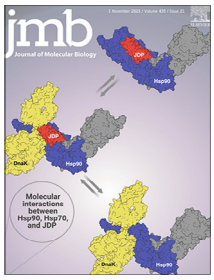

# CoMemMoRFPred: Sequence-based Prediction of MemMoRFs by Combining Predictors of Intrinsic Disorder, MoRFs and Disordered Lipid-binding Regions

**Sushmita Basu** [1], **Tamás Hegedűs** [2,3] **and Lukasz Kurgan** [1,*]

*1 - **Department of Computer Science,** Virginia Commonwealth University, USA*
*2 - **Department of Biophysics and Radiation Biology,** Semmelweis University, Budapest, Hungary*
*3 - **ELKH-SE Biophysical Virology Research Group,** Eötvös Loránd Research Network, Budapest, Hungary*

***Correspondence to Lukasz Kurgan:*** *Department of Computer Science, Virginia Commonwealth University, 401 West Main Street, Room E4225, Richmond, VA 23284, USA. lkurgan@vcu.edu (L. Kurgan)*
https://doi.org/10.1016/j.jmb.2023.168272
***Edited by Rita Casadio***

## Abstract

Molecular recognition features (MoRFs) are a commonly occurring type of intrinsically disordered regions (IDRs) that undergo disorder-to-order transition upon binding to partner molecules. We focus on recently characterized and functionally important membrane-binding MoRFs (MemMoRFs). Motivated by the lack of computational tools that predict MemMoRFs, we use a dataset of experimentally annotated MemMoRFs to conceptualize, design, evaluate and release an accurate sequence-based predictor. We rely on state-of-the-art tools that predict residues that possess key characteristics of MemMoRFs, such as intrinsic disorder, disorder-to-order transition and lipid-binding. We identify and combine results from three tools that include flDPnn for the disorder prediction, DisoLipPred for the prediction of disordered lipid-binding regions, and MoRFCHiBi$_{Light}$ for the prediction of disorder-to-order transitioning protein binding regions. Our empirical analysis demonstrates that combining results produced by these three methods generates accurate predictions of MemMoRFs. We also show that use of a smoothing operator produces predictions that closely mimic the number and sizes of the native MemMoRF regions. The resulting CoMemMoRFPred method is available as an easy-to-use webserver at http://biomine.cs.vcu.edu/servers/CoMemMoRFPred. This tool will aid future studies of MemMoRFs in the context of exploring their abundance, cellular functions, and roles in pathologic phenomena.



## Introduction

Intrinsically disordered proteins (IDPs) contain one or more intrinsically disordered regions (IDRs), which are sequence segments that lack stable structure under physiological conditions.[1–5] IDPs complement functional repertoire of rigid protein structures, expanding functional diversity and cellular complexity of proteomes.[6–9] IDPs contribute to a broad range of cellular processes, such as signaling, regulation and molecular recognition, to name just a few.[10–14] One of the key functional features of IDRs is facilitation of interactions with a variety of partner molecules that include proteins, peptides, nucleic acids and lipids.[15–23] In that context, IDRs offer a number of advantages over structured regions. They are capable of adopting different conformations upon binding to specific partners, allowing for interactions with multiple partners and enabling promiscuous and selective binding that depends on a cellular context[22–26]. IDRs can serve as "fuzzy" binding interfaces which lack a well-defined structure but can dynamically adjust and conform to an interacting partner[27]. Some IDRs possess short linear motifs that engage in specific protein–protein and protein-nucleic acid interac-

tions[13,15,28–31]. These motifs include Molecular Recognition Features (MoRFs), which are short regions embedded in longer IDRs and undergo disorder-to-order transition upon binding to proteins and peptides[15,28,13].

Recent studies suggest that IDRs are found in membrane-associated and transmembrane proteins, where a number of them transition from the disordered to the ordered state upon binding to lipids, sharing some characteristics of MoRFs.[20,32–34] These lipid-binding disorder-to-order transitioning IDRs are named MemMoRFs.[35] They are involved in regulation of apoptosis, phagocytosis, trafficking and shaping the cell membrane, and are associated with pathologies including neurodegeneration, viral infection, and toxicity.[35] A gold standard collection of experimentally annotated MemMoRFs was recently released in an online database.[35] While this database is currently limited to 107 proteins, many more MemMoRFs are expected to be found across proteomes, motivating the development of computational methods that would accurately predict MemMoRFs in protein sequences. These predictors could be used to support efforts to reduce the current MemMoRF annotation gap.

Practicality of developing computational sequence-based predictors is supported by the fact that disorder is encoded in the underlying sequences, i.e., disorder is intrinsic to the sequence.[5] In other words, IDRs have different sequence biases when compared to ordered/structured sequence regions,[3,36–38] making them predictable directly from their sequences. Consequently, well over 100 predictors of IDRs were developed.[39–47] Similarly, binding IDRs were also shown to have specific sequence biases,[36] suggesting that they can be also accurately identified in protein sequences. Correspondingly, close to 40 methods that predict particular functional subtypes of IDRs, such as MoRFs, RNA-binding, DNA-binding and lipid-binding IDRs, were released.[48] However, there are no methods that target prediction of MemMoRFs.

In addition to the MoRF-specific characteristics (i.e., intrinsic disorder and disorder-to-order transition upon binding), MemMoRF sequences interact with membranes. The latter characteristic is manifested via distinct sequence bias where MemMoRFs are enriched in positively charged Lys residues when compared to the other MoRFs. This enrichment can be explained by the fact that the intracellular leaflet of a membrane bilayer contains many lipids with negatively charged head groups.[49,50] While there are no predictors of MemMoRFs, there are methods that provide accurate predictions of residues that share some of the key characteristics that define MemMoRFs. Correspondingly, we investigate whether the current methods that predict the intrinsic disorder, MoRFs and disordered lipid-binding regions can be used to accurately predict MemMoRFs. We also combine results produced by these different types of methods to examine whether such approach would result in a more accurate prediction of MemMoRFs. Our overarching objective is to formulate a simple-to-implement sequence-based MemMoRF predictor that provides relatively high residue-level accuracy (i.e., can accurately identify amino acids that make up MemMoRFs) and that reproduces distribution of MemMoRF lengths (i.e., to ensure that the amino acids predicted as MemMoRFs form regions in the sequence that share similar distribution of their length when compared to the distribution of the length of native MemMoRF regions).

## Materials and methods

### Selection of relevant predictors

We identify suitable current disorder and disorder function predictors that cover the three defining aspects of MemMoRFs: intrinsic disorder, similarity to MoRFs, and binding to lipids. We rely on results from the recently completed community-organized Critical Assessment of Intrinsic disorder (CAID).[51,52] More specifically, CAID was organized and run by independent assessors (i.e., they did not submit methods for assessment) and comparatively evaluated a broad selection of predictors of IDRs and binding IDRs by comparing their predictions to native annotations using large blind datasets of IDPs (i.e., these proteins and annotations were not available to the participants before the assessment). This makes results produced by CAID arguably more reliable and less biased compared to the results of other studies that are produced by the authors of predictors and which utilize previously known ground truth annotations.[47,53–57] We use the main performance metric, the Area under the ROC curve (AUC_ROC) to quantify and compare predictive performance between different methods, with the underlying goal to select the most accurate tools.

Among the 32 participating disorder predictors and using the main DisProt dataset, CAID assessors identified flDPnn[58] as the most accurate tool.[51] This method secures AUC_ROC of 0.814, compared to the closest other predictors including RawMSA[59] (AUC_ROC of 0.780), ESpritz[60] (0.774), DisoMine[61] (0.765), and SPOT-Disorder2[62] (0.760). Moreover, flDPnn is also relatively fast, allowing to efficiently predict large collections of proteins. The average flDPnn's runtime is about 20 seconds per protein, compared to 250 seconds for RawMSA, 5 seconds for ESpritz, 3 seconds for DisoMine, and 2,000 seconds for SPOT-Disorder2.[51] These advantages of flDPnn were also highlighted in a commentary article for the CAID assessment,[52] further supporting our selection of this tool as a representative method for the accurate prediction of IDRs.

CAID also evaluated a broad collection of 11 predictors of binding IDRs.[51] The top five predictors in this category include ANCHOR2[63] (AUC_ROC of 0.742), DisoRDPbind[64,65] (0.729), MoRFCHiBi$_{Light}$[66] (0.720), MoRFCHiBi$_{Web}$[67] (0.702), and ANCHOR[68] (0.694). These methods are computationally efficient and complete predictions with an average runtime below 5 seconds per protein, except for MoRFCHiBi$_{Web}$ that takes about 100 seconds.[51] We select the most accurate ANCHOR2, which predicts IDRs that interact with proteins, and the top-ranked predictor of MoRFs, which is MoRFCHiBi$_{Light}$.

At present, there are only two methods that predict lipid-binding IDRs: DisoLipPred[69] and MemDis.[70] Both methods were released recently and after CAID was completed. They target different types of regions where the former predicts lipid-binding IDRs that exclude transmembrane regions (i.e., MemMoRFs are not localized in the transmembrane regions) and the later exclusively targets prediction of IDRs in the transmembrane proteins. While DisoLipPred is capable of making predictions for all protein sequences, MemDis is limited to the transmembrane proteins. Thus, MemDis cannot be used to make predictions for the membrane-associated proteins that also include MemMoRFs. Consequently, DisoLipPred is the only predictor of lipid-binding IDRs that is suitable for our study.

In summary, we selected four representative methods that produce predictions that are relevant to the identification of MemMoRFs: flDPnn, MoRFCHiBi$_{Light}$, ANCHOR2 and DisoLipPred. These methods take protein sequence as the sole input and generate a numeric score for each amino acid in the input sequence that quantifies its propensity to be disordered (flDPnn), to be part of a MoRF region (MoRFCHiBi$_{Light}$), to be disordered and bind proteins (ANCHOR2), and to be disordered and bind lipids (DisoLipPred).

## Test dataset

We use the MemMoRF database,[35] which includes 107 proteins, to develop a dataset of experimentally annotated MemMoRFs to test predictive performance of selected predictors. We ensure that the test proteins have low, below 25% sequence similarity to the training proteins that were used to develop the four selected predictors. This follows procedures used to evaluate related sequence-based predictors,[60,62,65,66,69,71,72] and is motivated by the fact that sequence alignment would not produce accurate predictions at these low levels of similarity. To this end, we combine the training dataset of the four methods with the proteins that we collect from the MemMoRF database and we cluster the resulting set of 17,579 proteins using the CD-Hit program[73] at 25% sequence similarity. Next, we select proteins from clusters that do not include any of the training proteins and we exclude sequences that are <30 amino acids in length (i.e., peptides). The resulting test dataset is composed of 41 proteins that include 684 MemMoRF residues and which share the low similarity with the training datasets of the four predictors.

## Evaluation metrics

The disorder and disorder function predictors produce two outputs for each amino acid in the input protein sequence: a numeric propensity value and a binary score. The propensities quantify likelihood for a given type of annotation (disorder, MoRF, MemMoRF, etc.). The binary scores are typically derived from the propensities based on a threshold, such that residues with propensities $\geq$ threshold are labelled as having a given annotation (i.e., positive residues) while the remaining amino acids are labelled as not having this annotation (i.e., negative residues). We evaluate the binary predictions using several popular metrics:

$$true\ \ positive\ \ rate\ \ (TPR) = sensitivity = \frac{TP}{TP + FN}$$

$$false\ \ positive\ \ rate\ \ (FPR) = 1 - specificity = \frac{FP}{FP + TN}$$

$$F1 = \frac{2TP}{2TP + FP + FN}$$

where TP, TN, FN and FP are the numbers of true positives (correctly predicted positives), true negatives (correctly predicted negatives), false negatives (positives incorrectly predicted as negatives), and false positives (negatives incorrectly predicted as positives), respectively. We calculate F1 at a threshold where the maximum value of F1 is obtained (F1max). We also compute sensitivity and F1 using the thresholds that produce low FPRs at 5% and 10%; these metrics quantify performance for predictions with a low rate of incorrect predictions of annotations.

We assess predictive quality of the propensity scores with the commonly used AUC_ROC metric. The ROC curve plots TPR versus FPR values when using every unique propensity value as the threshold; this metric offers a comprehensive evaluation across all possible propensity values and the corresponding binary predictions.[74] The AUC values range between 0 (perfectly incorrect/inverted predictions) and 1 (all predictions are correct), where 0.5 denotes random predictions and the expected values span interval between 0.5 and 1. Moreover, since the test dataset is highly imbalanced, where about 3% of the amino acids are MemMoRFs, we also separately evaluate a part of the ROC curves where the FPR values are relatively low, below 10%. This part corresponds to the predictions where the number of the putative MemMoRFs does not substantially exceed the rate

of the native MemMoRFs; the remaining parts of the ROC curve are arguably impractical because they correspond to significant over-predictions of MemMoRFs. Since the corresponding AUC values are small numbers that might be hard to interpret, we calculate ratio between the AUC for the predictions from a given tool and the AUC of a random prediction. Hence, rateAUC of 1 indicate that a given tool produces predictions that are equivalent to a random predictor and while values > 1 quantify the rate of improvement over a random predictor. The use of the rateAUC value is motivated by its applications in a number of related studies that feature similarly imbalanced test scenarios.[75–80]

### Statistical tests

We performed statistical tests to investigate whether differences in predictive performance between the best-performing and the other predictors are consistent across diverse datasets. To do that, we sample 100 sets of randomly selected 50% of the test proteins and compare the differences across the corresponding 100 evaluations. We use the student *t*-test when the corresponding data (i.e., measured values of the AUC, rateAUC, F1, and TPR metrics) are normal, and otherwise we apply the Wilcoxon rank-sum test. We determine normality using the Anderson-Darling test at $p$-value < 0.05.

## Results

### Comparative evaluation of the selected four predictors

We assess predictive performance for the four representative related predictors (flDPnn, DisoLipPred, MoRFCHiBi$_{Light}$ and ANCHOR2) when using their results to predict MemMoRFs in the test dataset. Moreover, we compare these results against a baseline produced by random predictions that mimic the distribution of the native MemMoRF regions. We setup the baseline by producing randomly generated scores between 0 to 1 such that using the threshold of 0.5 the number of the putative MemMoRFs matches with the number and the size of the native MemMoRF regions. Table 1 reports the results.

The baseline predictor obtains AUC = 0.52, rateAUC = 1.6 and $F1_{max}$ = 0.07, which as expected is near a random performance level. Interestingly, the four representative methods produce predictions with a broad range of predictive quality. Three of the four selected methods outperform the baseline while ANCHOR2 performs rather poorly with AUC = 0.52, rateAUC = 0.8 and $F1_{max}$ = 0.09. The ANCHOR2′s results can be explained by the fact that this tool focuses on predicting disordered protein-binding regions that share little in common with MemMoRFs. flDPnn produces the best and accurate predictions with AUC = 0.76, rateAUC = 2.7 and $F1_{max}$ = 0.15. The other two methods offer comparatively modest predictive quality, MoRFCHiBi$_{Light}$ with AUC = 0.69, rateAUC = 3.1 and $F1_{max}$ = 0.14, and DisoLipPred with AUC = 0.65, rateAUC = 2.7 and $F1_{max}$ = 0.12. The ROC curves provide further details (Figure 1). The curve of flDPnn is noticeably above the curves of the other three methods and the baseline for FPR > 0.1. The best option for the lower PFR values is MoRFCHiBi$_{Light}$ Statistical analysis based on the AUC values reveals that flDPnn significantly outperforms DisoLipPred, MoRFCHiBi$_{Light}$, ANCHOR2 and the baseline ($p$-value $\leq$ 0.01). Altogether, we find that three of the selected tools, flDPnn, DisoLipPred, and MoRFCHiBi$_{Light}$, provide predictions that can be used to relatively accurately identify MemMoRFs. They target prediction of different key characteristics of MemMoRFs including intrinsic disorder (flDPnn), MoRFs (MoRFCHiBi$_{Light}$) and lipid-binding (DisoLipPred), which can explain their relatively good performance.

### Selection of combination method

The fact that the three selected and well-performing predictors focus on different aspects of MemMoRFs suggests that combining their results might provide a more holistic and accurate MemMoRF prediction. We test this hypothesis by implementing and empirically evaluating all possible permutations of subsets of the four individual methods, i.e., we also include ANCHOR2 for completeness and to check whether combining it with the other tools could be helpful. There are total of ten permutations that include six pairs of methods, three combinations of three methods, and all four methods combined together. We implement the combinations in two steps. First, we standardize the numeric propensities generated by each tool into the unit range using the min–max normalization. In the second step we combine the normalized scores of the selected subset of methods and lastly, we normalize the resulting combined score to the unit range using the min–max approach. We combine the scores using five techniques: (1) SimpleProduct where we multiply the scores; (2) SimpleAverage where we average scores when each score is given equal importance; (3) WeightedAverage where score for a given method is multiplied by this method's AUC (weight) before calculating the average; 4) SimpleAverage *Minimum where we multiply the average by the minimal score selected among the considered methods; and (5) WeightedAverage *Minimum where we multiply the weighted average by the minimal score. The SimpleProduct implements an approach where the resulting prediction has high scores (i.e., predicts MemMoRFs) only when all contributing methods have high scores. The

Table 1 Predictive performance on the test dataset. Methods are sorted in the descending order by their AUC values. We use bold font to identify the best individual predictor and the best combined method based on their AUC values. We report medians of the metrics that we calculated over the 100 sampled test sets (see "Statistical test" section for details). We summarize results of the statistical significance analysis in the *x*/*y* format next to the reported median value, where *x* is for comparison against the best overall/combined method (CoMemMoRFpred) and y against the best single method (flDPnn), where ** and * denote statistically significant differences with *p*-values $\leq$ 0.01 and $\leq$0.05, respectively, and = denotes differences that are not statistically significant (*p*-value > 0.05).

| Predictors | AUC | rateAUC | F1_max | F1_FPR5% | F1_FPR10% | TPR_FPR5% | TPR_FPR10% |
|---|---|---|---|---|---|---|---|
| **CoMemMoRFpred** | **0.785 **/** | **3.8 **/** | **0.182 **/** | **0.135 **/** | **0.160 **/** | **0.19 **/** | **0.36 **/** |
| flDPnn + DisoLipPred + MoRFCHiBi$_{Light}$ | 0.778 */= | 3.6 **/= | 0.178 **/= | 0.115 **/** | 0.158 **/= | 0.16 **/** | 0.33 **/** |
| **flDPnn** | **0.761 /*** | **2.7 /**** | **0.155 /**** | **0.095 /**** | **0.114 /**** | **0.12 /**** | **0.23 /**** |
| MoRFCHiBi$_{Light}$ | 0.689 **/** | 3.1 **/** | 0.135 **/** | 0.119 **/** | 0.127 =/** | 0.16 **/** | 0.27 **/** |
| DisoLipPred | 0.650 **/** | 2.7 =/** | 0.118 **/** | 0.115 **/** | 0.103 =/** | 0.14 **/** | 0.22 =/** |
| Baseline | 0.523 **/** | 1.6 **/** | 0.073 **/** | 0.065 **/** | 0.660 **/** | 0.06 **/** | 0.06 **/** |
| ANCHOR2 | 0.515 **/** | 0.8 **/** | 0.086 **/** | 0.028 **/** | 0.057 **/** | 0.03 **/** | 0.12 **/** |

SimpleAverage predicts MemMoRFs when multiple but not necessarily all of the input scores are high. The WeightedAverage is a similar approach but it weights the scores by the overall predictive performance of the corresponding methods. The SimpleAverage*Minimum and WeightedAverage*Minimum techniques compute scores that hybridize averaging and multiplying, i.e., not all inputs scores must be high for the resulting score to be high (i.e., predict MemMoRF) but the lowest input score cannot be low.

We compare AUCs generated by each of the ten combinations and using the five techniques (total of 50 options) in Figure 2. We find that the best AUC = 0.778 is secured by the SimpleAverage*Minimum and WeightedAverage*Minimum techniques that combine predictions from flDPnn, MoRFCHiBi$_{Light}$, and DisoLipPred, closely followed by the SimpleProduct of the same three methods with AUC = 0.776. The highest AUCs for the SimpleAverage (AUC = 0.754) and WeightedAverage (AUC = 0.757) that are based on combining flDPnn and MoRFCHiBi$_{Light}$ are similar and substantially lower than the AUCs of the other three techniques. The finding that SimpleAverage and WeightedAverage produce similar quality of predictions suggests that weights are not helpful. Moreover, the observation that techniques that use multiplications are better than those that rely on average suggests that MemMoRFs are more accurately predicted when all contributing methods are required to produce high scores. This can be explained by the fact that the individual predictors cover somehow orthogonal characteristics of MemMoRFs (i.e., intrinsic disorder, lipid binding and MoRF) and only combing them all together reflect the "complete" nature of MemMoRFs. Moreover, our results reveal that combinations that involve ANCHOR2 are outperformed by those that exclude ANCHOR2, which is in line with the results for the individual predictors in Table 1.

Along with the AUC values in Figure 2, we compare the ROC curves of the ten combination methods obtained using the overall best-performing SimpleAverage*Minimum technique in Figure 1. The top-three combination methods including flDPnn + DisoLipPred + MoRFCHiBi$_{Light}$, flDPnn + MoRFCHiBi$_{Light}$, and flDPnn + Diso LipPred secure relatively similar AUCs (0.778, 0.773 and 0.767, respectively). Their ROC curves are better (i.e., positioned above) than the ROC curve of the best individual method, flDPnn, particularly for the low FPR region (Figure 1). This suggests that combining the disorder prediction with the prediction of MoRFs and/or lipid-binding regions results in a more accurate identification of putative MemMoRFs, confirming the validity of our hypothesis. Among the top three SimpleAverage*Minimum-based combinations, flDPnn + DisoLipPred + MoRFCHiBi$_{Light}$ is the best since it obtains the highest AUC value (Figure 2) and higher ROC curve in the low FPR region (Figure 1).

## CoMemMoRFPred

While the analysis based on Figures 1 and 2 shows that the flDPnn + DisoLipPred + MoRFCHi Bi$_{Light}$ combination that relies on the SimpleAverage*Minimum approach predicts MemMoRF residues with relatively high accuracy, we further investigate whether these residues compose sequence regions that mimic the distribution (i.e., number and sizes/lengths) of the native MemMoRF regions. Figure 3 reveals that the distribution of the predicted regions (WS 1, light-yellow line) is very different than the distribution for the native MemMoRFs (Native, dashed green line).

The predicted regions are much shorter and there are many of them, thus the difference between the two distributions is statistically significant (*p*-value $\leq$0.01). This issue can be fixed with window-based smoothing, i.e., the predicted propensity

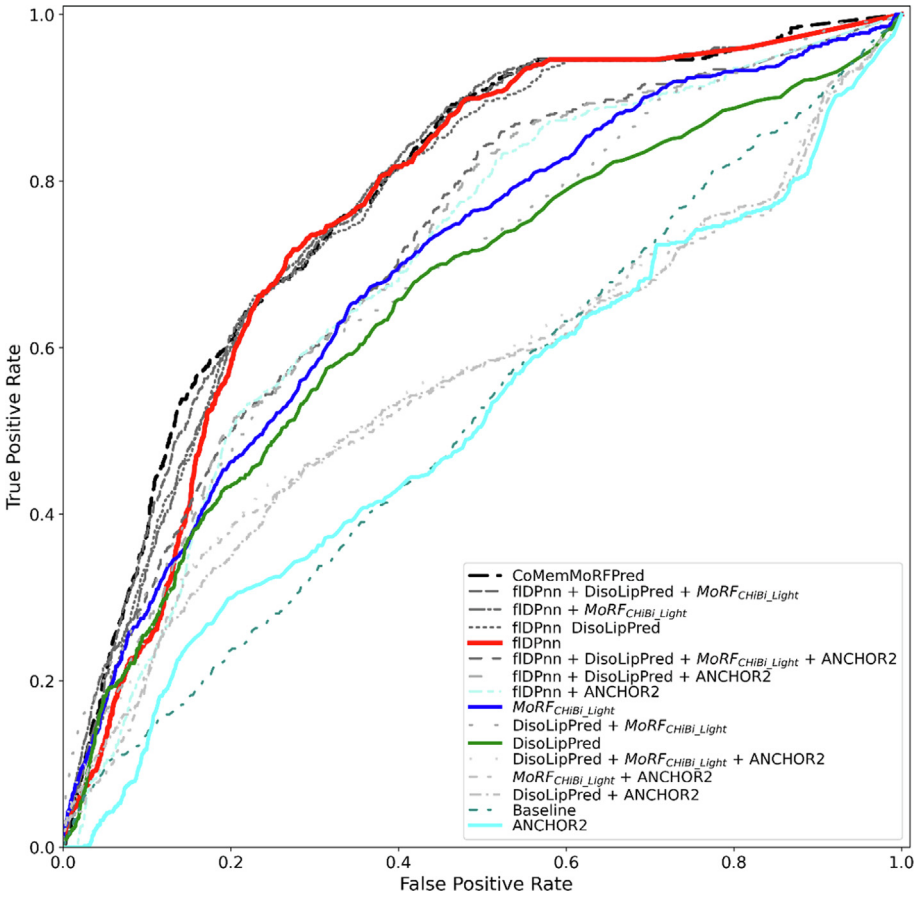


**Figure 1.** ROC curves computed on the test dataset. We include the baseline, the four selected predictors and ten combinations of these predictors that rely on the Simple Average*Minimum technique. Methods in the legend box are arranged in the descending order of their AUC values. The individual methods are shown using solid lines while the combination methods are in dashed lines.

values are averaged over a sliding sequence window to produce a new propensity for the residue located in the middle of the window. Smoothing was used in several related studies for the prediction of disorder and MoRFs.[81,82] We optimize the window size (WS) by comparing results for sizes = {3, 5, 7, 9, 11, 13, 15}, shown in Figure 3, where WS = 1 corresponds to predictions without smoothing. Distributions of the predicted MemMoRF regions with WS = 3, 13, and 15 are significantly different from the distribution for the native MemMoRFs ($p$-value $\leq$ 0.05). The differences are not statistically significant for WS = 5, 7, 9, and 11 ($p$-value > 0.05). Among these four window sizes, MemMoRFs generated using smoothing with WS = 9 (red colored line in Figure 3) are the closest to the native MemMoRFs (green colored line in Figure 3) in terms of size and number of MemMoRFs, i.e., this result converges to the distribution of the native MemMoRFs for longer region sizes.

Correspondingly, we formulate a new predictor of MemMoRFs based on combining flDPnn, DisoLipPred and MoRFCHiBi$_{Light}$ methods with smoothing using WS of 9. One of these tools (MoRFCHiBi$_{Light}$) was authored by another research group and we incorporate it into our solution with their permission. Table 1 compares this new computational tool, CoMemMoRFPred (**Co**mbined **MemMoRF Pred**ictor), with the other methods. CoMemMoRFPred secures the highest AUC = 0.785, rateAUC = 3.8 and F1$_{max}$ = 0.182. Its sensitivity at FPR = 0.05 is 0.19, which means that CoMemMoRFPred predicts the true positives at 0.19/0.05 = 3.8 higher rate than the false positives. These results are statistically better than the predictions of the best individual predictor, flDPnn (AUC = 0.761, rateAUC = 2.7 and F1$_{max}$ = 0.155; $p$-value $\leq$ 0.01). Moreover, the ROC curve of CoMemMoRFPred has the steepest slope for low FPR values (Figure 1), which suggests that it outperforms the other solutions where the amount of the predicted MemMoRFs does not substantially over-estimate the amount of native MemMoRFs. This is why CoMemMoRFPred obtains the highest rateAUC value in Table 1.

We further contrast predictions generated by a selected group of representative approaches that

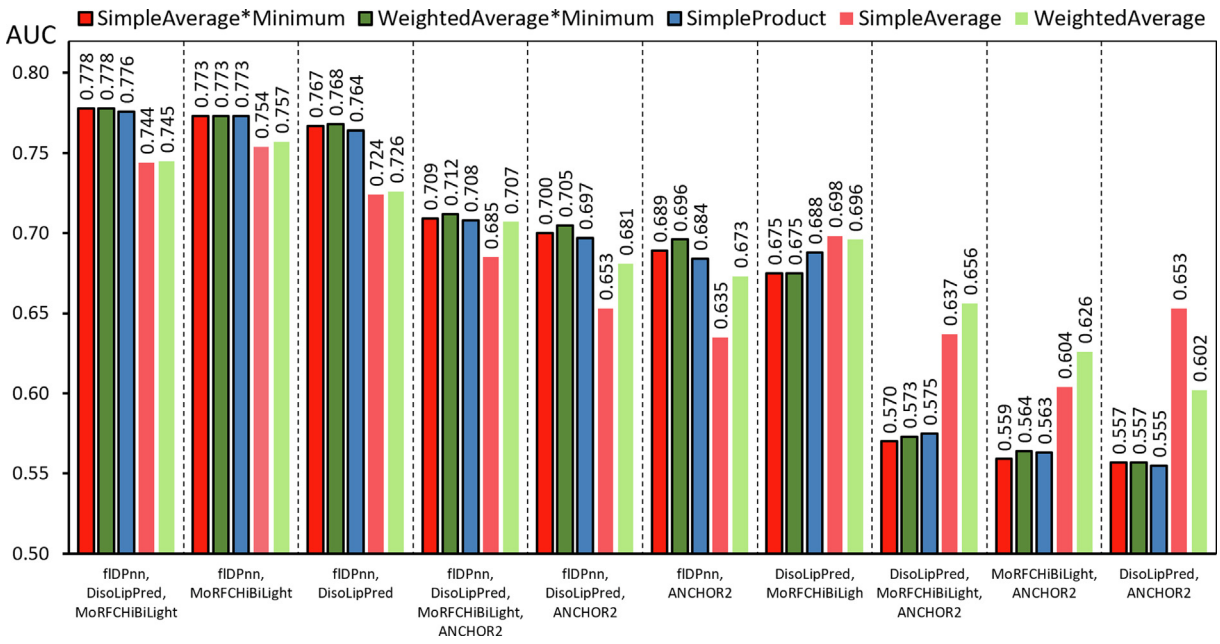


**Figure 2.** Comparison of the AUC values on the test dataset for the ten combinations of the four methods that are computed by using SimpleAverage (red bar with black border), WeightedAverage (green bar with black border), SimpleProduct (blue bar with black border), SimpleAverage*Minimum (light pink bar), and WeightedAverage*Minimum (light green bar) techniques. Results from the five techniques are grouped per specific combination, and they are sorted in descending order of their highest AUC value, which are shown at the top of the bars.

include the baseline, the best individual predictor fIDPnn, the best combined fIDPnn + DisoLipPred + MoRFCHiBi$_{Light}$ approach, and CoMemMoRFPred. In Figure 4, we compare distributions of propensity values produced by these approaches between the native MemMoRFs residues (in red) and the non-MemMoRF residues (in blue) from the test dataset. The results produced by the baseline predictor show no significant difference ($p$-value = 0.33). The progressively improved solutions, from the single predictor through the combination of methods to the inclusion of smoothing, register correspondingly smaller and smaller $p$-values. The overall best CoMemMoRFPred obtains $p$-value = $2.6 \times 10^{-140}$. These observations are in line with the results from Table 1 and demonstrate that CoMemMoRFPred offers accurate predictions of MemMoRFs. We conclude that this new tool provides high-quality residue-level propensities and also generates putative MemMoRF regions that closely replicate the number and sizes of the native MemMoRFs in the test dataset. This was accomplished by utilizing an innovative design that synergistically combines results from three existing tools that address predictions of different biophysical aspects of MemMoRFs and the application of the smoothing operator.

## CoMemMoRFPred web server

We provide CoMemMoRFPred as a freely available and easy to use web server at http://biomine.cs.vcu.edu/servers/CoMemMoRFPred. We implemented the front-end using HTML and JavaScript, while the back-end is based on PHP, Java, Python and the MySQL database. Users do not need to install any software beside a web browser.

Our web server features an easy to navigate input interface (Figure 5), where users need to provide a FASTA-formatted protein sequence as the input. While it is not mandatory, we encourage users to provide an email address where link to the results is emailed after the prediction is completed. Otherwise, the browser window must stay open during the prediction in order to access the web page with the results. Prediction is launched by clicking on the 'Run' button, which redirects to the processing page, followed by the results' page. The entire prediction process, which takes about 4 minutes for an average size protein sequence (about 300 residues long), runs automatically on the server side. The processing page provides updates on the process, including the current status (acceptance of the job, position in the queue of job, and processing the prediction) and completion.

The server provides results in two complementary formats including a text file, in which the data are in an easy-to-parse comma-separated format, and an interactive graphical output. The text file contains the raw scores and normalized (using the min–max approach) scores produced by CoMemMoRFPred, fIDPnn, MoRFCHiBi$_{Light}$ and DisoLipPred, and the binary predictions from

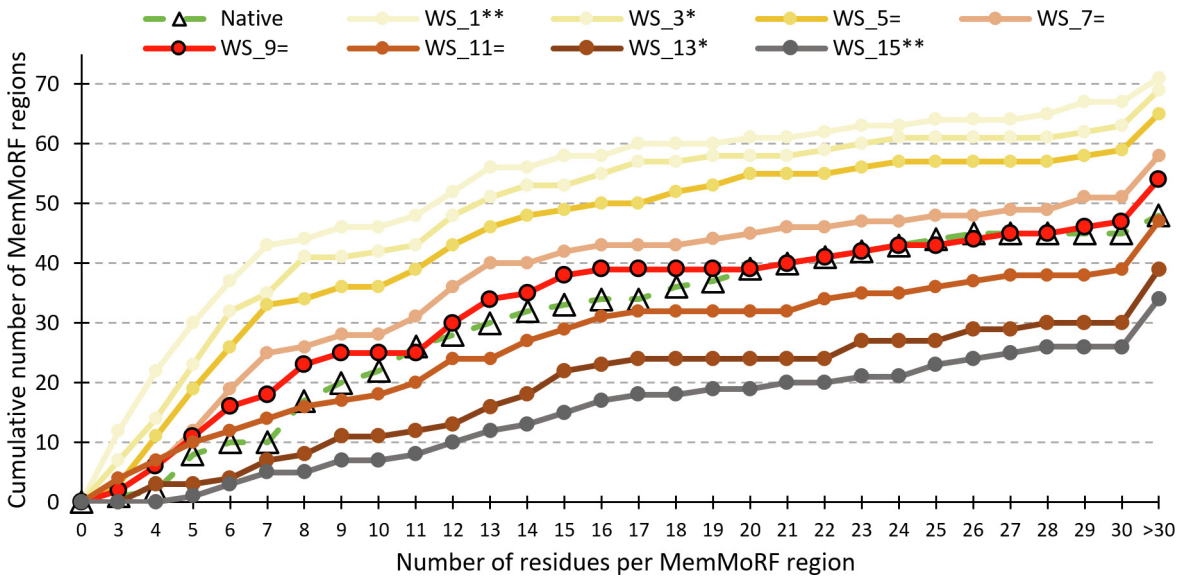


**Figure 3.** Distribution of MemMoRF region sizes for the native MemMoRFs and the putative MemMoRFs predicted by the flDPnn + DisoLipPred + MoRFCHiBi$_{Light}$ combination that relies on the SimpleAverage*Minimum approach that applies smoothing with variable window size (WS) values. The *x*-axis shows the MemMoRF size that we quantify with the number of residues. The *y*-axis is the cumulative count of MemMoRF regions with size that is defined by the *x*-axis value. Window sizes are shown in the figure legend and they vary between 1 (no window; light yellow) and 15 (gray). Statistical significance of differences between the distribution for the native MemMoRF regions and each of the predictions is given in the figure legend, where * and ** denote that the differences is statistically significant with *p*-value $\leq$ 0.01 and $\leq$0.05, respectively, and where = denotes differences that are not statistically significant (*p*-value > 0.05).

CoMemMoRFPred. The threshold we use to binarize the predictions corresponds to a low FPR = 0.05 (i.e., specificity = 0.95). The text file contains a header explaining the data format. The server generates the graphical output directly in the web browser window. This graphical panel visualizes the predicted MemMoRF regions based on the binary predictions from CoMemMoRFPred together with the corresponding normalized putative propensity values, which are shown along the input protein sequence. We also plot the propensity values generated by flDPnn, DisoLipPred and MoRFCHiBi$_{Light}$. The graphical panel is interactive with zoom-in and zoom-out features, panning along the horizontal axis (protein sequence), and ability to reset to the default view. Users can obtain details, such as numeric values of propensities and location of the putative MemMoRF regions, on the mouse hover. Moreover, we include an option to generate and download an image of the graphical panel.

## Case study

We explain CoMemMoRFPred's predictions using an example test protein, the cell division topological specificity factor MinE (UniProt accession number: P0A734). The Min protein system is crucial for cell division in *E. coli* and consists of the MinC, MinD and MinE proteins.[83] Briefly, dynamic oscillations of these proteins from one pole of the bacterial cell to the other pole determines placement of the central division septum during the *E. coli*'s cell division process.[84] As part of this process, a MemMoRF region at the N-terminus of the MinE sequence (positions 2–9) binds to the cell membrane inducing changes in the membrane topology and facilitating detachment of MinD from membranes during disassembly stage of the oscillation cycle.[85] Upon binding the membrane, this MemMoRF undergoes a conformational change by folding into an amphipathic helix which drives the deformation of the membrane.[86]

We use the MinE sequence to generate prediction of MemMoRF with CoMemMoRFPred (Figure 6), which predicts a MemMoRF region at the N-terminus (positions 4 and 15). This prediction closely matches the native MemMoRF (positions 2–9).[86] While the region predicted by CoMemMoRFPred extends beyond the native MemMoRF region, three basic amino acids that are part of this extension (R10, K11 and K12) were reported to undergo folding upon binding to the membrane.[85] This observation suggests that our prediction might be in fact correct. The MemMoRF predictions by CoMemMoRFPred are facilitated by the scores of the three methods that it combines: flDPnn, DisoLipPred and MoRFCHiBi$_{Light}$. The putative propensities produced by each of the three methods (gray lines in Figure 6) are relatively high

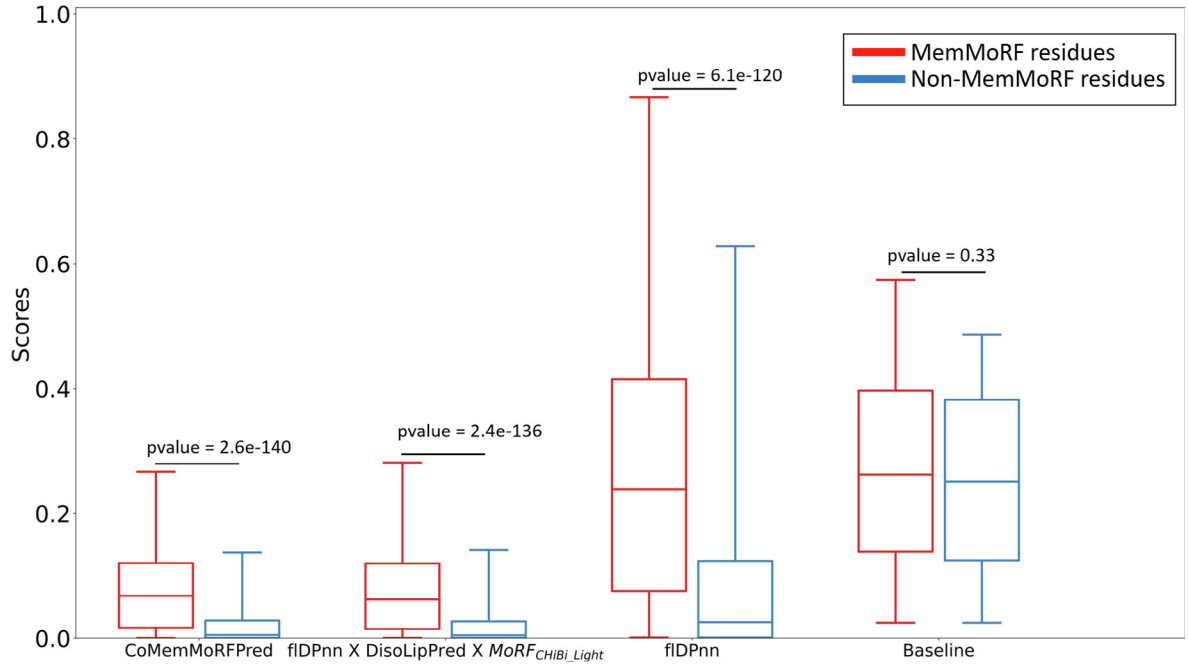


**Figure 4.** Boxplots that compare distributions of the predicted propensity values for the native MemMoRFs residues (in red) and the non-MemMoRF residues (in blue) from the test dataset. We compare the results produced by CoMemMoRFPred, the best combined method that does not apply smoothing (flDPnn + DisoLipPred + MoRFCHiBi$_{Light}$ that relies on the SimpleAverage*Minimum technique), the best individual method (flDPnn) and a random baseline. The whiskers denote 1st and 99th percentiles and the solid lines inside the boxes correspond to the median. We evaluate the statistical significance of differences between the two corresponding distributions using the Wilcoxon-rank sum test since all distributions are non-normal, as tested with the Anderson-Darling test at *p*-value of 0.05. The corresponding *p*-values are placed above each pair of box plots.

for the putative MemMoRF region (green horizontal bar in Figure 6). This observation reveals that the prediction of this regions is supported by high putative propensities for disorder (flDPnn), for MoRF (MoRFCHiBi$_{Light}$) and for lipid binding (DisoLipPred). Having high values for only two or one of these putative propensities is insufficient to predict MemMoRFs. These observations illustrate why the SimpleAverage*Minimum technique that we use to combine predictors provides strong results. Comparison of the red and green line plots in Figure 6 explains the smoothing performed by CoMemMoRFPred. The output of the CoMemMoRFPred (green line) is based on averaging and produces a smoother curve when compared to the red line produced by the flDPnn + DisoLipPred + MoRFCHiBi$_{Light}$ combination. Moreover, smoothing also potentially eliminates spurious predictions of very short MemMoRFs that might be generated by the flDPnn + DisoLipPred + MoRFCHiBi$_{Light}$ combination. We note that this combination generates such spurious prediction near the C-terminus (residues 83 and 84) where the propensities exceed the value of the threshold. Smoothing lowers the values of propensities in that region. More generally, predictions from CoMemMoRFPred for residues that have either high or low putative propensities (i.e., propensities that are substantially higher or lower than the threshold used by CoMemMoRFPred; green dotted line in Figure 6), should be considered as more accurate than the predictions associated with propensities near the threshold value. Our example demonstrates how to understand and interpret results produced by CoMemMoRFPred, and how these predictions are generated from the results output by flDPnn, DisoLipPred, and MoRFCHiBi$_{Light}$.

## Summary

IDRs interact with a variety of partner molecules, such as peptides, proteins, nucleic acids and lipids, by adopting partner-specific conformations upon binding. We focus on recently introduced MemMoRF regions, which are lipid-binding MoRFs that were found in the membrane-associated and transmembrane proteins. Motivated by functional importance and a relatively small number of the experimentally annotated MemMoRFs and the lack of computational tools for prediction of these regions, we apply an empirical approach to develop an accurate sequence-based predictor of these regions. We use an experimentally validated collection of low similarity proteins with MemMoRFs to select and combine results

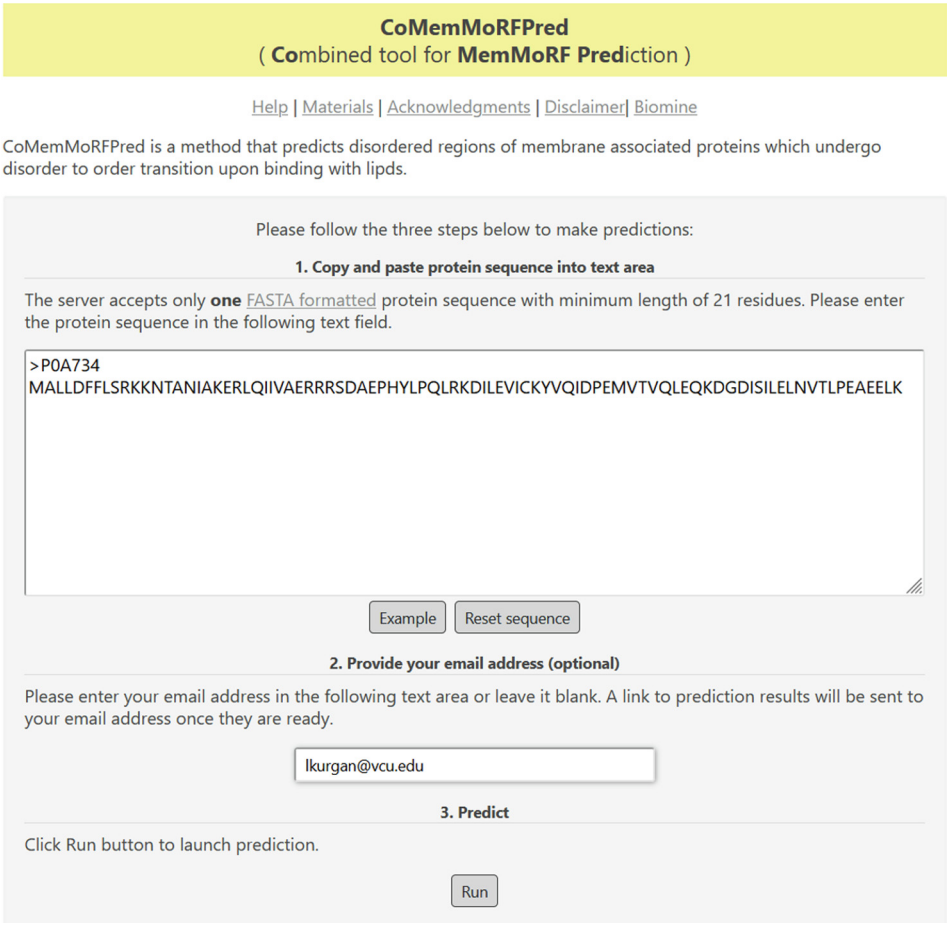


**Figure 5.** Input interface for the CoMemMoRFPred webserver at http://biomine.cs.vcu.edu/servers/CoMemMoRFPred.

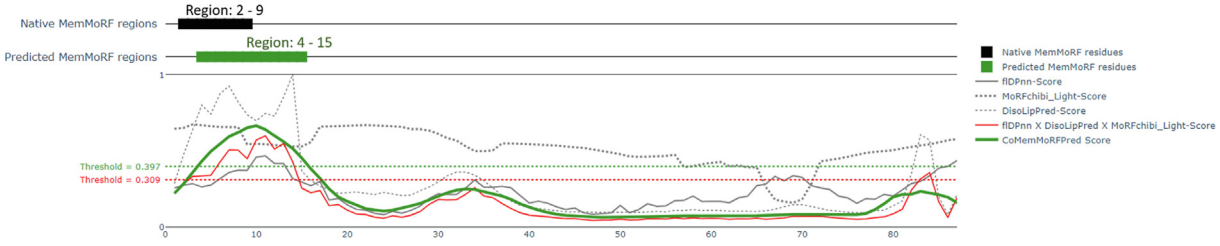


**Figure 6.** MemMoRF predictions for the MinE protein from *E. coli* (UniProt: P0A734). We show the native MemMoRF region (MemMoRF database ID: 9247) and the putative MemMoRF generated by CoMemMoRFPred as the black and green horizontal bars, respectively. The line plots show putative propensities from CoMemMoRFPred (solid green line), flDPnn + DisoLipPred + MoRFCHiBi$_{Light}$ combination that relies on the SimpleAverage*Minimum technique (solid red line), flDPnn (solid gray line), MoRFCHiBI$_{Light}$ (dotted dark gray line) and DisoLipPred (dotted light gray line). We visualize the thresholds that we use to generate the binary predictions for CoMemMoRFpred and flDPnn + DisoLipPred + MoRFCHiBi$_{Light}$ combinations as dotted horizontal lines in green and red color, respectively. Both thresholds correspond to the correct prediction rate computed on the test dataset (i.e., the number of putative MemMoRFs is set to be the same as the number of native MemMoRFs).

produced by a representative group of relevant state-of-the-art predictors of the intrinsic disorder (flDPnn), MoRFs (MoRFCHiBi$_{Light}$), and disordered lipid-binding regions (DisoLipPred). The resulting CoMemMoRFPred method generates accurate predictions of MemMoRF residues, with AUC_ROC = 0.785, F1_max = 0.182 and TPR = 0.36 at FPR of 10%.

Moreover, we show that the inclusion of the smoothing operator in CoMemMoRFPred results in the prediction of MemMoRF regions that closely resemble the distribution of the length of the native MemMoRF regions. We release CoMemMoRFPred as a convenient and freely available web server at http://biomine.cs.vcu.edu/servers/CoMemMoRFPred. This resource automates and performs the entire prediction process on the server side, and can be used without the need to install any software. It provides predictions in two complementary formats, text files that can be easily parsed to acquire the underlying raw predictions and an interactive graphical interface that visualizes the predictions directly in the web browser window. We also introduce a case study that demonstrates how to interpret and use the CoMemMoRFPred's predictions. Similar to the existing predictors that target other types of disordered binding regions,[48] our tool can be used to support efforts to identify MemMoRFs in a time and resource-efficient manner. We note that the currently characterized MemMoRF regions were found in the α-helical transmembrane proteins.[35] As part of our future work we plan to collect data on MemMoRFs in the β-barrel membrane proteins, expand our MemMoRF database,[35] and evaluate CoMemMoRFPred on these proteins. Ultimately, we believe that this work will aid exploration of the functional roles of MemMoRFs in cellular processes and pathologic phenomena related to membrane bilayer interactions.

## Funding

This work was funded in part by the National Science Foundation (DBI2146027 and IIS2125218), the Robert J. Mattauch Endowment funds to LK, and the NRDIO/NKFIH grant (K127961) to TH.

## CRediT authorship contribution statement

**Sushmita Basu:** Data curation, Formal analysis, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Tamás Hegedűs:** Conceptualization, Data curation, Funding acquisition, Investigation, Writing – original draft. **Lukasz Kurgan:** Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

## DATA AVAILABILITY

The data is available publically and the details how to access it are provided in the article.

## DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

*Received 5 July 2023;*
*Accepted 7 September 2023;*
*Available online 12 September 2023*

***Keywords***:
molecular recognition features;
intrinsic disorder;
lipid-binding;
prediction;
membrane proteins

## References

[1]. Dunker, A.K., Lawson, J.D., Brown, C.J., Williams, R.M., Romero, P., Oh, J.S., Oldfield, C.J., Campen, A.M., Ratliff, C.M., Hipps, K.W., et al., (2001). Intrinsically disordered protein. *J. Mol. Graph. Model.* **19**, 26–59. https://doi.org/10.1016/S1093-3263(00)00138-8.

[2]. Oldfield, C.J., Uversky, V.N., Dunker, A.K., Kurgan, L., (2019). Introduction to intrinsically disordered proteins and regions. *Dyn. Bind. Funct. Intrinsically Disord. Proteins*.

[3]. Lieutaud, P., Ferron, F., Uversky, A.V., Kurgan, L., Uversky, V.N., Longhi, S., (2016). How disordered is my protein and what is its disorder for? A guide through the "dark side" of the protein universe. *Intrinsically Disord Proteins* **4**, e1259708.

[4]. Habchi, J., Tompa, P., Longhi, S., Uversky, V.N., (2014). Introducing protein intrinsic disorder. *Chem. Rev.* **114**, 6561–6588. https://doi.org/10.1021/cr400514h.

[5]. Dunker, A.K., Babu, M.M., Barbar, E., Blackledge, M., Bondos, S.E., Dosztányi, Z., Dyson, H.J., Forman-Kay, J., Fuxreiter, M., Gsponer, J., et al., (2013). What's in a name? Why these proteins are intrinsically disordered. *Intrinsically Disord. Proteins* **1**, e24157.

[6]. Babu, M.M., (2016). The contribution of intrinsically disordered regions to protein function, cellular complexity, and human disease. *Biochem. Soc. Trans.* **44**, 1185–1200. https://doi.org/10.1042/BST20160172.

[7]. Gao, C., Ma, C., Wang, H., Zhong, H., Zang, J., Zhong, R., He, F., Yang, D., (2021). Intrinsic disorder in protein domains contributes to both organism complexity and clade-specific functions. *Sci. Rep.* **11**, 2985. https://doi.org/10.1038/s41598-021-82656-9.

[8]. Holguin-Cruz, J.A., Foster, L.J., Gsponer, J., (2022). Where protein structure and cell diversity meet. *Trends Cell Biol.* **32**, 996–1007. https://doi.org/10.1016/j.tcb.2022.04.004.

[9]. Peng, Z., Yan, J., Fan, X., Mizianty, M.J., Xue, B., Wang, K., Hu, G., Uversky, V.N., Kurgan, L., (2015). Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life. *Cell. Mol. Life Sci.* **72**, 137–151. https://doi.org/10.1007/s00018-014-1661-9.

[10]. Bondos, S.E., Dunker, A.K., Uversky, V.N., (2022). Intrinsically disordered proteins play diverse roles in cell

signaling. *Cell Commun. Signal* **20**, 20. https://doi.org/10.1186/s12964-022-00821-7.
[11]. Tantos, A., Kalmar, L., Tompa, P., (2015). The role of structural disorder in cell cycle regulation, related clinical proteomics, disease development and drug targeting. *Expert Rev. Proteomics* **12**, 221–233. https://doi.org/10.1586/14789450.2015.1042866.
[12]. Uversky, V.N., Oldfield, C.J., Dunker, A.K., (2005). Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling. *J. Mol. Recognit.* **18**, 343–384. https://doi.org/10.1002/jmr.747.
[13]. Vacic, V., Oldfield, C.J., Mohan, A., Radivojac, P., Cortese, M.S., Uversky, V.N., Dunker, A.K., (2007). Characterization of molecular recognition features, MoRFs, and their binding partners. *J. Proteome Res.* **6**, 2351–2366. https://doi.org/10.1021/pr0701411.
[14]. Verkhivker, G.M., Bouzida, D., Gehlhaar, D.K., Rejto, P. A., Freer, S.T., Rose, P.W., (2003). Simulating disorder-order transitions in molecular recognition of unstructured proteins: where folding meets binding. *PNAS* **100**, 5148–5153. https://doi.org/10.1073/pnas.0531373100.
[15]. Yan, J., Dunker, A.K., Uversky, V.N., Kurgan, L., (2016). Molecular recognition features (MoRFs) in three domains of life. *Mol. Biosyst.* **12**, 697–710. https://doi.org/10.1039/c5mb00640f.
[16]. Wang, C., Uversky, V.N., Kurgan, L., (2016). Disordered nucleiome: Abundance of intrinsic disorder in the DNA- and RNA-binding proteins in 1121 species from Eukaryota, Bacteria and Archaea. *Proteomics* **16**, 1486–1498. https://doi.org/10.1002/pmic.201500177.
[17]. Basu, S., Bahadur, R.P., (2016). A structural perspective of RNA recognition by intrinsically disordered proteins. *Cell. Mol. Life Sci.* **73**, 4075–4084. https://doi.org/10.1007/s00018-016-2283-1.
[18]. Calabretta, S., Richard, S., (2015). Emerging Roles of Disordered Sequences in RNA-Binding Proteins. *Trends Biochem. Sci* **40**, 662–672. https://doi.org/10.1016/j.tibs.2015.08.012.
[19]. Jamecna, D., Antonny, B., (2021). Intrinsically disordered protein regions at membrane contact sites. *Biochim. Biophys. Acta Mol. Cell Biol. Lipids* **1866**, https://doi.org/10.1016/j.bbalip.2021.159020 159020.
[20]. Kjaergaard, M., Kragelund, B.B., (2017). Functions of intrinsic disorder in transmembrane proteins. *Cell. Mol. Life Sci.* **74**, 3205–3224. https://doi.org/10.1007/s00018-017-2562-5.
[21]. Wu, Z., Hu, G., Yang, J., Peng, Z., Uversky, V.N., Kurgan, L., (2015). In various protein complexes, disordered protomers have large per-residue surface areas and area of protein- DNA- and RNA-binding interfaces. *FEBS Lett* **589**, 2561–2569. https://doi.org/10.1016/j.febslet.2015.08.014.
[22]. Hsu, W.L., Oldfield, C.J., Xue, B., Meng, J., Huang, F., Romero, P., Uversky, V.N., Dunker, A.K., (2013). Exploring the binding diversity of intrinsically disordered proteins involved in one-to-many binding. *Protein Sci.* **22**, 258–273. https://doi.org/10.1002/pro.2207.
[23]. Hu, G., Wu, Z., Uversky, V.N., Kurgan, L., (2017). Functional analysis of human hub proteins and their interactors involved in the intrinsic disorder-enriched interactions. *Int. J. Mol. Sci.* **18** https://doi.org/10.3390/ijms18122761.
[24]. Dunker, A.K., Brown, C.J., Lawson, J.D., Iakoucheva, L. M., Obradovic, Z., (2002). Intrinsic disorder and protein function. *Biochemistry* **41**, 6573–6582.
[25]. Dyson, H.J., Wright, P.E., (2005). Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.* **6**, 197–208. https://doi.org/10.1038/nrm1589.
[26]. Uversky, V.N., (2013). Intrinsic Disorder-based Protein Interactions and their Modulators. *Curr Pharm Design* **19**, 4191–4213.
[27]. Tompa, P., Fuxreiter, M., (2008). Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. *Trends Biochem. Sci* **33**, 2–8. https://doi.org/10.1016/j.tibs.2007.10.003.
[28]. Mohan, A., Oldfield, C.J., Radivojac, P., Vacic, V., Cortese, M.S., Dunker, A.K., Uversky, V.N., (2006). Analysis of molecular recognition features (MoRFs). *J. Mol. Biol.* **362**, 1043–1059. https://doi.org/10.1016/j.jmb.2006.07.087.
[29]. Dyson, H.J., (2012). Roles of intrinsic disorder in protein-nucleic acid interactions. *Mol. Biosyst.* **8**, 97–104. https://doi.org/10.1039/c1mb05258f.
[30]. Shammas, S.L., (2017). Mechanistic roles of protein disorder within transcription. *Curr. Opin. Struct. Biol.* **42**, 155–161. https://doi.org/10.1016/j.sbi.2017.02.003.
[31]. Balcerak, A., Trebinska-Stryjewska, A., Konopinski, R., Wakula, M., Grzybowska, E.A., (2019). RNA-protein interactions: disorder, moonlighting and junk contribute to eukaryotic complexity. *Open Biol.* **9**, https://doi.org/10.1098/rsob.190096 190096.
[32]. Burgi, J., Xue, B., Uversky, V.N., van der Goot, F.G., (2016). Intrinsic disorder in transmembrane proteins: roles in signaling and topology prediction. *Plos One* **11**, https://doi.org/10.1371/journal.pone.0158594 e0158594.
[33]. Tusnady, G.E., Dobson, L., Tompa, P., (1848). Disordered regions in transmembrane proteins. *Bba-Biomembranes* **2015**, 2839–2848. https://doi.org/10.1016/j.bbamem.2015.08.002.
[34]. Cornish, J., Chamberlain, S.G., Owen, D., Mott, H.R., (2020). Intrinsically disordered proteins and membranes: a marriage of convenience for cell signalling? *Biochem. Soc. Trans.* **48**, 2669–2689. https://doi.org/10.1042/BST20200467.
[35]. Csizmadia, G., Erdos, G., Tordai, H., Padanyi, R., Tosatto, S., Dosztanyi, Z., Hegedus, T., (2021). The MemMoRF database for recognizing disordered protein regions interacting with cellular membranes. *Nucleic Acids Res.* **49**, D355–D360. https://doi.org/10.1093/nar/gkaa954.
[36]. Zhao, B., Kurgan, L., (2022). Compositional bias of intrinsically disordered proteins and regions and their predictions. *Biomolecules* **12** https://doi.org/10.3390/biom12070888.
[37]. Yan, J., Cheng, J., Kurgan, L., Uversky, V.N., (2020). Structural and functional analysis of “non-smelly” proteins. *Cell. Mol. Life Sci.* **77**, 2423–2440. https://doi.org/10.1007/s00018-019-03292-1.
[38]. Campen, A., Williams, R.M., Brown, C.J., Meng, J., Uversky, V.N., Dunker, A.K., (2008). TOP-IDP-scale: a new amino acid scale measuring propensity for intrinsic disorder. *Protein Pept. Lett.* **15**, 956–963.
[39]. Zhao, B., Kurgan, L., (2023). Machine learning for intrinsic disorder prediction. *In: Machine Learning in*

*Bioinformatics of Protein Sequences*. https://doi.org/10.1142/9789811258589_0008 pp. 205-236.

[40]. Zhao, B., Kurgan, L., (2021). Surveying over 100 predictors of intrinsic disorder in proteins. *Expert Rev. Proteomics* **18**, 1019–1029. https://doi.org/10.1080/14789450.2021.2018304.

[41]. Liu, Y., Wang, X., Liu, B., (2019). A comprehensive review and comparison of existing computational methods for intrinsically disordered protein and region prediction. *Brief. Bioinform.* **20**, 330–346. https://doi.org/10.1093/bib/bbx126.

[42]. Meng, F., Uversky, V.N., Kurgan, L., (2017). Comprehensive review of methods for prediction of intrinsic disorder and its molecular functions. *Cell. Mol. Life Sci.* **74**, 3069–3090. https://doi.org/10.1007/s00018-017-2555-4.

[43]. Punta, M., Simon, I., Dosztanyi, Z., (2015). Prediction and analysis of intrinsically disordered proteins. *Methods Mol. Biol.* **1261**, 35–59. https://doi.org/10.1007/978-1-4939-2230-7_3.

[44]. Deng, X., Eickholt, J., Cheng, J., (2012). A comprehensive overview of computational protein disorder prediction methods. *Mol. Biosyst.* **8**, 114–121. https://doi.org/10.1039/c1mb05207a.

[45]. He, B., Wang, K., Liu, Y., Xue, B., Uversky, V.N., Dunker, A.K., (2009). Predicting intrinsic disorder in proteins: an overview. *Cell Res.* **19**, 929–949. https://doi.org/10.1038/cr.2009.87.

[46]. Kurgan, L., (2022). Resources for computational prediction of intrinsic disorder in proteins. *Methods* **204**, 132–141. https://doi.org/10.1016/j.ymeth.2022.03.018.

[47]. Zhao, B., Kurgan, L., (2022). Deep learning in prediction of intrinsic disorder in proteins. *Comput. Struct. Biotechnol. J.* **20**, 1286–1294. https://doi.org/10.1016/j.csbj.2022.03.003.

[48]. Basu, S., Kihara, D., Kurgan, L., (2023). Computational prediction of disordered binding regions. *Comput. Struct. Biotechnol. J.* **21**, 1487–1497. https://doi.org/10.1016/j.csbj.2023.02.018.

[49]. McLaughlin, S., (1989). The electrostatic properties of membranes. *Annu. Rev. Biophys. Biophys. Chem.* **18**, 113–136. https://doi.org/10.1146/annurev.bb.18.060189.000553.

[50]. von Heijne, G., (1992). Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. *J. Mol. Biol.* **225**, 487–494. https://doi.org/10.1016/0022-2836(92)90934-c.

[51]. Necci, M., Piovesan, D., Predictors, C., DisProt, C., Tosatto, S.C.E., (2021). Critical assessment of protein intrinsic disorder prediction. *Nat. Methods* **18**, 472–481. https://doi.org/10.1038/s41592-021-01117-3.

[52]. Lang, B., Babu, M.M., (2021). A community effort to bring structure to disorder. *Nat. Methods* **18**, 454–455. https://doi.org/10.1038/s41592-021-01123-5.

[53]. Katuwawala, A., Oldfield, C.J., Kurgan, L., (2020). Accuracy of protein-level disorder predictions. *Brief. Bioinform.* **21**, 1509–1522. https://doi.org/10.1093/bib/bbz100.

[54]. Katuwawala, A., Kurgan, L., (2020). Comparative assessment of intrinsic disorder predictions with a focus on protein and nucleic acid-binding proteins. *Biomolecules* **10** https://doi.org/10.3390/biom10121636.

[55]. Necci, M., Piovesan, D., Dosztanyi, Z., Tompa, P., Tosatto, S.C.E., (2018). A comprehensive assessment of long intrinsic protein disorder from the DisProt database. *Bioinformatics* **34**, 445–452. https://doi.org/10.1093/bioinformatics/btx590.

[56]. Walsh, I., Giollo, M., Di Domenico, T., Ferrari, C., Zimmermann, O., Tosatto, S.C., (2015). Comprehensive large-scale assessment of intrinsic protein disorder. *Bioinformatics* **31**, 201–208. https://doi.org/10.1093/bioinformatics/btu625.

[57]. Peng, Z.L., Kurgan, L., (2012). Comprehensive comparative assessment of in-silico predictors of disordered regions. *Curr. Protein Pept. Sci.* **13**, 6–18.

[58]. Hu, G., Katuwawala, A., Wang, K., Wu, Z., Ghadermarzi, S., Gao, J., Kurgan, L., (2021). flDPnn: Accurate intrinsic disorder prediction with putative propensities of disorder functions. *Nat. Commun.* **12**, 4438. https://doi.org/10.1038/s41467-021-24773-7.

[59]. Mirabello, C., Wallner, B., (2019). rawMSA: End-to-end deep learning using raw multiple sequence alignments. *PLoS One* **14** https://doi.org/10.1371/journal.pone.0220182.

[60]. Walsh, I., Martin, A.J., Di Domenico, T., Tosatto, S.C., (2012). ESpritz: accurate and fast prediction of protein disorder. *Bioinformatics* **28**, 503–509. https://doi.org/10.1093/bioinformatics/btr682.

[61]. Orlando, G., Raimondi, D., Codice, F., Tabaro, F., Vranken, W., (2022). Prediction of disordered regions in proteins with recurrent neural networks and protein dynamics. *J. Mol. Biol.* **434**, https://doi.org/10.1016/j.jmb.2022.167579 167579.

[62]. Hanson, J., Paliwal, K.K., Litfin, T., Zhou, Y., (2020). SPOT-Disorder 2: Improved protein intrinsic disorder prediction by ensembled deep learning. *Genom. Proteom. Bioinf.*. https://doi.org/10.1016/j.gpb.2019.01.004.

[63]. Meszaros, B., Erdos, G., Dosztanyi, Z., (2018). IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.* **46**, W329–W337. https://doi.org/10.1093/nar/gky384.

[64]. Peng, Z.L., Wang, C., Uversky, V.N., Kurgan, L., (2017). Prediction of disordered RNA, DNA, and protein binding regions using DisoRDPbind. *Predict. Protein Second. Struct.* **1484**, 187–203. https://doi.org/10.1007/978-1-4939-6406-2_14.

[65]. Peng, Z., Kurgan, L., (2015). High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder. *Nucleic Acids Res.* **43**, e121.

[66]. Malhis, N., Jacobson, M., Gsponer, J., (2016). MoRFchibi SYSTEM: software tools for the identification of MoRFs in protein sequences. *Nucleic Acids Res.* **44**, W488–W493. https://doi.org/10.1093/nar/gkw409.

[67]. Malhis, N., Wong, E.T., Nassar, R., Gsponer, J., (2015). Computational identification of MoRFs in protein sequences using hierarchical application of Bayes rule. *PLoS One* **10**, e0141603. https://doi.org/10.1371/journal.pone.0141603.

[68]. Dosztanyi, Z., Meszaros, B., Simon, I., (2009). ANCHOR: web server for predicting protein binding regions in disordered proteins. *Bioinformatics* **25**, 2745–2746. https://doi.org/10.1093/bioinformatics/btp518.

[69]. Katuwawala, A., Zhao, B., Kurgan, L., (2022). DisoLipPred: accurate prediction of disordered lipid-binding residues in protein sequences with deep recurrent networks and transfer learning. *Bioinformatics*

**38**, 115–124. https://doi.org/10.1093/bioinformatics/btab640.

[70]. Dobson, L., Tusnady, G.E., (2021). MemDis: Predicting disordered regions in transmembrane proteins. *Int. J. Mol. Sci.* **22**, https://doi.org/10.3390/ijms222212270 12270.

[71]. Peng, Z., Li, Z., Meng, Q., Zhao, B., Kurgan, L., (2023). CLIP: accurate prediction of disordered linear interacting peptides from protein sequences using co-evolutionary information. *Brief. Bioinform.* **24** https://doi.org/10.1093/bib/bbac502.

[72]. Zhang, F., Li, M., Zhang, J., Shi, W., Kurgan, L., (2023). DeepPRObind: Modular deep learner that accurately predicts structure and disorder-annotated protein binding residues. *J. Mol. Biol.*, 167945. https://doi.org/10.1016/j.jmb.2023.167945.

[73]. Fu, L., Niu, B., Zhu, Z., Wu, S., Li, W., (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152. https://doi.org/10.1093/bioinformatics/bts565.

[74]. Fawcett, T., (2006). An introduction to ROC analysis. *Pattern Recogn. Lett.* **27**, 861–874. https://doi.org/10.1016/j.patrec.2005.10.010.

[75]. Zhang, J., Ma, Z., Kurgan, L., (2019). Comprehensive review and empirical analysis of hallmarks of DNA-, RNA- and protein-binding residues in protein chains. *Brief. Bioinform.* **20**, 1250–1268. https://doi.org/10.1093/bib/bbx168.

[76]. Yan, J., Kurgan, L., (2017). DRNApred, fast sequence-based method that accurately predicts and discriminates DNA- and RNA-binding residues. *Nucleic Acids Res.* **45**, e84.

[77]. Su, H., Liu, M., Sun, S., Peng, Z., Yang, J., (2019). Improving the prediction of protein-nucleic acids binding residues via multiple sequence profiles and the consensus of complementary methods. *Bioinformatics* **35**, 930–936. https://doi.org/10.1093/bioinformatics/bty756.

[78]. Zhang, J., Kurgan, L., (2019). SCRIBER: accurate and partner type-specific prediction of protein-binding residues from proteins sequences. *Bioinformatics* **35**, i343–i353. https://doi.org/10.1093/bioinformatics/btz324.

[79]. Wang, K., Hu, G., Wu, Z., Su, H., Yang, J., Kurgan, L., (2020). Comprehensive survey and comparative assessment of RNA-binding residue predictions with analysis by RNA type. *Int. J. Mol. Sci.* **21**, 6879.

[80]. Zhang, J., Ghadermarzi, S., Katuwawala, A., Kurgan, L., (2021). DNAgenie: Accurate prediction of DNA-type-specific binding residues in protein sequences. *Brief. Bioinform.* **22** https://doi.org/10.1093/bib/bbab336.

[81]. Erdos, G., Pajkos, M., Dosztanyi, Z., (2021). IUPred3: Prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of evolutionary conservation. *Nucleic Acids Res.* **49**, W297–W303. https://doi.org/10.1093/nar/gkab408.

[82]. Fang, C., Noguchi, T., Tominaga, D., Yamana, H., (2013). MFSPSSMpred: Identifying short disorder-to-order binding regions in disordered proteins based on contextual local evolutionary conservation. *BMC Bioinf.* **14**, 300. https://doi.org/10.1186/1471-2105-14-300.

[83]. de Boer, P.A., Crossley, R.E., Rothfield, L.I., (1988). Isolation and properties of minB, a complex genetic locus involved in correct placement of the division site in *Escherichia coli*. *J. Bacteriol.* **170**, 2106–2112. https://doi.org/10.1128/jb.170.5.2106-2112.1988.

[84]. Lutkenhaus, J., (2008). Min oscillation in bacteria. *Adv. Exp. Med. Biol.* **641**, 49–61. https://doi.org/10.1007/978-0-387-09794-7_4.

[85]. Hsieh, C.W., Lin, T.Y., Lai, H.M., Lin, C.C., Hsieh, T.S., Shih, Y.L., (2010). Direct MinE-membrane interaction contributes to the proper localization of MinDE in *E. coli*. *Mol. Microbiol.* **75**, 499–512. https://doi.org/10.1111/j.1365-2958.2009.07006.x.

[86]. Shih, Y.L., Huang, K.F., Lai, H.M., Liao, J.H., Lee, C.S., Chang, C.M., Mak, H.M., Hsieh, C.W., Lin, C.C., (2011). The N-terminal amphipathic helix of the topological specificity factor MinE is associated with shaping membrane curvature. *Plos One* **6**, https://doi.org/10.1371/journal.pone.0021425 e21425.