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

Molecular Recognition Features (MoRFs) are segments within disordered protein regions (IDRs) that undergo a disorder-to-order transition upon binding to their partners. Identifying MoRFs remains a significant challenge. 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 logistic regression convolutional neural network 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: An online server: https://mc2.msl.ubc.ca/index.xhtml and code: https://github.com/NawarMalhis/MC2.git.

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 [6, 7].

As of the latest release of the DisProt database [8] 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 [9]. 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 [10-46].

A subfamily of protein-binding segments in IDRs is known as molecular recognition features (MoRFs) [47-49]. 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 [13], MoRFpred [18], fMoRFpred [23], MoRFchibi_light [25], MoRFchibi_web [21], OPAL [31], MoRFPred-plus [29], and MoRFcnn [44]. Since MoRFs represent only a small fraction of all residues in protein sequences, i.e., low "priors" probabilities, we require strong evidence, i.e., high-quality predictors, to achieve acceptable posterior probabilities [50], i.e., precision.

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 [51, 52] and are available among the CAID Portal set of predictors [53]. MoRFchibi light predictions are included in the DescribePROT database [54], discussed in several review articles [55-59], and utilized in various applications [60-63]. 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 [18]. 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 MoRFchibi 2.0, a redesigned MoRF predictor that encompasses an ensemble of four logistic regression convolutional neural network (LR-CNN) 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 [64], MFIB [65], and IDEAL [66]. 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, including protein and nucleotide binding, as well as disorder-to-order (DtO) transitions. DIBS and IDEAL are databases that 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 using experimental data and restricted the 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 yielded 1,217 MoRF segments (HQ and LQ) across 1,030 sequences, totaling 52,560 residues. Thus, 769 sequences are annotated by HQ MoRFs and possibly LQ MoRFs, and the remaining 261 sequences are annotated by LQ MoRFs only.

A high degree of redundancies in the testing data can lead to biased evaluations. In the training data, it leads the optimization process to over-weight these redundant sequences at the expense of others. Thus, we used CD-Hit [67] to cluster these sequences at an 80% identity. To avoid unnecessarily losing some training MoRF annotations, 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 a combined 46,504 HQ and LQ MoRF residues. This included 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 the total number of HQ MoRF residues in each cluster. 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 number of HQ MoRF residues.

We used four subsets containing both LQ & HQ MoRF annotations to train four cross-validated LR-CNN models and the remaining fifth subset for testing. We further filtered out sequences in the test dataset with an identity higher than 30% to those in the four training datasets, resulting in a test dataset with 143 sequences (Test-143). Although sequences in the test dataset are less than 30% identical to those used in the training, short test regions can still have local homology to training sequences. Thus, we used HAM [68] to identify and mask out regions with 10 or more residues in the test sequences that are homologs (with an identity of greater than 80%) to the training sequences. HAM identified three regions (totaling 33 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 22 residues. Thus, the test set (Test-143) contains 143 sequences with 148 MoRFs, totaling 7,757 MoRF residues. For completeness, we also created Test-143 (unmasked), for which all previously masked residues, except those masked by HAM, were unmasked. In addition, we created the Test-128 (AF2) subset, which is limited to Test-143 sequences with AlphaFold2 structures in the AlphaFoldDB database [69]. We also created the Test-106 (SR) subset, which only contains Test-143 sequences with both MoRF and non-MoRF residues. Specifically, we excluded all sequences from Test-143 that do not have at least five residues for each class, i.e., all sequences must have five or more MoRF residues and five or more non-MoRF residues. Moreover, as SR is not sensitive to under-annotated sequences and to avoid significantly reducing the test data size, we unmasked non-MoRF sequences outside DisProt. **Table 1** describes the details of all four training datasets, the Test-143 dataset, and the three variations.

Table 1: Training and testing datasets. Four subsets, cv1, cv2, cv3, and cv4, utilize LQ annotations for training and HQ annotations for validation. And the fifth only used HQ annotations for testing.

	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-143	HQ	126/143	148	46,744	7,757	20,037
Test-143 (unmasked)	HQ	126/143	148	66,759	7,757	22
Test-128 (AF2)	HQ	113/128	132	33,923	7,246	16,449
Test-106 (SR)	HQ	106/106	128	49,676	5,823	1,163

2.2 Input Features

We used the seven features introduced in [45] 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 [70]. Specifically, the first interaction energy feature is the sum of the pairwise energies between the residue at k and all residues within a window of 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 residues 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 [45].

2.3 Model selection and optimization

We used the ensemble window-in/window-out structure utilizing four LR-CNN models introduced in [45]. An input sliding window size ten is padded by the 250 residues on each side, totaling 510 window-in residues. If not enough residues are available for padding in the input sequence, blank residues are used in their place. A blank residue is a hypothetical residue with all its features are zeros. We trained four LR-CNN models, each on three CV subsets, and validated them on the fourth (see **Table 1** for details**)**. Each model comprises three convolutional layers with a kernel size of seven, separated by two average pooling layers with a kernel size of seven, followed by two fully connected (FC) layers with the output size of the first FC layer set to 150. All the above hyperparameters were selected using a grid search to minimize the loss against the validation sets. We used a binary cross-entropy with a sigmoid (BCEWithLogitsLoss) as the training loss and applied a sigmoid function to the evaluation output of each LR-CNN model. Thus, each LR-CNN is a logistic regression 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 LR-CNN models, each minimizing the validation loss. To avoid overfitting the validation subsets, we then selected the instance with the third-best validation AUC.

Finally, we need to highlight two crucial design issues: first, we are using a large padding size, 250 residues on each side, to provide the model with a larger receptive field, thus enabling a larger environment to be

used in the evaluation whenever needed. Second, we employed an ensemble of four LR-CNNs trained on overlapping datasets to increase prediction consistency and mitigate potential biases that could arise from the limitations of our training data size.

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 represent the counts of True Positives, False Positives, True Negatives, and False Negatives, respectively. We use the area under the ROC curve (AUC) values to evaluate the overall performance of a predictor in a single value. The diagonal line is the ROC curve for a random (naïve) classifier. The AUC for a random classifier is 0.5, and the better the classifier is, the higher its AUC value will be. 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 [71]. We used the Precision-Recall curves to chart the relationship between precision and recall where:

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

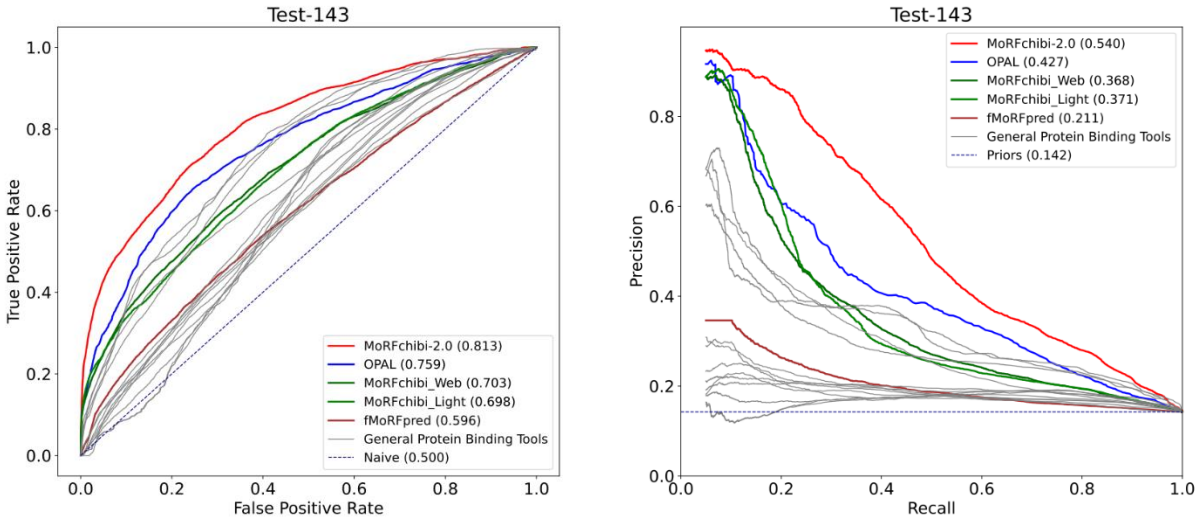
We included the average precision score (APS) [52] for each MoRF predictor in the legends of the PR curves. We also used the success rate (SR) to evaluate our predictor. SR measures the percentage of protein sequences with the average score of the positive class residues in the sequence (MoRF residues) 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 the contribution of protein-level features that adjust the scores of residues for entire proteins based on the overall protein activity. Using protein-level features, such as protein-protein interaction data, can boost the AUC value of predictors by upscoring entire sequences of hub proteins. However, the outputs of predictors relying on protein-level features are less precise in identifying the location of MoRF residues within each sequence. Predictors with higher SR scores are more suitable for determining the position of the MoRF(s) 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 intrinsically disordered regions (IDRs). Specifically, we used the complete Test-143 set to compare MC2 predictions with those of the original MoRFchibi web and light methods, as well as those MoRF predictors that participated in the CAID1 or CAID2 competitions, such as fMoRFpred and OPAL. In addition, we compared predictions with those of thirteen methods that identify protein-binding segments in IDRs, independent of whether the segments undergo DtO transitions upon binding. **Figure 1** and **Table 2** reveal that MoRFchibi 2.0 outperforms all 17 tested predictors in terms of AUC and APS. In the Test-143 set, we masked all low-quality MoRF sites to reduce the likelihood of false negatives. To align with test datasets commonly used in the field and assess 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-143 (unmasked) dataset show a slight decrease in the AUC and APS of most predictors, including all specialized MoRF predictors, but no substantial change in the ranking 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 [42, 72]. 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 [42, 45]. Thus, we excluded sequences in the Test-143 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-106 (SR), which only contains sequences with both MoRF and non-MoRF residues. Results in **Figure 1** and **Table 2** reveal that MoRFchibi 2.0 provides predictions for the Test-106 (SR) subset, with the highest AUC and SR values for all predictors tested. Note that while DisoFLAG-PR (dotted Brown line) has an AUC for this subset that almost 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).



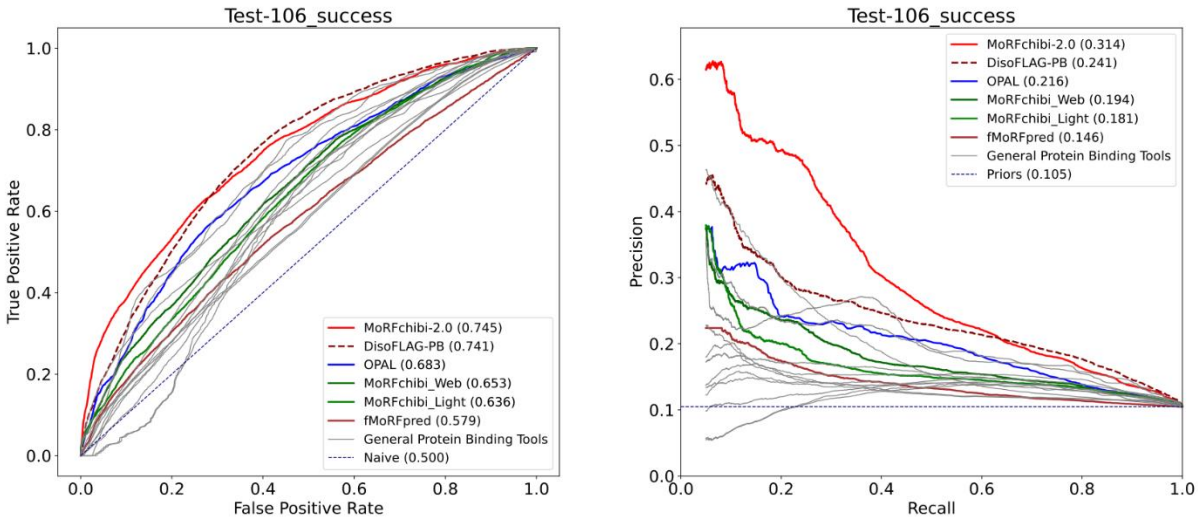


Figure 1: ROC and PR curves generated with predictions of MoRFchibi 2.0 and various methods. **Top)** MoRFchibi 2.0 performance (left: ROC and right: Precision-Recall) against the Test-143 dataset compared to a set of MoRF predictors: OPAL, MoRFchibi_web, MoRFchibi_light, and fMoRFpred, and thirteen general predictors of protein-binding segments in IDR (in gray lines). **Bottom)** the performance of these predictors against the SR subset, where the general binding predictor DisoFLAG-PB is in the doted Broun line to highlight MoRFchibi 2.0 higher performance at the most critical lower left corner of the ROC curve and the left of the PR curve. In the PR curves (Top and Bottom), to improve the clarity of the figures, we excluded points with recall values less than 0.05. The dotted gray lines represent 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-143, the complete unmasked dataset, Test-143(unmask), the subset of the test dataset with AlphaFold structures, Test-128(AF2), and the subset of sequences with at least 5 MoRF and 5 non-MoRF residues, Test-106(SR). We also used the Test-106 (SR) subset to compare the success rates of these predictors. MoRF predictors are in bold, and the highest AUC and Success Rate values are underlined.

Measure	AUC	AUC	AUC	AUC	Success Rate
Test Data	Complete	Unmasked	Subset (AF2)	Subset (SR)	Subset (SR)
MoRFchibi 2.0	**<u>0.813</u>**	**<u>0.803</u>**	**<u>0.806</u>**	**<u>0.745</u>**	**<u>0.736</u>**
OPAL	**0.759**	**0.752**	**0.744**	**0.683**	**0.660**
MoRFchibi_web	**0.703**	**0.702**	**0.710**	**0.653**	**0.717**
MoRFchibi_light	**0.698**	**0.694**	**0.700**	**0.636**	**0.689**
fMoRFpred	**0.597**	**0.591**	**0.597**	**0.579**	**0.670**
IPA_all	0.750	0.722	0.729	0.703	0.708
IPA_protein	0.741	0.716	0.720	0.705	0.717
DisoFLAG-PB	0.727	0.715	0.726	0.741	0.726
ProBiPred_protein	0.675	0.676	0.673	0.627	0.575
AIUPred-Binding	0.629	0.610	0.623	0.645	0.632

ENSHROUD_all	0.628	0.606	0.645	0.644	0.632
ENSHROUD_protein	0.601	0.577	0.614	0.617	0.642
DeepDISObind_all	0.608	0.591	0.591	0.588	0.557
DeepDISObind_protein	0.598	0.582	0.581	0.581	0.632
bindEmbed21IDR_idrGeneral	0.609	0.618	0.604	0.576	0.481
bindEmbed21IDR_rawGeneral	0.601	0.611	0.595	0.569	0.377
DisoRDPbind	0.580	0.563	0.578	0.609	0.670
ANCHOR2	0.575	0.556	0.566	0.600	0.613
AlphaFold_binding			0.657		
IPA-AF2_protein			0.712		

3.2 Case studies

To demonstrate MC2's output and help understand its predictions, we present prediction examples for six of the Test-143 proteins in **Figure 2**. For the example in **Figure 2A**, MC2 successfully identifies the 157 residues annotated by DisProt as a MoRF. It is clear from the ROC and PR curves in **Figure 1** 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 2B**, 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. Not all cases are clear-cut; in **Figure 2C**, MC2 only identifies half of the MoRF annotated in DisProt (the 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' [73], 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 2D.** 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 2E**, MC2 scores suggest a MoRF between residues 378 and 439. However, DisProt only annotated residues 378-393 and IDEAL residues 435-439 as MoRFs, leaving the region 394-434 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 3F** 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 zinc finger not stabilized by binding to the protein partner but by zinc ions [74].

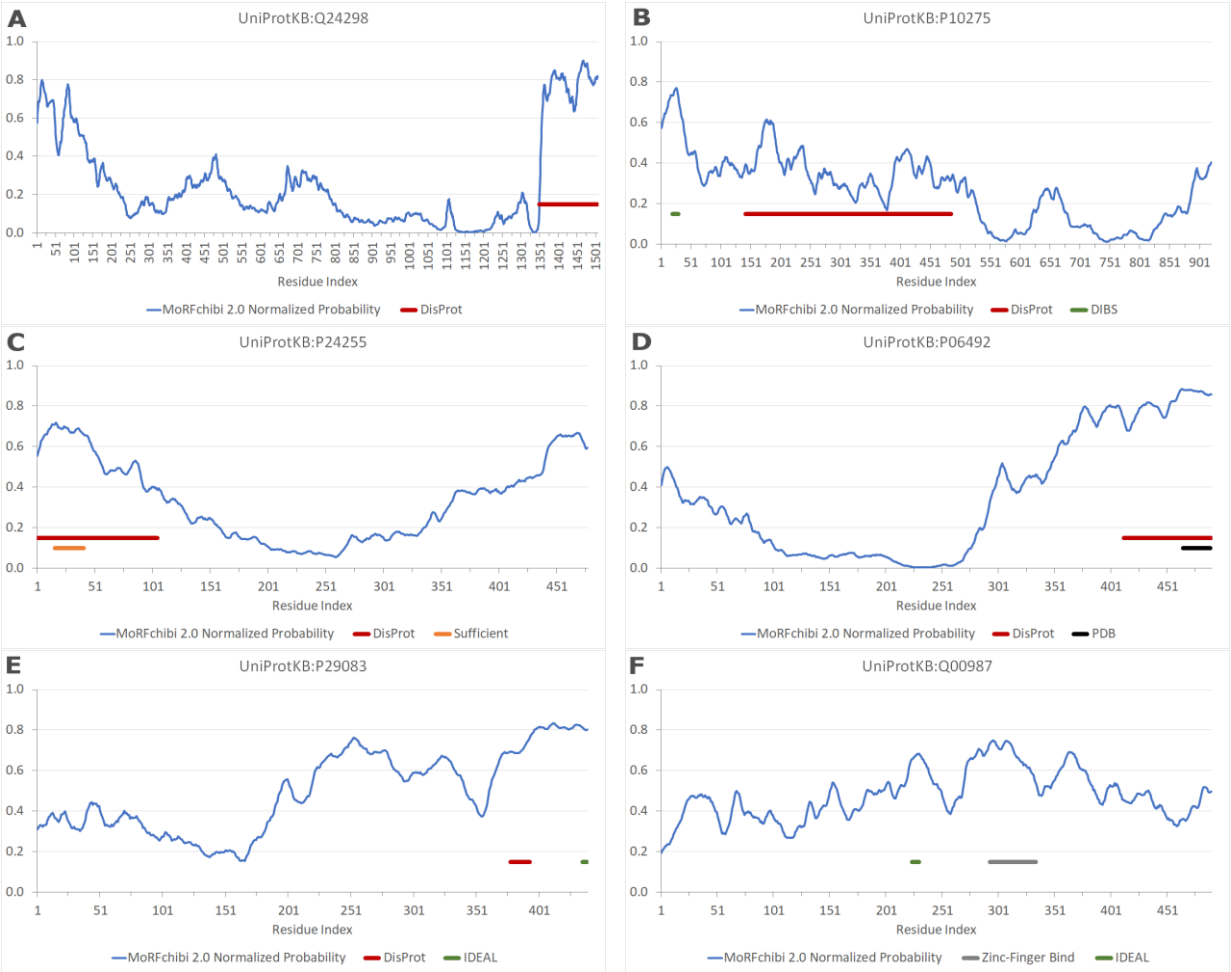


Figure 2: MoRFchibi 2.0 Normalized Probability scores in blue compared to the DisProt MoRF annotations in red, IDEAL and DIBS annotations in green, and those derived from PDB structures are in black. In (C), we also included a subset of DisProt annotations provided by the source publication, which were deemed 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 [75]. 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 [72]. Regardless, neither MoRFchibi 2.0 predictions nor MoRF annotations are perfect. We provide examples that demonstrate MC2 predictions in various

scenarios, including accurate predictions, under-predictions, over-predictions, and predictions for regions with ambiguous or incomplete annotations. Most importantly, MC2, relying solely on amino acid sequences, predicts regions with a low disorder 'signature' that bind to proteins and are likely to fold upon binding as MoRFs. However, we highlighted in **Figure 2F** that protein binding and protein folding can be influenced by environmental factors, such as ions or post-translational modifications, which can trigger a protein region to fold independently and mislead MC2 predictions.

MoRFchibi 2.0 generates output scores using an ensemble of four LR-CNN models followed by a reverse Bayes Rule. Consequently, these four scores, along with their average, represent probabilities normalized for the priors in the training data. As a result, MC2 scores are individually interpretable, enabling direct comparison and integration with other scores following the same output framework, such as IPA [45]. We provide an HTML server and downloadable code 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.

Author Contributions

Nawar Malhis: Dataset design and assembly; software design and implementation; writing (original draft, review, and editing). **Joerg Gsponer:** writing (review and editing); funding**.**

References:

1. Tompa, P., & Fersht, A. (2009). Structure and Function of Intrinsically Disordered Proteins (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781420078930
2. Van Roey K, Uyar B, Weatheritt RJ, Dinkel H, Seiler M, Budd A, Gibson TJ, Davey NE. Short linear motifs: ubiquitous and functionally diverse protein interaction modules directing cell regulation. Chem Rev. 2014 Jul 9;114(13):6733-78. doi: 10.1021/cr400585q. Epub 2014 Jun 13. PMID: 24926813.
3. Wright PE, Dyson HJ. Intrinsically disordered proteins in cellular signaling and regulation. Nat Rev Mol Cell Biol. 2015 Jan; 16(1):18–29. pmid:25531225.
4. Dunker AK, Bondos SE, Huang F, Oldfield CJ. Intrinsically disordered proteins and multicellular organisms. Semin Cell Dev Biol. 2015 Jan;37:44-55. doi: 10.1016/j.semcdb.2014.09.025. Epub 2014 Oct 13. PMID: 25307499.
5. Necci M, Piovesan D, Tosatto SC. Large-scale analysis of intrinsic disorder flavors and associated functions in the protein sequence universe. Protein Sci. 2016 Dec;25(12):2164-2174. doi: 10.1002/pro.3041. Epub 2016 Oct 25. PMID: 27636733; PMCID: PMC5119570.
6. Freedman SJ, Sun ZY, Poy F, Kung AL, Livingston DM, Wagner G, Eck MJ. Structural basis for recruitment of CBP/p300 by hypoxia-inducible factor-1 alpha. Proc Natl Acad Sci U S A. 2002 Apr 16;99(8):5367-72. doi: 10.1073/pnas.082117899. PMID: 11959990; PMCID: PMC122775.
7. Dyson HJ, Wright PE. Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol. 2005 Mar;6(3):197-208. doi: 10.1038/nrm1589. PMID: 15738986.
8. 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.
9. 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.
10. 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.

11. 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.
12. 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.
13. 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.
14. 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.
15. 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.
16. 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.
17. 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.
18. 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.
19. Malhis N, and Gsponer J. Computational Identification of MoRFs in Protein Sequences. Bioinformatics (2015) 31 (11): 1738-1744. pmid: 25637562.
20. 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.
21. 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.
22. 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.
23. 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.
24. 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.
25. 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.
26. 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.
27. 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.
28. 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.
29. 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.
30. 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.

31. 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.
32. 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.
33. 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.
34. 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.
35. 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.
36. 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.
37. 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.
38. 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.
39. 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.
40. 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.
41. 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.
42. 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.
43. 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.
44. 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.
45. Malhis, N. 2024. Probabilistic Annotations of Protein Sequences for Intrinsically Disordered Features. bioRxiv doi: 10.1101/2024.12.18.629275.
46. 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.
47. 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.
48. Mohan A, Oldfield CJ, Radivojac P, et al. Analysis of molecular recognition features (MoRFs). J Mol Biol 2006;362:1043–59.
49. 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.
50. Bayes, Thomas & Price, Richard (1763). "An Essay towards solving a Problem in the Doctrine of Chance. By the late Rev. Mr. Bayes, communicated by Mr. Price, in a letter to John Canton, A.M.F.R.S." Philosophical Transactions of the Royal Society of London. 53: 370–418. doi:10.1098/rstl.1763.0053.

51. 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
52. 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.
53. 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.
54. 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.
55. 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.
56. 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.
57. 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.
58. 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.
59. 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).
60. 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.
61. 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.
62. 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.
63. 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.
64. 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.
65. 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.
66. 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.
67. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012 Dec 1;28(23):3150-2. doi: 10.1093/bioinformatics/bts565. Epub 2012 Oct 11. PMID: 23060610; PMCID: PMC3516142.
68. Malhis, N. 2024. Pre-processing annotated homologous regions in protein sequences concerning machine-learning applications. bioRxiv doi: 10.1101/2024.10.25.620288.

69. 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.
70. 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.
71. Fawcett T. (2006); An introduction to ROC analysis, Pattern Recognition Letters, 27, 861–874.
72. 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.
73. 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.
74. 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.
75. https://caid.idpcentral.org/challenge/results