# Predicting molecular recognition features in protein sequences with MoRFchibi 2.0

Nawar Malhis[1], Jörg Gsponer[1, 2]

[1]: Michael Smith Laboratories, University of British Columbia, Vancouver, BC V6T 1Z4, Canada

[2]: Department of Biochemistry and Molecular Biology, University of British Columbia, Vancouver, BC V6T 1Z3, Canada

## Abstract

Identifying sites within intrinsically disordered regions (IDR) that bind to other proteins remains a significant challenge. Molecular Recognition Features (MoRFs) are a subset of segments in IDR that bind to other proteins, undergoing a disorder-to-order transition upon binding. This paper introduces MoRFchibi 2.0, a specialized prediction tool designed to identify the locations of MoRFs within protein sequences. Our results show that MoRFchibi 2.0 outperforms all existing MoRF and general predictors of protein-binding sites within IDRs, including top-performing models from CAID rounds 1, 2, and 3. Remarkably, MoRFchibi 2.0 surpasses predictors that utilize AlphaFold data and state-of-the-art protein language models, achieving superior ROC and Precision-Recall curves and higher success rates. MoRFchibi 2.0 generates output scores using an ensemble of convolutional neural network logistic regression models, followed by a reverse Bayes Rule to adjust for priors in the training data. These scores reflect MoRF probabilities normalized for the priors in the training data, making them individually interpretable and compatible with other tools utilizing the same scoring framework.

**Availability:** https://mc2.msl.ubc.ca/index.xhtml

## 1 Introduction

Intrinsically disordered regions (IDRs) in proteins are segments that lack a defined tertiary structure under physiological conditions [1]. Despite this, IDRs are functionally significant and often play key roles in cellular processes by interacting with other proteins, DNA, RNA, or small molecules [2-5]. Their structural flexibility and dynamic nature allow IDRs to adapt and mold themselves around their interaction partners.

As of the latest release of the DisProt database [6] in June 2024, 641 out of 3,022 proteins annotated with IDRs include one or more experimentally validated binding segments. However, this is a small number, especially considering that approximately 32% of the human proteome consists of proteins with more than 30% disordered residues [7]. Consequently, the computational identification of binding regions within IDRs has emerged as a significant challenge. Of particular interest are regions within IDRs that bind to other proteins, and numerous computational tools have been developed to aid in identifying these protein-binding segments [8-44].

A subfamily of protein-binding segments in IDRs is known as molecular recognition features (MoRFs) [45-47]. These segments transition from a disordered to an ordered state upon binding to their interaction partners. Typically, MoRFs are short, though they can sometimes span 50 or more residues. Many MoRF-mediated interactions are characterized by low affinity but high specificity, making them well-suited for reversible regulatory interactions in signaling pathways. The functional significance of MoRFs, along with the challenges posed by their experimental identification and characterization, has led to the development of several computational tools designed to predict the locations of MoRFs in protein

sequences. Notable examples include Retro-MoRFs [11], MoRFpred [16], fMoRFpred [21], MoRFchibi_light [23], MoRFchibi_web [19], OPAL [29], MoRFPred-plus [27], and MoRFcnn [42]. Since MoRFs represent only a small fraction of all residues in protein sequences, enhancing the quality of computational approaches is essential to minimize false positives and improve prediction accuracy.

MoRFchibi predictors, light and web, are among the most accurate MoRF predictors available today. They are ranked among the leading IDR-protein binding predictors in the first two rounds of the CAID competition [48, 49] and are available among the CAID Portal set of predictors [50]. MoRFchibi light predictions are included in the DescribePROT database [51], discussed in many review articles [52 .. 56], and used in many applications [57-60]. However, the accuracy of predictions depends heavily on the quality and size of the training datasets, and MoRFchibi predictors were developed using training data collected in 2008 [16]. Since then, several databases have been created that contain curated and annotated IDR protein-binding segments, including MoRF segments. The availability of more extensive and accurate training data enabled us to redesign MoRFchibi and increase the quality of its prediction. Here, we introduce this redesigned MoRF predictor, MoRFchibi 2.0, which encompasses an ensemble of four convolutional neural network models. We also provide newly assembled datasets with MoRF annotations for the community to optimize and test other models.

## 2 Methods

### 2.1 Datasets

We assembled a dataset of protein sequences annotated with disordered regions that undergo disorder-to-order transitions upon binding to other proteins, drawing from four manually curated databases: DisProt, DIBS [61], MFIB [62], and IDEAL [63]. Two key factors that impact the quality of MoRF annotations are the availability of evidence supporting the disorder of a sequence segment in isolation and proof of folding induced by binding to a protein target. The extent to which this critical information is covered varies across the four databases.

DisProt, a database for disordered protein sequences, provides disorder annotations and details on IDR properties, such as protein and nucleotide binding and disorder-to-order (DtO) transitions. DIBS and IDEAL databases focus on interactions between binding segments in IDRs and their ordered protein partners. In these databases, the disordered state of binding segments is supported either by experimental data or inferred from homology. While all interacting IDR segments in IDEAL undergo a disorder-to-order transition upon binding, a small fraction of these segments in DIBS do not fold upon binding. The MFIB database, on the other hand, includes IDRs that bind to other IDRs and fold. The majority of interactions in MFIB are backed by experimental evidence, with some based on homology.

Using the latest DisProt release (24_06), we annotated all disordered protein-binding residues that overlap with DtO transitions as MoRF residues. Since DtO annotations are likely incomplete, we masked out annotated protein-binding residues not associated with DtO, meaning we considered them neither MoRFs nor non-MoRFs. We included IDEAL and DIBS binding segments that were validated through experimental data and restricted DIBS segments to those associated with DtO annotations in DisProt. Additionally, we incorporated all MFIB binding annotations. This process resulted in 899 high-quality MoRF segments (HQ MoRFs) across 769 sequences, encompassing 46,694 residues. Alongside these HQ MoRF annotations, we included low-quality MoRF annotations (LQ MoRFs) for training. These LQ MoRFs

are binding segments from IDEAL and DIBS, supported by evidence from homology and DIBS binding segments not corroborated by DisProt DtO annotations. This approach provided 1,217 MoRF segments (HQ & LQ) across 1,030 sequences, amounting to 52,560 residues.

We used CD-Hit to cluster these sequences at 80% identity. Within each cluster, MoRF annotations for sequences with >95% identity to cluster centers or >92% identity and equal length to cluster centers are transferred to cluster centers as LQ annotations, i.e., to be used for training only. Then, we only used cluster centers, resulting in 907 sequences with 1,081 MoRFs totaling 46,504 HQ & LQ MoRF residues, including 674 sequences with 790 HQ MoRF segments totaling 40,657 MoRF residues. As sequences sourced from the DisProt database are likely to be more completely annotated than those from IDEAL, DIBS, and MFIB, the non-MoRF residues in DisProt sequences are less likely to include false negatives. Thus, we masked out non-MoRF residues in sequences not sourced from DisProt.

We divided the total 907 MoRF sequences into five subsets such that sequences in each subset are less than 40% identity to those in other subsets. For that, we clustered the sequences at 40% identity using CD-Hit. We sorted these clusters in descending order based on each cluster's total number of HQ MoRF residues. Then, starting with five empty subsets, we iteratively moved clusters to these subsets such that at each iteration, we moved the cluster with the highest number of HQ MoRF residues to the subset with the least total HQ MoRF residues.

We used four subsets containing both LQ & HQ MoRF annotations to train four cross-validated CNN models, and the remaining fifth subset for testing (Test-151). Although sequences in the test dataset are less than 40% identical to those used in the training, short test regions can still have local homology to training sequences. Thus, we used HAM [64] to identify and mask out such local homologs. HAM identified four regions (totaling 44 residues) in the testing dataset as homologs of the training data, one of which is a homolog of a masked training region of 11 residues. Therefore, HAM only masked the remaining 33 residues. Thus, the test set (Test-151) contains 151 sequences with 156 MoRFs, totaling 8,082 MoRF residues. For completeness, we also created Test-151 (unmasked), for which all previously masked residues, except those masked by HAM, were unmasked. In addition, we created the Test-151 (AF2) subset, which is limited to sequences with AlphaFold2 structures in the AlphaFoldDB database [65]. We also created the Test-151 (SR) subset, which only contains sequences with both MoRF and non-MoRF residues. Specifically, we excluded all sequences from Test-151 that do not have at least five residues for each class, i.e., all sequences must have 5+ MoRF residues and 5+ non-MoRF residues. Moreover, as SR is not sensitive to under-annotated sequences and in order not to reduce the test data size significantly, we unmasked non-MoRF outside DisProt sequences. **Table 1** describes the details of all four training datasets, the Test-151 dataset, and the three variations of the Test-151 dataset.

**Table 1:** Training and testing datasets.

| | MoRFs quality | Sequences with MoRFs/Total | MoRFs | Non-MoRF Residues | MoRF Residues | Masked Residues |
|---|---|---|---|---|---|---|
| **Training (CV1)** | LQ & HQ | 141/141 | 172 | 28,812 | 8,902 | 47,440 |
| **Training (CV2)** | LQ & HQ | 182/182 | 208 | 57,212 | 9,267 | 43,553 |
| **Training (CV3)** | LQ & HQ | 192/192 | 232 | 51,231 | 9,347 | 45,729 |
| **Training (CV4)** | LQ & HQ | 198/198 | 237 | 52,908 | 9,532 | 58,255 |
| **Test-151** | HQ | 133/151 | 156 | 49,103 | 8,082 | 21,057 |
| **Test-151 (unmasked)** | HQ | 133/151 | 156 | 70,127 | 8,082 | 33 |
| **Test-151 (AF2)** | HQ | 119/135 | 139 | 36,097 | 7,556 | 17,403 |
| **Test-151 (SR)** | HQ | 112/112 | 135 | 52,603 | 6,066 | 1,240 |

### 2.2 Input Features

We used the seven features introduced in [43] to encode amino acid sequences. The first two features estimate the energy associated with intrachain interactions of each residue by using the previously derived 20 x 20 energy predictor matrix [66]. Specifically, the first interaction energy feature is the sum of the pairwise energies between the residue at k and all residues within a window w1=10 on either side of k. The second interaction energy feature is the sum of this energy for all residue pairs between w1 and w2=25 on either side of k. The additional five features were obtained from the **F5** amino acid enrichment matrix. **F5** is a 20 x 5 matrix encoding the enrichment of each amino acid within IDRs, linker regions, IDR nucleic binding sites, IDR protein binding sites, and folded domains, as described in [43].

### 2.3 Model selection and optimization

We used the ensemble window-in/window-out structure utilizing four logistic regression convolutional neural network (CNN) models introduced in [43]. Each model comprises three convolutional layers with kernel size seven, separated by two average pooling layers with kernel size seven, followed by two fully connected (FC) layers with the output size of the first FC layer equal to 150 (**Figure 1**). An input sliding window size ten is padded by 250 residues on each side, totaling 510 input residues. We trained four CNN models, each on three CV subsets, and validated them on the fourth ( see **Table 1** for details**)**. We used a binary cross entropy with a sigmoid (BCEWithLogitsLoss) as training LOSS and applied a sigmoid function to the evaluation output of each CNN model. We applied reverse Bayes to each model output to factor out the training priors, i.e., the imbalances in the percentages of MoRF residues between the training data subsets. Then, we averaged the four scores and used a smoothing window of equal weights size 5 to smooth the final output score. We trained 15 instances for each of the four CNN models, each minimizing validation loss, and then, to avoid overfitting the validation subsets, we selected the instance with the third-best validation AUC.

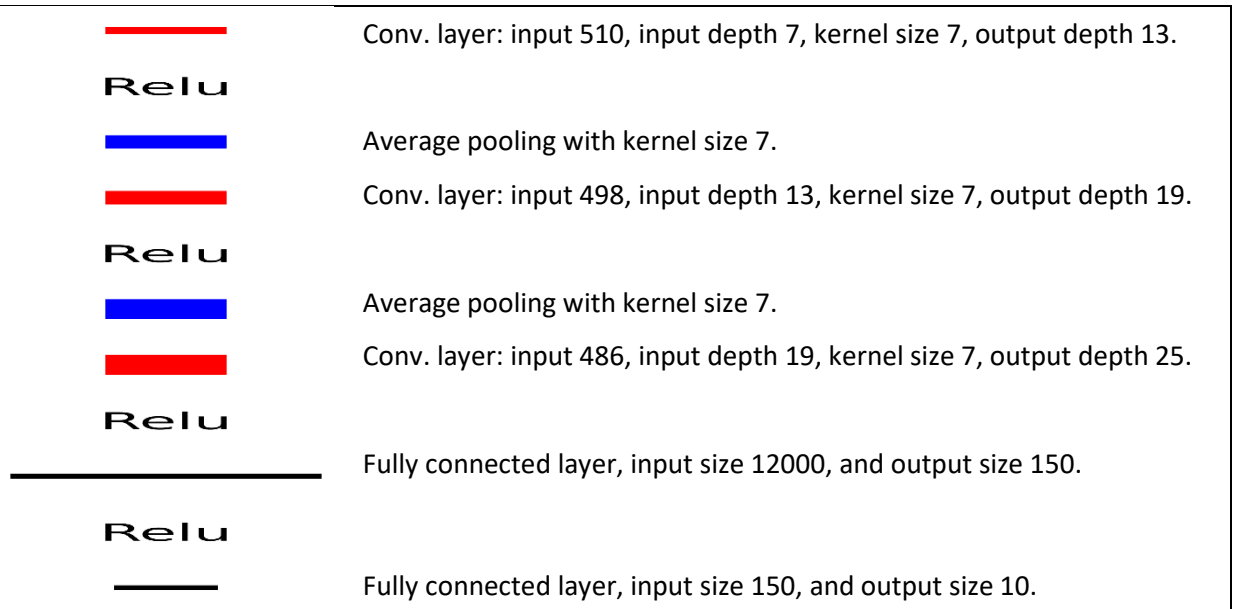


**Figure 1:** The overall architecture of the CNN is used with convolutional kernels in red, average pooling kernels in blue, and fully connected layers in black.

### 2.4 Evaluation measures

We calculated Receiver Operating Characteristic (**ROC)** curves to evaluate the performance of tools independent of the binary threshold cut-off values. ROC curves chart the relationship between the percentage of the positive class passing any score (true-positive rate, TPR) and the percentage of the negative class passing that score (false-positive rate, FPR). The formal definitions of TPR and FPR are:

$$TPR = \frac{TP}{TP+FN} \quad \& \quad FPR = \frac{FP}{FP+TN}$$

TP, FP, TN, and FN are the counts of True Positive, False Positive, True Negative, and False Negative, respectively. We use the area under the ROC curve values **AUC** to evaluate the overall performance of a predictor in a single value. The diagonal line is the ROC curve for a random (naive) classifier. The AUC for a random classifier is 0.5, and the better the classifier is, the higher the AUC value is. Since ROC curves are insensitive to dataset imbalance, so are AUC values. For a randomly selected positive class instance and a randomly selected negative class instance, the AUC is the probability of ranking the positive instance higher than the negative one [67]. We used the Precision-Recall curves to chart the relationship between precision and recall where:

$$Precision = \frac{TP}{(TP+FP)} \quad \& \quad Recall = \frac{TP}{(TP+FN)}$$

We also used the success rate (SR) to evaluate our predictor. SR measures the percentage of protein sequences with the average score of the positive class residues in the sequence (MoRF residues) higher than the average score of the negative class (non-MoRF residues). SR scores are essential as they are insensitive to under-annotation in the data and factor out protein-level features' contribution. Using protein level features, e.g., protein-protein interaction data, can boost the AUC value of predictors by up-scoring entire sequences of hub proteins. However, the output of these predictors is less precise in identifying the location of MoRF residues within each sequence. Predictors with higher SR scores are more appropriate for locating the MoRF(s) position in a protein sequence.

## 3 Results

### 3.1 Prediction assessment

We compared MoRFchibi 2.0 (MC2) predictions with those available for MoRF-specific and more general predictors of protein-binding regions in IDRs. Specifically, we used the complete Test-151 set to compare MC2 predictions with those of the original MoRFchibi web and light methods, as well as the state-of-the-art MoRF predictors fMoRFpred and OPAL. In addition, we compared predictions with those of twelve methods that identify protein-binding segments in IDRs, independent of whether the segments undergo DtO transitions upon binding. **Figure 2** and **Table 2** reveal that MoRFchibi 2.0 outperforms all 16 tested predictors in terms of AUC. In the Test-151 set, we masked all low-quality MoRF sites to reduce the likelihood of false negatives. To be aligned with test datasets classically used in the field and test the impact of this procedure, we unmasked the LQ MoRF sites as well as non-MoRF residues outside DisProt sequences and created an additional test dataset. Results for this Test-151 (unmasked) dataset show a slight decrease in the AUC of most predictors, including all specialized MoRF predictors, but no substantial change in the raking in AUC performance. A recent study has shown that AlphaFold2 often assigns higher pLDDT scores to residues in MoRFs, residues that undergo DtO upon binding, and proposed that AlphaFold2 can identify MoRFs with high precision [40, 68]. Therefore, we included two additional predictors (AlphaFold-Bind and IPA-AF2_protein) that rely on AlphaFold-2 structures in their predictions, specifically AlphaFold-2's pLDDT scores [40, 43]. Thus, we excluded sequences in the Test-151 dataset that do not have a structure available in the AlphaFoldDB (see **Table 1**). Results in **Table 2** show that MoRFchibi 2.0 continues to outperform all predictors, including those that use AlphaFold-2 scores. Finally, we compared predictors using success rates. To do so, we used the test dataset, Test-151 (SR), which only contains sequences with both MoRF and non-MoRF residues. Results in **Figure 2** and **Table 2** reveal that MoRFchibi 2.0 provides predictions for the Test-151 (SR) subset, with the highest AUC and SR values for all predictors tested. Note that while DisoFLAG-PR (dotted orange line) has an AUC for this subset that equals that of MoRFchibi 2.0, the DisoFLAG-PR AUC value is the result of a better performance in identifying non-MoRF residues (top right corner) when compared to MoRFchibi 2.0, which is better in identifying MoRF residues (lower left corner of the ROC curve).

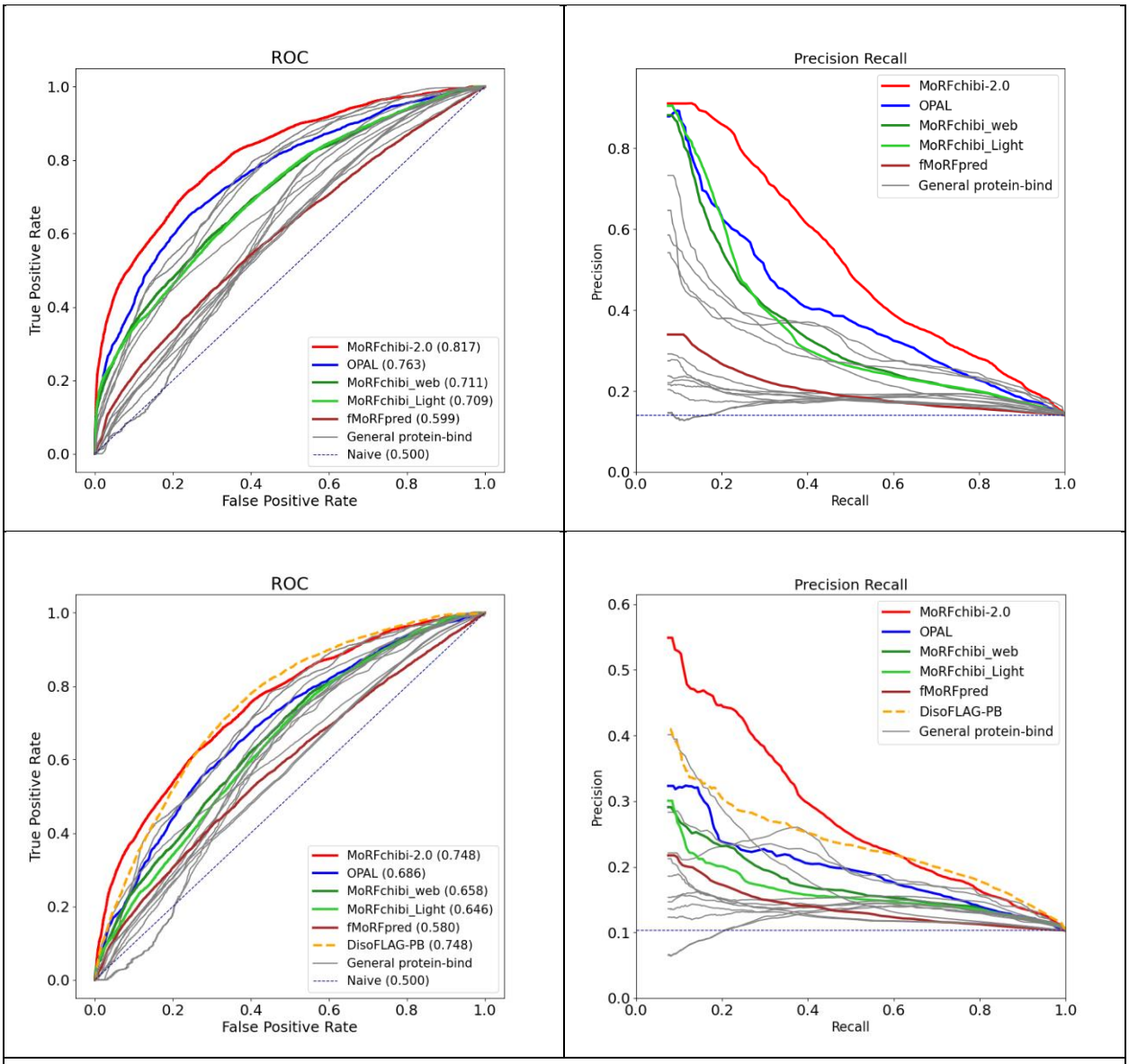


**Figure 2: ROC and PR curves generated with predictions of MoRFchibi 2.0 and various state-of-the-art methods**. Top) MoRFchibi 2.0 performance (left: ROC and right: Precision-Recall) against the Test-151 dataset compared to a set of MoRF predictors: OPAL, MoRFchibi_web, MoRFchibi_light, and fMoRFpred, and twelve general predictors of protein-binding segments in IDR (in gray lines). Bottom) the performance of these predictors against the SR subset. The dotted gray line represents the performance of a naïve predictor.

**Table 2:** Comparison of prediction performances. The AUC values for MoRFchibi 2.0 and other predictors using the complete test dataset, Test-151, the complete unmasked dataset, Test-151(unmask), the subset of the test dataset with AlphaFold structures, Test-151(AF2), and the subset of sequences with at least 5 MoRF and 5 non-MoRF residues, Test-151(SR). We also used the Test-151(SR) subset to compare these predictors' success rates. MoRF predictors are in bold, and the highest AUC values are underlined.

| Measure | AUC | AUC | AUC | AUC | Success Rate |
|---|---|---|---|---|---|
| Test Data | Complete | Unmasked | Subset(AF2) | Subset(SR) | Subset(SR) |
| **MoRFchibi 2.0** | **0.817** | **0.804** | **0.813** | **0.748** | **0.732** |
| **OPAL** | **0.763** | **0.753** | **0.752** | **0.685** | **0.661** |
| **MoRFchibi_web** | **0.711** | **0.707** | **0.721** | **0.658** | **0.714** |
| **MoRFchibi_light** | **0.709** | **0.702** | **0.715** | **0.646** | **0.688** |
| **fMoRFpred** | **0.599** | **0.593** | **0.600** | **0.580** | **0.670** |
| IPA_all | 0.757 | 0.729 | 0.741 | 0.709 | 0.705 |
| IPA_protein | 0.746 | 0.721 | 0.730 | 0.708 | 0.714 |
| DisoFLAG-PB | 0.735 | 0.723 | 0.739 | 0.748 | 0.732 |
| ProBiPred_protein | 0.670 | 0.671 | 0.667 | 0.621 | 0.571 |
| ENSHROUD_all | 0.636 | 0.614 | 0.657 | 0.651 | 0.625 |
| ENSHROUD_protein | 0.611 | 0.587 | 0.628 | 0.627 | 0.634 |
| DeepDISObind_all | 0.617 | 0.601 | 0.607 | 0.599 | 0.562 |
| DeepDISObind_protein | 0.606 | 0.591 | 0.595 | 0.592 | 0.625 |
| bindEmbed21IDR_idrGeneral | 0.606 | 0.614 | 0.598 | 0.573 | 0.500 |
| bindEmbed21IDR_rawGeneral | 0.596 | 0.607 | 0.589 | 0.564 | 0.402 |
| ANCHOR2 | 0.591 | 0.572 | 0.589 | 0.616 | 0.616 |
| DisoRDPbind | 0.590 | 0.573 | 0.592 | 0.621 | 0.652 |
| AlphaFold_binding | | | 0.666 | | |
| IPA-AF2_protein | | | 0.723 | | |

### 3.2 Case studies

To demonstrate MC2's output and help understand its predictions, we present prediction examples for six of the Test-151 proteins in **Figure 3**. For the example in **Figure 3 A**, MC2 successfully identifies the 157 residues annotated by DisProt as MoRF. It is clear from the ROC and PR curves in **Figure 2** that such "perfect" predictions are not always possible, and there are scenarios where MoRFchibi 2.0 is not able to identify MoRF sites, as illustrated in **Figure 3 B**, where MC2 only identifies residues 20-30 that are annotated in the DIBS database as MoRF and fails to identify the extended site 142-485 annotated as MoRF in DisProt. However, not all cases are clear-cut. In the case of the prediction shown in **Figure 3 C,** MC2 only identifies half of the MoRF annotated in DisProt (residues 1-105 case). However, the paper cited by DisProt states that 'the minimal AID construct σ54(16-41) is sufficient to form the activator complex' [69], i.e., indicating that the first half of this site is more important than the second, notably consistent with MC2's score distribution for this site. A similar case is shown in **Figure 3 D.** While the segment 412-490 is annotated as MoRF in DisProt, a structure of the interacting MoRF is only available for the segment 465-490, where the MC2 score is the highest. Perhaps MoRF annotations should not be binary. In **Figure 3 E**, MC2 scores suggest a MoRF between residues 378-439. However, DisProt only annotated residues 378-

393 and IDEAL residues 435-439 as MoRFs, leaving the region 394 to 434 as an over-prediction by MC2 or an under-annotated region. Interestingly, MC2 seems to predict regions that undergo DtO upon binding to proteins and other molecules, which can be considered over-predictions. **Figure 3 F** shows a case where MC2 scores identify the site around the 41 residue-long segment 294-334 as a MoRF. However, this 41 residue site, which binds to a protein partner, constitutes a zink-finger not stabilized by the binding to the protein partner but by zink ions [70].

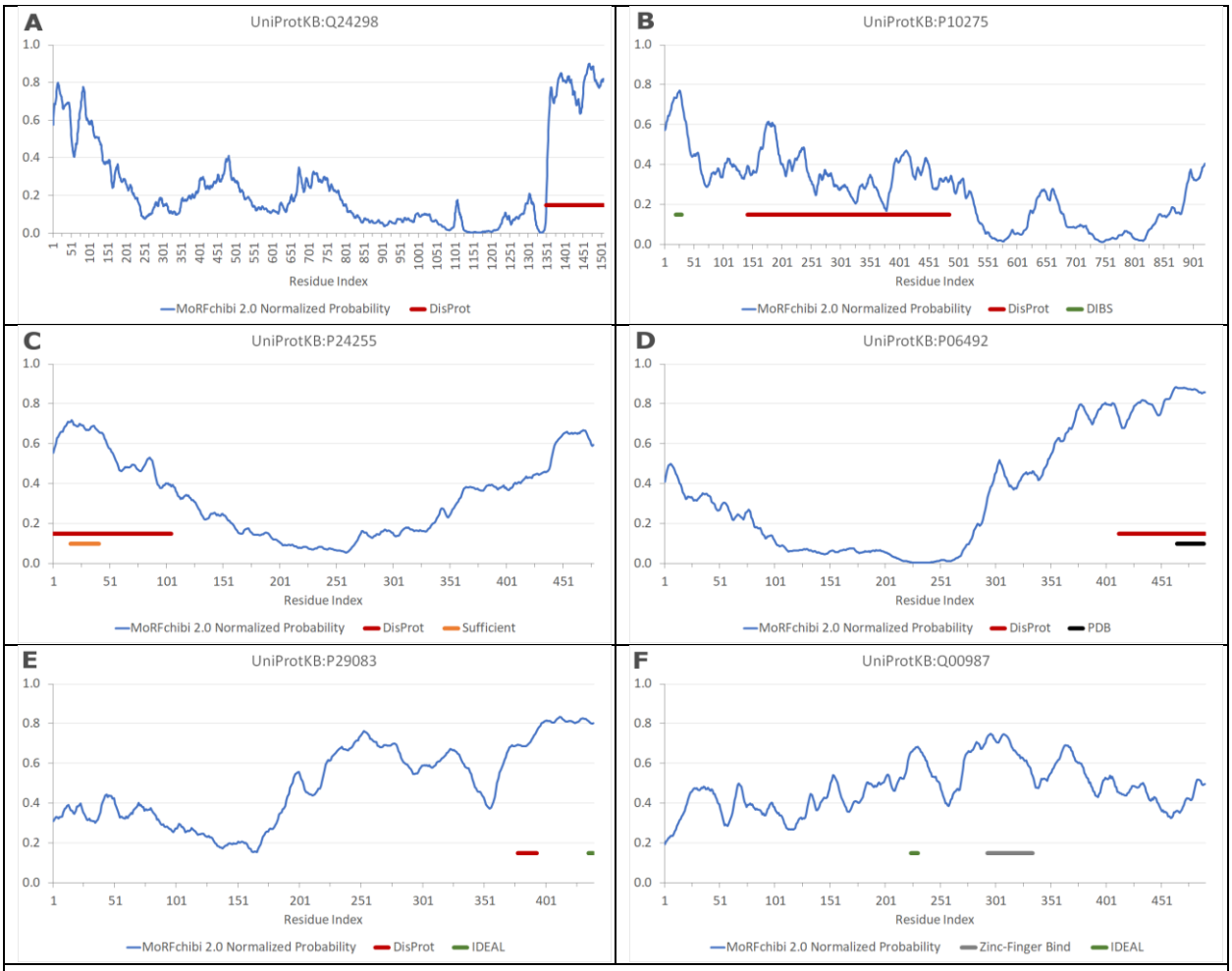


**Figure 3:** MoRFchibi 2.0 Normalized Probability scores in blue compared to the DisProt MoRF annotations in red, IDEAL and DIBS annotations in green, and those derived from PDB structures are in black. In (C), we also included a subset of DisProt annotations provided by the source publication as sufficient to form the protein complex in Orange.

## 4 Discussion

We introduce MoRFchibi 2.0 (MC2), a specialized prediction tool designed to identify the locations of Molecular Recognition Features (MoRFs) in protein sequences. MoRFs constitute a subset of all binding regions within IDRs, those undergoing a disorder-to-order transition upon binding to protein partners. Our results demonstrate that MC2 outperforms existing MoRF predictors and predictors that identify

protein-binding segments more broadly within IDRs. The latter group of predictors includes those that were ranked top in the community-driven critical assessments of protein intrinsic disorder predictions (CAID) rounds 1, 2, and 3 [71]. Importantly, MC2 prediction performance surpasses even those methods that leverage AlphaFold output and use advanced protein language models, delivering superior ROC and Precision-Recall curves and higher success rates. This result is notable, given recent studies showing that segments within IDRs that undergo DtO transitions often have higher pLDDT scores than the surrounding segments or IDRs that do not generally have MoRFs [68]. Regardless, neither MoRFchibi 2.0 predictions nor MoRF annotations are perfect. We provide examples demonstrating MC2 predictions in different scenarios, such as appropriate prediction, under-prediction, over-prediction, and predictions for regions with ambiguous or incomplete annotations.

MoRFchibi 2.0 generates output scores using an ensemble of logistic regression convolutional neural network models followed by a reverse Bayes Rule, which normalizes for the priors in the training data. Consequently, these scores represent probabilities normalized for the priors in the training data. As a result, the scores are individually interpretable, enabling direct comparison and integration with other scores following the same output framework, such as IPA [43]. We provide an HTML server for the broader community to use this new method and make the newly assembled dataset with MoRF annotations accessible to all developers: https://gsponerlab.msl.ubc.ca/software/morf_chibi/mc2.

**References:**


1. Haynes C, Oldfield CJ, Ji F, et al. Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. PLoS Comput Biol 2006; 2 (8):e100.
2. Oldfield CJ, Uversky VN, Dunker AK, et al. Chapter 1 - introduction to intrinsically disordered proteins and regions. In: Salvi N (ed). Intrinsically Disordered Proteins. Academic Press, Cambridge, Massachusetts, 2019, 1–34.
3. Wright PE, Dyson HJ. Intrinsically disordered proteins in cellular signalling and regulation. Nat Rev Mol Cell Biol. 2015 Jan; 16(1):18–29. pmid:25531225.
4. Oldfield CJ, Dunker AK. Intrinsically disordered proteins and intrinsically disordered protein regions. Annu Rev Biochem. 2014; 83:553–84. pmid:24606139.
5. Van der Lee R, Buljan M, Lang B, et al. Classification of intrinsically disordered regions and proteins. Chem Rev. 2014 Jul 9; 114(13):6589–631. pmid:24773235.
6. Aspromonte MC, Nugnes MV, Quaglia F, Bouharoua A; DisProt Consortium; Tosatto SCE, Piovesan D. DisProt in 2024: improving function annotation of intrinsically disordered proteins. Nucleic Acids Res. 2024 Jan 5;52(D1):D434-D441. doi: 10.1093/nar/gkad928. PMID: 37904585; PMCID: PMC10767923.
7. Deiana A, Forcelloni S, Porrello A, Giansanti A. Intrinsically disordered proteins and structured proteins with intrinsically disordered regions have different functional roles in the cell. PLoS One. 2019 Aug 19;14(8):e0217889. doi: 10.1371/journal.pone.0217889. PMID: 31425549; PMCID: PMC6699704.
8. Cheng Y, Oldfield CJ, Meng J, et al. Mining alpha-helix-forming molecular recognition features with cross species sequence alignments. Biochemistry. 2007 Nov 27;46(47):13468-77. doi: 10.1021/bi7012273. Epub 2007 Nov 1. PMID: 17973494; PMCID: PMC2570644.
9. Dosztányi Z, Mészáros B, Simon I. ANCHOR: web server for predicting protein binding regions in disordered proteins. Bioinformatics. 2009 Oct 15;25(20):2745-6. doi: 10.1093/bioinformatics/btp518. Epub 2009 Aug 28. PMID: 19717576; PMCID: PMC2759549.
10. Mészáros B, Simon I, Dosztányi Z. Prediction of protein binding regions in disordered proteins. PLoS Comput Biol. 2009 May;5(5):e1000376. doi: 10.1371/journal.pcbi.1000376. Epub 2009 May 1. PMID: 19412530; PMCID: PMC2671142.
11. Xue B, Dunker AK, Uversky VN. Retro-MoRFs: identifying protein binding sites by normal and reverse alignment and intrinsic disorder prediction. Int J Mol Sci. 2010 Sep 29;11(10):3725-47. doi: 10.3390/ijms11103725. PMID: 21152297; PMCID: PMC2996789.

12. Miri Disfani F, Hsu WL, Mizianty MJ, et al. MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins. 2012; Bioinformatics, 28(12): i75-i83.
13. Mooney C, Pollastri G, Shields DC, et al. Prediction of short linear protein binding regions. J Mol Biol. 2012 Jan 6;415(1):193-204. doi: 10.1016/j.jmb.2011.10.025. Epub 2011 Oct 21. PMID: 22079048.
14. Khan W, Duffy F, Pollastri G, et al. Predicting binding within disordered protein regions to structurally characterised peptide-binding domains. PLoS One. 2013 Sep 3;8(9):e72838. doi: 10.1371/journal.pone.0072838. PMID: 24019881; PMCID: PMC3760854.
15. Fang C, Noguchi T, Tominaga D, et al. MFSPSSMpred: identifying short disorder-to-order binding regions in disordered proteins based on contextual local evolutionary conservation. BMC Bioinformatics. 2013 Oct 4;14:300. doi: 10.1186/1471-2105-14-300. PMID: 24093637; PMCID: PMC3853019.
16. Disfani FM, Hsu WL, Mizianty MJ, Oldfield CJ, Xue B, Dunker AK, Uversky VN, Kurgan L. MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins. Bioinformatics. 2012 Jun 15;28(12):i75-83. doi: 10.1093/bioinformatics/bts209. PMID: 22689782; PMCID: PMC3371841.
17. Malhis N, and Gsponer J. Computational Identification of MoRFs in Protein Sequences. Bioinformatics (2015) 31 (11): 1738-1744. pmid: 25637562.
18. Jones DT, Cozzetto D. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. Bioinformatics. 2015 Mar 15;31(6):857-63. doi: 10.1093/bioinformatics/btu744. Epub 2014 Nov 12. PMID: 25391399; PMCID: PMC4380029.
19. Malhis N, Wong TCE, Nassar R, et al. Computational Identification of MoRFs in Protein Sequences Using Hierarchical Application of Bayes Rule. PLOS ONE (2015), DOI: 10.1371/journal.pone.0141603. pmid: 26517836.
20. Palopoli N, Lythgow KT, Edwards RJ. QSLiMFinder: improved short linear motif prediction using specific query protein data. Bioinformatics. 2015 Jul 15;31(14):2284-93. doi: 10.1093/bioinformatics/btv155. Epub 2015 Mar 19. PMID: 25792551; PMCID: PMC4495300.
21. Yan J, Dunker AK, Uversky VN, et al. Molecular recognition features (MoRFs) in three domains of life. Mol Biosyst. 2016 Mar;12(3):697-710. doi: 10.1039/c5mb00640f. PMID: 26651072.
22. Peng Z, Kurgan LA, 2015. High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder. Nucleic Acids Research, 43(18): e121.
23. Malhis N, Jacobson M, and Gsponer J. MoRFchibi SYSTEM: Software Tools for the Identification of MoRFs in Protein sequences. Nucleic Acids Research (2016), doi: 10.1093/nar/gkw409. Pmid: 27174932.
24. Sharma R, Kumar S, Tsunoda T, et al. Predicting MoRFs in protein sequences using HMM profiles. BMC Bioinformatics. 2016 Dec 22;17(Suppl 19):504. doi: 10.1186/s12859-016-1375-0. PMID: 28155710; PMCID: PMC5259822.
25. Krystkowiak I, Davey NE. SLiMSearch: a framework for proteome-wide discovery and annotation of functional modules in intrinsically disordered regions. Nucleic Acids Res. 2017 Jul 3;45(W1):W464-W469. doi: 10.1093/nar/gkx238. PMID: 28387819; PMCID: PMC5570202.
26. Peng Z, Wang C, Uversky VN, et al. Prediction of Disordered RNA, DNA, and Protein Binding Regions Using DisoRDPbind. Methods Mol Biol. 2017;1484:187-203. doi: 10.1007/978-1-4939-6406-2_14. PMID: 27787828.
27. Sharma R, Bayarjargal M, Tsunoda T, et al. MoRFPred-plus: Computational Identification of MoRFs in Protein Sequences using Physicochemical Properties and HMM profiles. J Theor Biol. 2018 Jan 21;437:9-16. doi: 10.1016/j.jtbi.2017.10.015. Epub 2017 Oct 16. PMID: 29042212.
28. Mészáros B, Erdos G, Dosztányi Z, IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding, Nucleic Acids Research, Volume 46, Issue W1, Jul 2 2018, Pages W329–W337, https://doi.org/10.1093/nar/gky384.
29. Sharma R, Raicar G, Tsunoda T, et al. OPAL: prediction of MoRF regions in intrinsically disordered protein sequences. Bioinformatics. 2018 Jun 1;34(11):1850-1858. doi: 10.1093/bioinformatics/bty032. PMID: 29360926.
30. Sharma R, Sharma A, Raicar G, et al. OPAL+: Length-Specific MoRF Prediction in Intrinsically Disordered Protein Sequences. Proteomics. 2019 Mar;19(6):e1800058. doi: 10.1002/pmic.201800058. Epub 2018 Nov 2. PMID: 30324701.
31. Fang C, Moriwaki Y, Tian A, et al. Identifying short disorder-to-order binding regions in disordered proteins with a deep convolutional neural network method. J Bioinform Comput Biol. 2019 Feb;17(1):1950004. doi: 10.1142/S0219720019500045. PMID: 30866736.

32. He H, Zhao J, Sun G. Computational prediction of MoRFs based on protein sequences and minimax probability machine. BMC Bioinformatics. 2019 Oct 28;20(1):529. doi: 10.1186/s12859-019-3111-z. PMID: 31660849; PMCID: PMC6819637.
33. Fang C, Moriwaki Y, Li C, et al. MoRFPred_en: Sequence-based prediction of MoRFs using an ensemble learning strategy. J Bioinform Comput Biol. 2019 Dec;17(6):1940015. doi: 10.1142/S0219720019400158. PMID: 32019410.
34. He H, Zhao J, Sun G. Prediction of MoRFs in Protein Sequences with MLPs Based on Sequence Properties and Evolution Information. Entropy (Basel). 2019 Jun 27;21(7):635. doi: 10.3390/e21070635. PMID: 33267349; PMCID: PMC7515128.
35. Hanson J, Litfin T, Paliwal K, et al. Identifying molecular recognition features in intrinsically disordered regions of proteins by transfer learning. Bioinformatics. 2020 Feb 15;36(4):1107-1113. doi: 10.1093/bioinformatics/btz691. PMID: 31504193.
36. He H, Zhou Y, Chi Y, et al. Prediction of MoRFs based on sequence properties and convolutional neural networks. BioData Min. 2021 Aug 14;14(1):39. doi: 10.1186/s13040-021-00275-6. PMID: 34391457; PMCID: PMC8364704.
37. Hu G, Katuwawala A, Wang K, et al. flDPnn: Accurate intrinsic disorder prediction with putative propensities of disorder functions. Nat Commun. 2021 Jul 21;12(1):4438. doi: 10.1038/s41467-021-24773-7. PMID: 34290238; PMCID: PMC8295265.
38. Zhang F, Zhao B, Shi W, et al. DeepDISOBind: accurate prediction of RNA-, DNA- and protein-binding intrinsically disordered residues with deep multi-task learning. Brief Bioinform. 2022 Jan 17;23(1):bbab521. doi: 10.1093/bib/bbab521. PMID: 34905768.
39. Katuwawala A, Zhao B, Kurgan L. DisoLipPred: accurate prediction of disordered lipid-binding residues in protein sequences with deep recurrent networks and transfer learning. Bioinformatics. 2021 Dec 22;38(1):115-124. doi: 10.1093/bioinformatics/btab640. PMID: 34487138.
40. Piovesan D, Monzon AM, Tosatto SCE. Intrinsic protein disorder and conditional folding in AlphaFoldDB. Protein Sci. 2022 Nov;31(11):e4466. doi: 10.1002/pro.4466. PMID: 36210722; PMCID: PMC9601767.
41. Peng Z, Li Z, Meng Q, et al. CLIP: accurate prediction of disordered linear interacting peptides from protein sequences using co-evolutionary information. Brief Bioinform. 2023 Jan 19;24(1):bbac502. doi: 10.1093/bib/bbac502. PMID: 36458437.
42. He H, Zhou Y, Chi Y, He J. Prediction of MoRFs based on sequence properties and convolutional neural networks. BioData Min. 2021 Aug 14;14(1):39. doi: 10.1186/s13040-021-00275-6. PMID: 34391457; PMCID: PMC8364704.
43. Malhis, N. 2024. Probabilistic Annotations of Protein Sequences for Intrinsically Disordered Features. bioRxiv doi: 10.1101/2024.12.18.629275.
44. Pang Y, Liu B. DisoFLAG: accurate prediction of protein intrinsic disorder and its functions using graph-based interaction protein language model. BMC Biol. 2024 Jan 2;22(1):3. doi: 10.1186/s12915-023-01803-y. PMID: 38166858; PMCID: PMC10762911.
45. Oldfield CJ, Cheng Y, Cortese MS, et al. Coupled folding and binding with alpha-helix-forming molecular recognition elements. Biochemistry 2005; 44 (37):12454-12470. doi:10.1021/bi050736e.
46. Mohan A, Oldfield CJ, Radivojac P, et al. Analysis of molecular recognition features (MoRFs). J Mol Biol 2006;362:1043–59.
47. Vacic V, Oldfield CJ, Mohan A, et al. Characterization of molecular recognition features, MoRFs, and their binding partners. J Proteome Res 2007;6:2351–66.
48. Necci M, Piovesan D, Hoque T, et al. (2021). Critical assessment of protein intrinsic disorder prediction. Nature Methods, 18(5), 472-481. https://doi.org/10.1038/s41592-021-01117-3
49. Conte AD, Mehdiabadi M, Bouhraoua A, Miguel Monzon A, Tosatto SCE, Piovesan D. Critical assessment of protein intrinsic disorder prediction (CAID) - Results of round 2. Proteins. 2023 Dec;91(12):1925-1934. doi: 10.1002/prot.26582. Epub 2023 Aug 25. PMID: 37621223.
50. Del Conte A, Bouhraoua A, Mehdiabadi M, et al. (2023). CAID prediction portal: A comprehensive service for predicting intrinsic disorder and binding regions in proteins. Nucleic Acids Research. https://doi.org/10.1093/nar/gkad430.
51. Basu S, Zhao B, Biró B, Faraggi E, Gsponer J, Hu G, Kloczkowski A, Malhis N, Mirdita M, Söding J, Steinegger M, Wang D, Wang K, Xu D, Zhang J, Kurgan L. DescribePROT in 2023: more, higher-quality and experimental annotations and improved data download options. Nucleic Acids Res. 2024 Jan 5;52(D1):D426-D433. doi: 10.1093/nar/gkad985. PMID: 37933852; PMCID: PMC10767971.

52. Chen R, Li X, Yang Y, Song X, Wang C, Qiao D. Prediction of protein-protein interaction sites in intrinsically disordered proteins. Front Mol Biosci. 2022 Sep 30;9:985022. doi: 10.3389/fmolb.2022.985022. PMID: 36250006; PMCID: PMC9567019.
53. Han B, Ren C, Wang W, Li J, Gong X. Computational Prediction of Protein Intrinsically Disordered Region Related Interactions and Functions. Genes (Basel). 2023 Feb 8;14(2):432. doi: 10.3390/genes14020432. PMID: 36833360; PMCID: PMC9956190.
54. Basu S, Kihara D, Kurgan L. Computational prediction of disordered binding regions. Comput Struct Biotechnol J. 2023 Feb 10;21:1487–1497. doi: 10.1016/j.csbj.2023.02.018. PMID: 36851914; PMCID: PMC9957716.
55. Kurgan L, Hu G, Wang K, Ghadermarzi S, Zhao B, Malhis N, Erdős G, Gsponer J, Uversky VN, Dosztányi Z. Tutorial: a guide for the selection of fast and accurate computational tools for the prediction of intrinsic disorder in proteins. Nat Protoc. 2023 Nov;18(11):3157–3172. doi: 10.1038/s41596-023-00876-x. Epub 2023 Sep 22. PMID: 37740110.
56. Malhis, N, Gsponer, J. Computational Prediction of Linear Interacting Peptides. Springer Natures. Methods in Molecular Biology book, "Prediction of Protein Secondary Structure" second edition, chapter 14. (2024).
57. Juniku B, Mignon J, Carême R, Genco A, Obeid AM, Mottet D, Monari A, Michaux C. Intrinsic disorder and salt-dependent conformational changes of the N-terminal region of TFIP11 splicing factor. Int J Biol Macromol. 2024 Jul 30;277(Pt 3):134291. doi: 10.1016/j.ijbiomac.2024.134291. Epub ahead of print. PMID: 39089542.
58. Molzahn C, Kuechler ER, Zemlyankina I, Nierves L, Ali T, Cole G, Wang J, Albu RF, Zhu M, Cashman NR, Gilch S, Karsan A, Lange PF, Gsponer J, Mayor T. Shift of the insoluble content of the proteome in the aging mouse brain. Proc Natl Acad Sci U S A. 2023 Nov 7;120(45):e2310057120. doi: 10.1073/pnas.2310057120. Epub 2023 Oct 31. PMID: 37906643; PMCID: PMC10636323.
59. Sundar A, Umashankar P, Sankar P, Ramasamy K, Venkataraman S. Intrinsic disorder in flaviviral capsid proteins and its role in pathogenesis. J Biosci. 2024;49:57. PMID: 38783793.
60. Sun X, Malhis N, Zhao B, Xue B, Gsponer J, Rikkerink EHA. Computational Disorder Analysis in Ethylene Response Factors Uncovers Binding Motifs Critical to Their Diverse Functions. Int J Mol Sci. 2019 Dec 20;21(1):74. doi: 10.3390/ijms21010074. PMID: 31861935; PMCID: PMC6981732.
61. Schad E, Fichó E, Pancsa R, Simon I, Dosztányi Z, Mészáros B. DIBS: a repository of disordered binding sites mediating interactions with ordered proteins. Bioinformatics. 2018 Feb 1;34(3):535-537. doi: 10.1093/bioinformatics/btx640. PMID: 29385418; PMCID: PMC5860366.
62. Fichó E, Reményi I, Simon I, Mészáros B. MFIB: a repository of protein complexes with mutual folding induced by binding. Bioinformatics. 2017 Nov 15;33(22):3682-3684. doi: 10.1093/bioinformatics/btx486. PMID: 29036655; PMCID: PMC5870711.
63. Fukuchi S, Amemiya T, Sakamoto S, Nobe Y, Hosoda K, Kado Y, Murakami SD, Koike R, Hiroaki H, Ota M. IDEAL in 2014 illustrates interaction networks composed of intrinsically disordered proteins and their binding partners. Nucleic Acids Res. 2014 Jan;42(Database issue):D320-5. doi: 10.1093/nar/gkt1010. Epub 2013 Oct 30. PMID: 24178034; PMCID: PMC3965115.
64. Malhis, N. 2024. Pre-processing annotated homologous regions in protein sequences concerning machine-learning applications. bioRxiv doi: 10.1101/2024.10.25.620288.
65. Varadi M, Anyango S, Deshpande M, et al. AlphaFold protein structure database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. 2022;50:D439–D444.
66. Dosztányi, Z., et al. 2005 The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. J. Mol. Biol.347827–839.
67. Fawcett T. (2006); An introduction to ROC analysis, Pattern Recognition Letters, 27, 861–874.
68. Alderson TR, Pritišanac I, Kolarić Đ, Moses AM, Forman-Kay JD. Systematic identification of conditionally folded intrinsically disordered regions by AlphaFold2. Proc Natl Acad Sci U S A. 2023 Oct 31;120(44):e2304302120. doi: 10.1073/pnas.2304302120. Epub 2023 Oct 25. PMID: 37878721; PMCID: PMC10622901.
69. Siegel AR, Wemmer DE. Role of the σ54 Activator Interacting Domain in Bacterial Transcription Initiation. J Mol Biol. 2016 Nov 20;428(23):4669-4685. doi: 10.1016/j.jmb.2016.10.007. Epub 2016 Oct 11. PMID: 27732872; PMCID: PMC5116005.

70. Yu GW, Allen MD, Andreeva A, Fersht AR, Bycroft M. Solution structure of the C4 zinc finger domain of HDM2. Protein Sci. 2006 Feb;15(2):384-9. doi: 10.1110/ps.051927306. Epub 2005 Dec 29. PMID: 16385008; PMCID: PMC2242465.
71. https://caid.idpcentral.org/challenge/results