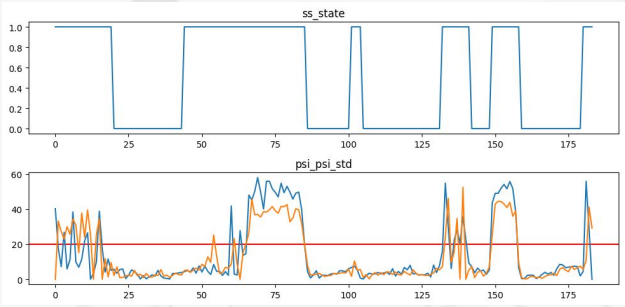

Estimated energy (IUPRED)



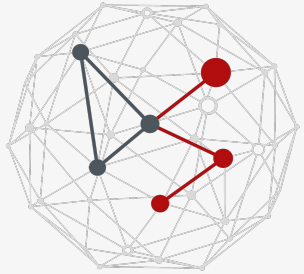



SS conservation




PHI / PSI


DIPARTIMENTO
MATEMATICA



DATA SCIENCE
UNIVERSITY OF PADOVA

# MobiDB

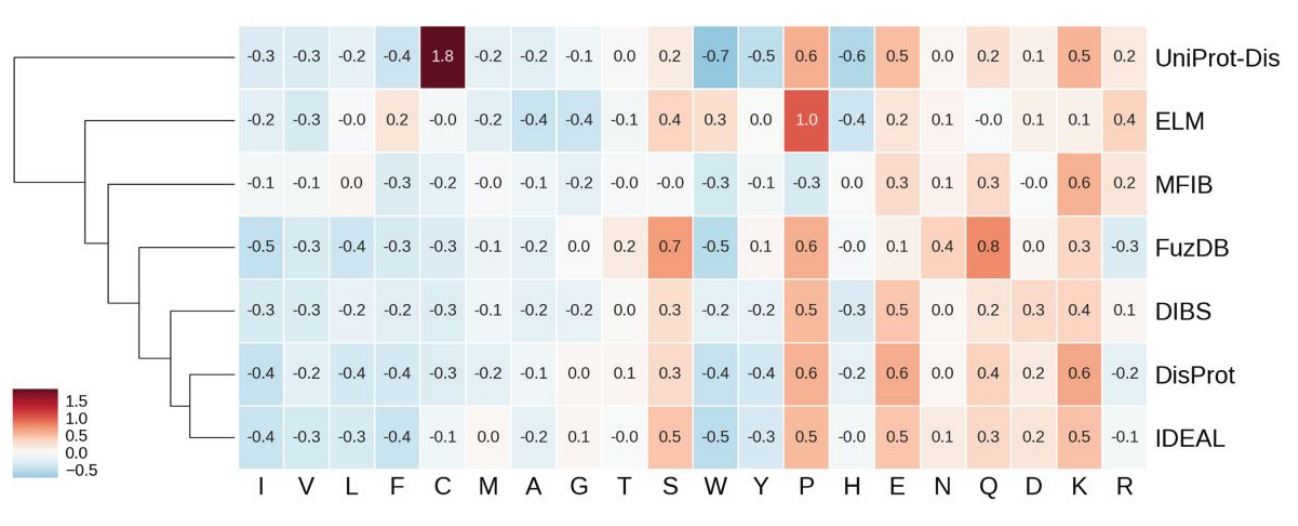
Master of Science in Data Science

Damiano Piovesan

# Amino acid composition in DBs

DIPARTIMENTO MATEMATICA

Fold increase (or decrease) compared to the whole UniProt (TrEMBL) amino acids frequencies




- Background TrEMBL
- Red fold increase
- Blue fold decrease

*Necci, Piovesan, Tosatto. 2018 Database*

# IDP ecosystem

DIPARTIMENTO MATEMATICA

Databases

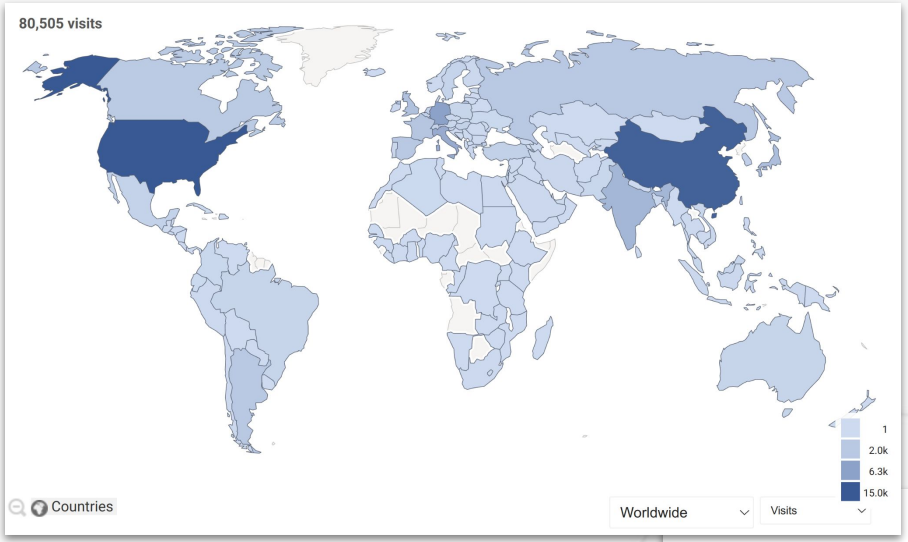
Organizations

Projects

Communities

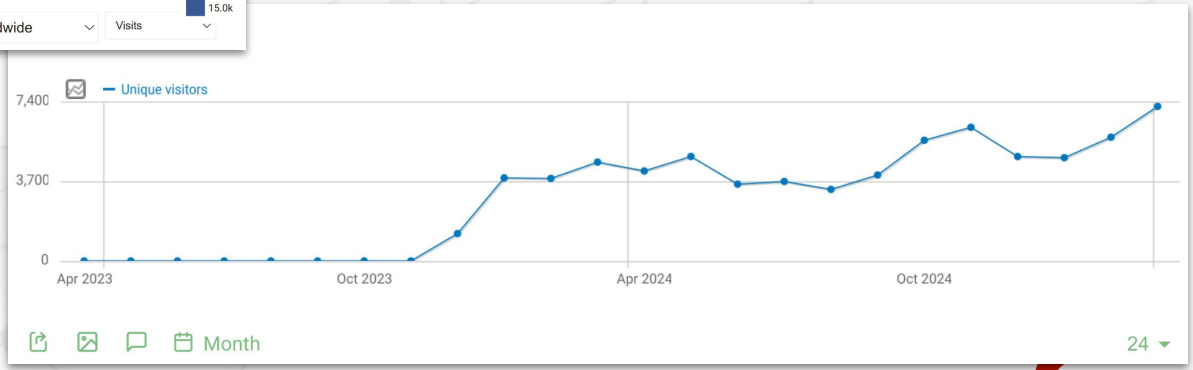
# MobiDB visits 2024

DIPARTIMENTO MATEMATICA



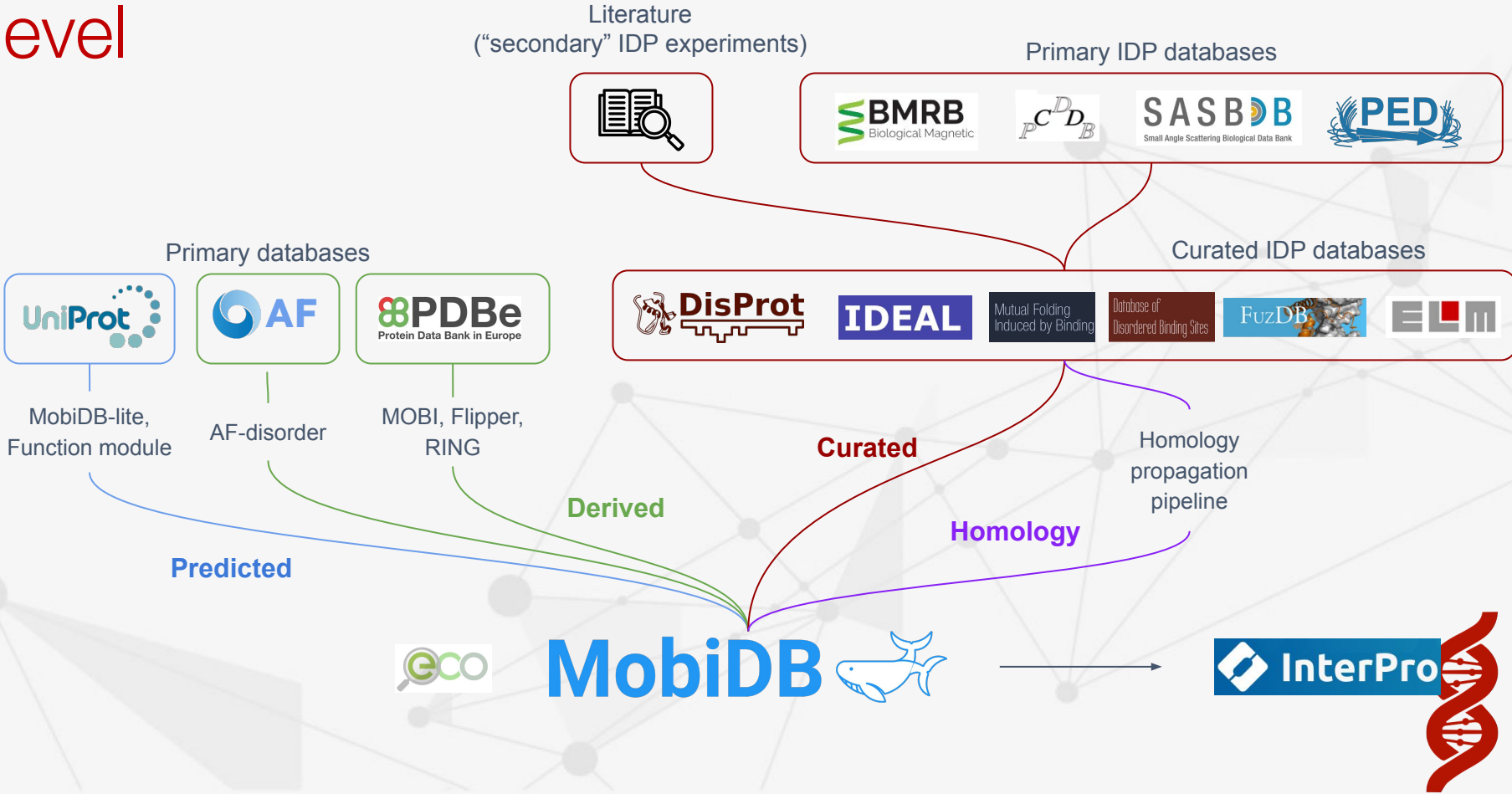

MobiDB

MOBIDB in 2025: integrating ensemble properties and function annotations for intrinsically disordered proteins
*Piovesan, Del Conte, Mehdiabadi, Aspromonte, Blum, Tesei, von Bülow, Lindorff-Larsen, Tosatto*
**Nucleic Acids Research, 2025**

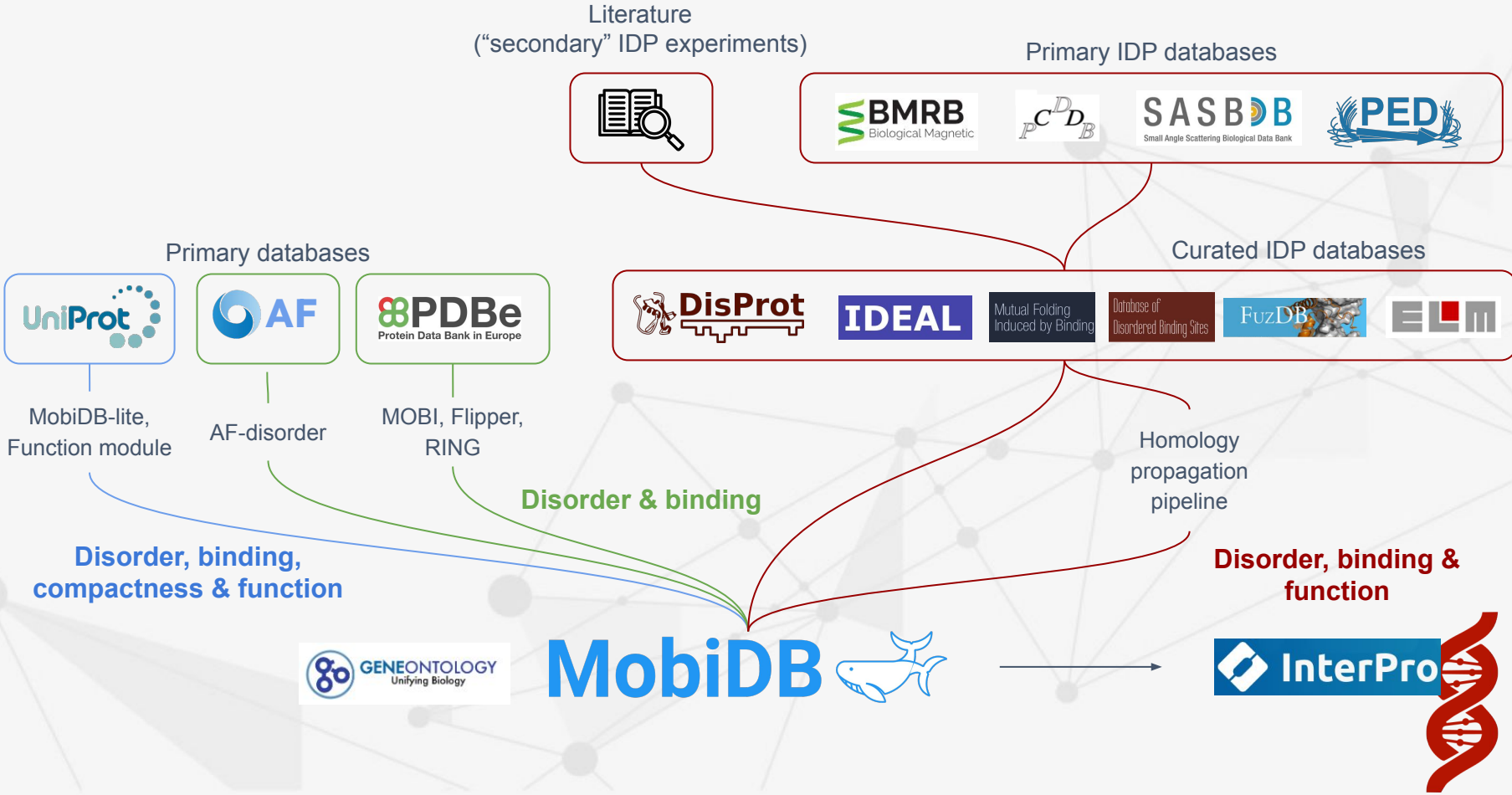



matomo

# Evidence level



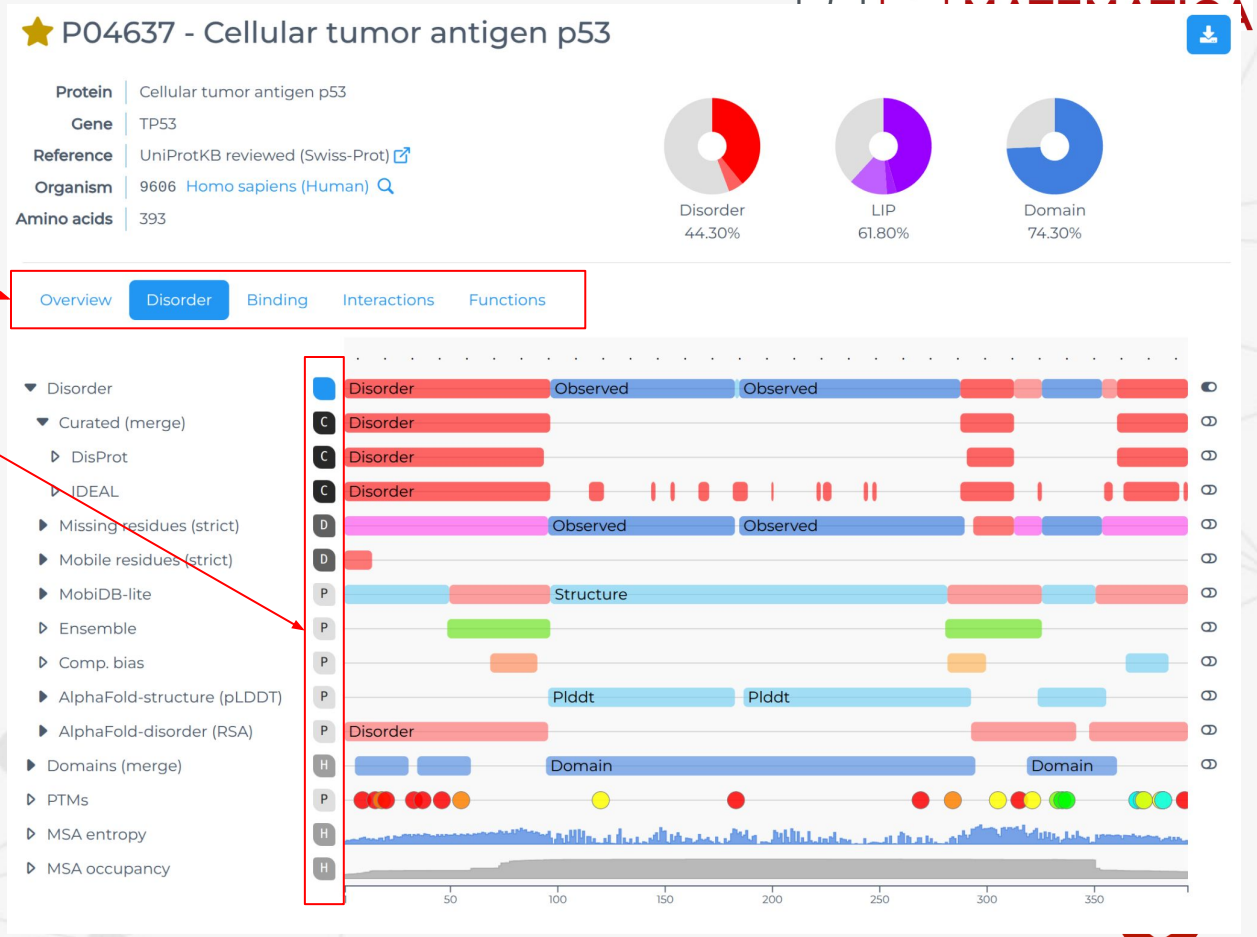
# Evidence type



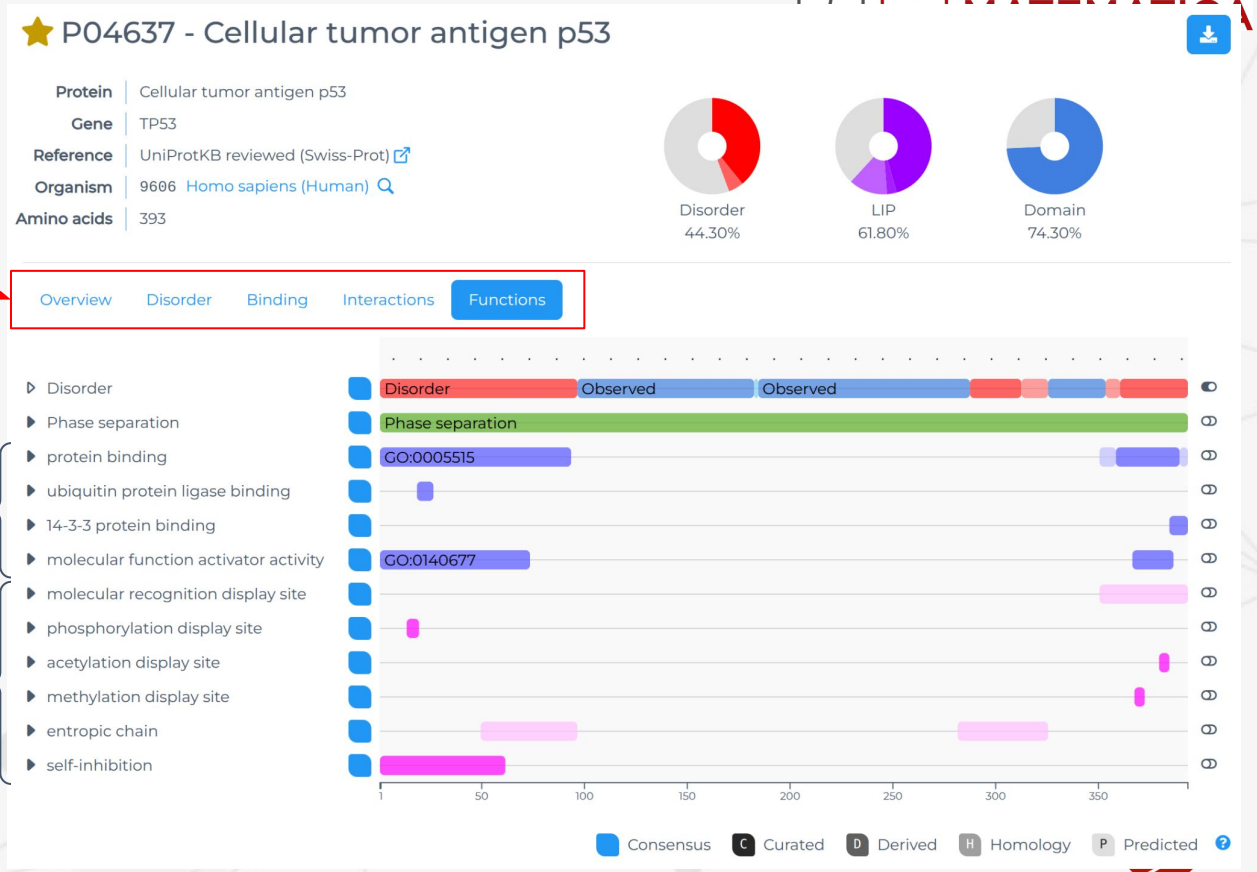
# Evidence level

Evidence “type”

Evidence “level”



# ID function




Evidence “type”

Molecular function

Disorder function

# ID function - Training

DIPARTIMENTO MATEMATICA

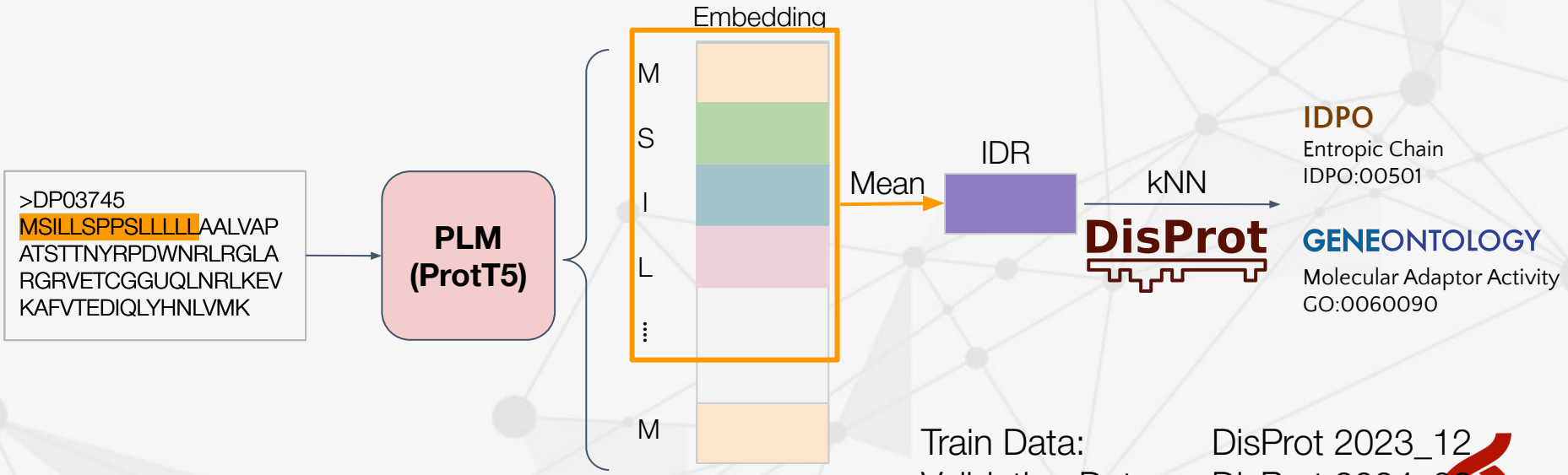
DisProt

| Ontology | Aspect | Terms | Proteins | Regions |
|---|---|---|---|---|
| GO | Biological process | 415 | 157 | 320 |
| GO | Cellular component | 36 | 17 | 37 |
| GO | Molecular function | 165 | 980 | 2879 |
| IDPO | Disorder function | 20 | 491 | 735 |

# Function Prediction for IDRs

DIPARTIMENTO MATEMATICA

Idea: “Close IDRs in the embedding space have Similar functions. ”




Train Data: DisProt 2023_12

Validation Data: DisProt 2024_06

# MobiDB Statistics

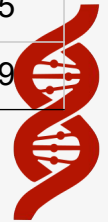
DIPARTIMENTO MATEMATICA

| Namespace | Term | Term name | Curated | Homology | Prediction |
|---|---|---|---|---|---|
| **Disorder Function (IDPO)** | IDPO:00502 | flexible linker/spacer | 317 | 31,714 | 23,200,290 |
| | IDPO:00501 | entropic chain | 24 | 1,603 | 15,123,574 |
| | IDPO:00504 | flexible C-terminal tail | 30 | 2,700 | 3,967,252 |
| | IDPO:00503 | flexible N-terminal tail | 26 | 2,363 | 3,634,519 |
| | IDPO:00025 | phosphorylation display site | 122 | 5,104 | 3,191,872 |
| | IDPO:00024 | molecular recognition display site | 13 | 1,239 | 3,023,196 |
| | IDPO:00505 | self-regulatory activity | 4 | 542 | 2,344,651 |
| | IDPO:00506 | self-inhibition | 24 | 1,620 | 1,053,900 |
| | IDPO:00026 | acetylation display site | 13 | 199 | 223,003 |
| | IDPO:00508 | self-assembly | 24 | 1,234 | 156,345 |

# MobiDB Statistics

DIPARTIMENTO MATEMATICA

| Namespace | Term | Term name | Curated | Homology | Prediction |
|---|---|---|---|---|---|
| Molecular Function (GO) | GO:0005515 | protein binding | 642 | 45,318 | 7,259,030 |
| | GO:0003676 | nucleic acid binding | 43 | 3,862 | 3,255,313 |
| | GO:0060090 | molecular adaptor activity | 151 | 10,959 | 2,725,151 |
| | GO:0098772 | molecular function regulator activity | 112 | 8,314 | 2,280,943 |
| | GO:0005488 | binding | 0 | 0 | 1,198,483 |
| | GO:0140693 | molecular condensate scaffold activity | 41 | 2,570 | 570,237 |
| | GO:0003723 | RNA binding | 40 | 1,906 | 466,434 |
| | GO:0003677 | DNA binding | 41 | 2,713 | 285,069 |
| | GO:0036094 | small molecule binding | 35 | 3,027 | 195,155 |
| | GO:0001069 | regulatory region RNA binding | 6 | 489 | 182,389 |



# DisProt IDR Encodings

DIPARTIMENTO MATEMATICA

**DP00677r003** Disorder function — ↑ Show on Feature-Viewer

**Term** flexible C-terminal tail, IDPO:00504 **Fragment** 1 - 42

**Evidence** analytical ultracentrifugation evidence used in manual assertion, ECO:0006275

**Reference** Crystal structure and assembly of a eukaryotic small heat shock protein. *van Montfort RL, Basha E, Friedrich KL, Slingsby C, Vierling E.* Nat Struct Biol, 2001, pmid:11702068

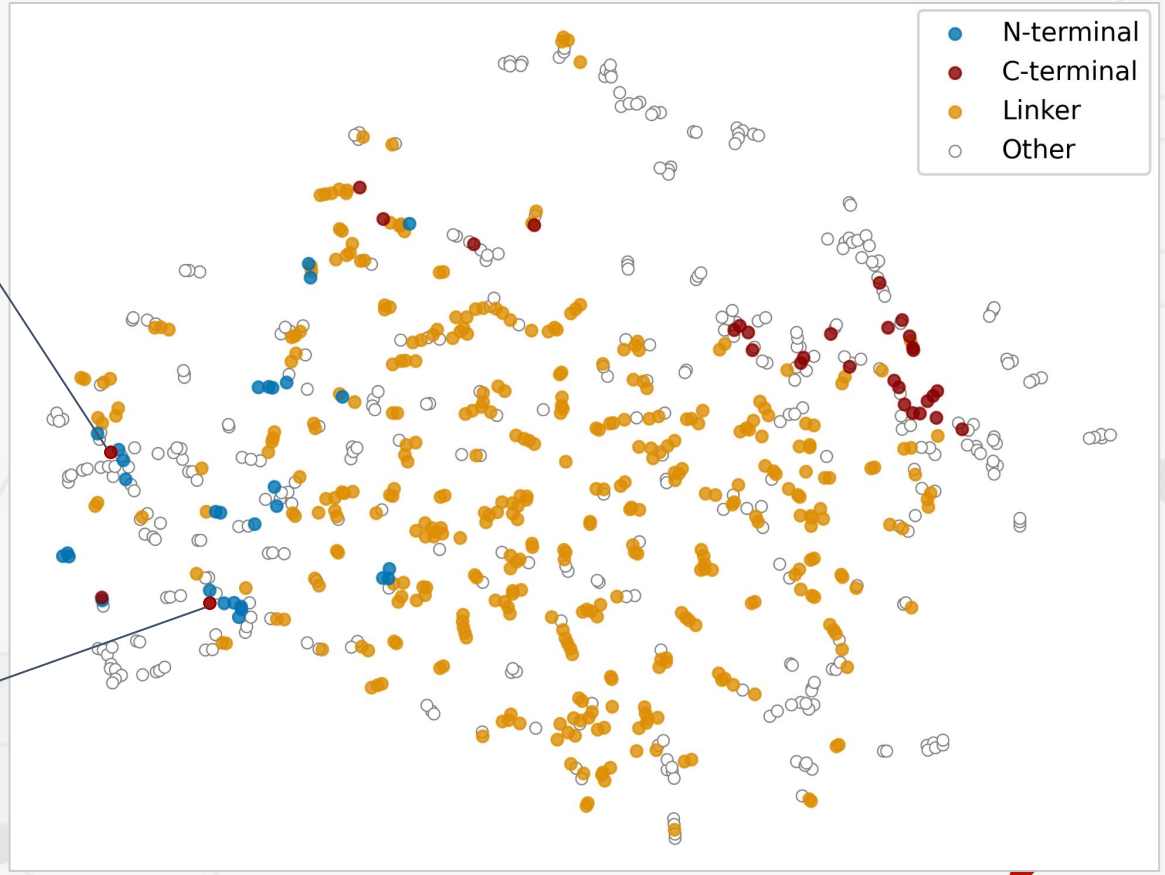
**DP00051r010** Disorder function — ↑ Show on Feature-Viewer

**Term** flexible C-terminal tail, IDPO:00504 **Fragment** 1 - 32

**Evidence** X-ray crystallography-based structural model with missing residue coordinates used in manual assertion, ECO:0006220

**Cross references** PDB:2TOD PDB:1QU4

**Reference** X-ray structure of ornithine decarboxylase from Trypanosoma brucei: the native structure and the structure in complex with alpha-difluoromethylornithine. *Grishin NV, Osterman AL, Brooks HB, Phillips MA, Goldsmith EJ.* Biochemistry, 1999, pmid:10563800

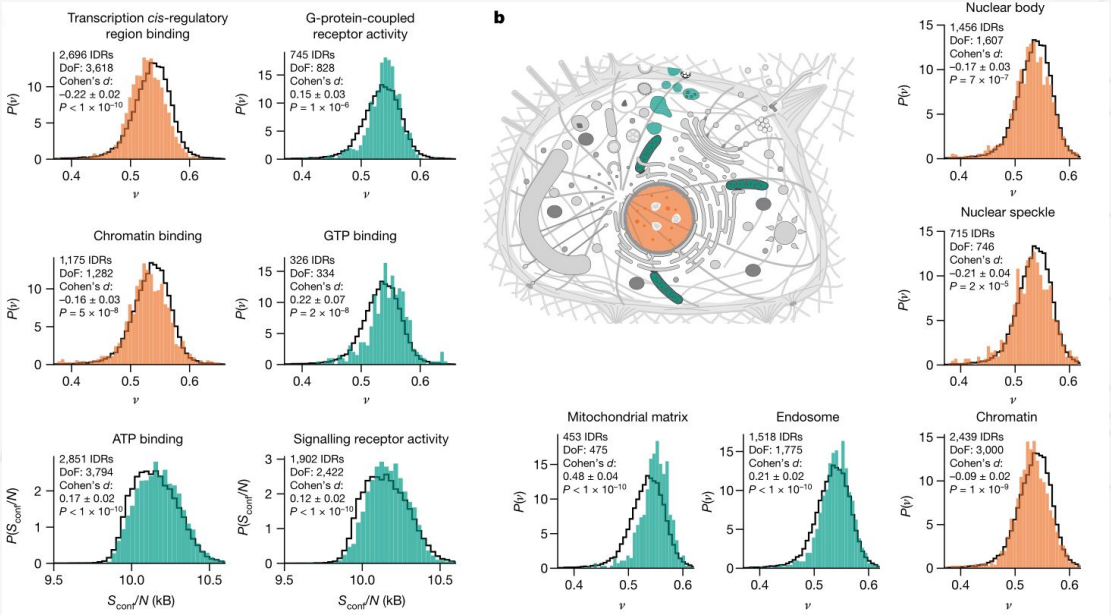


# New features - Compactness

DIPARTIMENTO MATEMATICA

Introduction of “**compact**” and “**extended**” region labels

Calculated only for disordered regions with **at least 30 residues**




Conformational ensembles of the human intrinsically disordered proteome
*Tesei, Trolle, Jonsson et al.*
**Nature, 2024**