CLIP: Accurate prediction of disordered linear interacting peptides (LIPs) from protein sequences using co-evolutionary information

Zhenling Peng[1, 2*], Zixia Li[3], Qiaozhen Meng[4], Bi Zhao[5] and Lukasz Kurgan[5*]
[1]Research Center for Mathematics and Interdisciplinary Sciences, Shandong University, Qingdao, 266237, China.
[2]Frontier Science Center for Nonlinear Expectations, Ministry of Education, Qingdao, 266237, China.
[3]Center for Applied Mathematics, Tianjin University, Tianjin, 300072, China.
[4]College of Intelligence and Computing, Tianjin University, Tianjin, 300072, China.
[5]Department of Computer Science, Virginia Commonwealth University, Richmond, VA 23284, USA.
[*]Corresponding authors: Zhenling Peng at zhenling@email.sdu.edu.cn and Lukasz Kurgan at lkurgan@vcu.edu

Abstract

One of key features of intrinsically disordered regions (IDRs) is facilitation of protein-protein and protein-nucleic acids interactions. These disordered binding regions include molecular recognition features (MoRFs), short linear motifs (SLiMs) and longer binding domains. Vast majority of current predictors of disordered binding regions target MoRFs, with a handful of methods that predict SLiMs and disordered protein-binding domains. A new and broader class of disordered binding regions, linear interacting peptides (LIPs), was introduced recently and applied in the MobiDB resource. LIPs are segments in protein sequences that undergo disorder-to-order transition upon binding to a protein or a nucleic acid, and they cover MoRFs, SLiMs and disordered protein-binding domains. While current predictors of MoRFs and disordered protein-binding regions could be used to identify some LIPs, there are no dedicated sequence-based predictors of LIPs. To this end, we introduce CLIP, a new predictor of LIPs that utilizes robust logistic regression model to combine three complementary types of inputs: co-evolutionary information derived from multiple sequence alignments, physicochemical profiles, and disorder predictions. Ablation analysis suggests that the co-evolutionary information is particularly useful for this prediction and that combining the three inputs provides substantial improvements when compared to using these inputs individually. Comparative empirical assessments using low-similarity test datasets reveal that CLIP secures AUC of 0.8 and substantially improves over the results produced by the closest current tools that predict MoRFs and disordered protein-binding regions. The webserver of CLIP is freely available at http://biomine.cs.vcu.edu/servers/CLIP/ and the standalone code can be downloaded from http://yanglab.qd.sdu.edu.cn/download/CLIP/.

Zhenling Peng is a Professor at the Shandong University. Her research includes studies of intrinsically disordered proteins, protein structure and function prediction, and applications of machine learning in bioinformatics problems.

Zixia Li is a graduate student in the Center for Applied Mathematics at the Tianjin University. Her research focuses on the sequence-based prediction of functions of intrinsically disordered regions.

Qiaozhen Meng is a graduate student in the College of Intelligence and Computing at the Tianjin University. Her research focuses on the ab-initio prediction of protein structure.

Bi Zhao earned PhD from the University of South Florida in 2019, completed postdoctoral studies in the Computer Science department at the Virginia Commonwealth University in 2022, and currently is an Assistant Professor and Director of the Computational Core at the University of South Florida. She spearheaded development of multiple bioinformatics resources for protein disorder and disorder function prediction.

Lukasz Kurgan is a Fellow of AIMBE and AAIA, Member of European Academy of Sciences and Arts, and Robert J. Mattauch Endowed Professor of Computer Science at the Virginia Commonwealth University. His research encompasses structural and functional characterization of proteins. He serves on the Editorial Board of *Bioinformatics* and as Associate Editor-in-Chief of *Biomolecules*. More details at http://biomine.cs.vcu.edu/.

Key Points

- Linear interacting peptides (LIPs) are a new and broad class of disordered binding regions
- Current predictors of disordered binding regions offer modest levels of performance
- CLIP is a new and accurate sequence-based predictor of LIPs
- Co-evolutionary information is useful to predicts disordered binding regions
- CLIP's webserver is available at http://biomine.cs.vcu.edu/servers/CLIP/

1 Introduction

Intrinsically disordered regions (IDRs) are segments in the protein sequence that lack a stable equilibrium structure under physiological conditions [1-3]. Proteins with IDRs carry out numerous cellular functions and were shown to be prevalent in nature [4-13]. One of their key functions is facilitation of protein-protein and protein-nucleic acids interactions. The binding IDRs typically undergo disorder-to-order transitions concomitant with binding [14-20] and their structural flexibility allows them to interact with multiple partners by folding into different conformations [14, 21, 22]. These interactions were investigated from the protein sequence and structural points of view, leading to the discovery of multiple classes of binding IDRs. They include molecular recognition features (MoRFs) that are defined as short disordered segments (5 to 25 consecutive residues) that undergo coupled binding and folding when interacting with proteins and peptides [23-25]; short linear motifs (SLiMs), which are typically short sequence segments that are defined by regular expressions and that include residues directly interacting with binding partners [26-29]; protean segments (ProSs) that are specific to the IDEAL database [30, 31]; and long (>30 residues in length) disordered protein-binding domains [32]. Since propensity for disorder is intrinsic to the underlying sequence [33, 34], it should be possible to predict disordered binding regions from the protein sequences. Importance of prediction of disordered binding regions was emphasized in the recently completed Critical Assessment of protein Intrinsic Disorder (CAID) experiment, where this category of predictions was introduced for the first time [35]. Moreover, recent reviews identify about 20 computational predictors of the disordered binding regions [36, 37]. About 75% of them target prediction of MoRFs [36, 37]. Example MoRF predictors include (in chronological order) MoRFpred [38], MoRFchibi [39-41], OPAL [42], OPAL+ [43], MoRFMLP [44], SPOT-MoRF [45], and MoRF$_{\text{CNN}}$ [46]. While being the largest group of predictors, they are limited to the prediction of these short protein-binding regions. SLiMs can be identified using either databases of these motifs, such as ELM [47, 48], or using motif-finding algorithms, such as MnM [49-51], SLiMSearch [52, 53] and QuasiMotifFinder [54]. However, these methods may generate a substantial number of false positive predictions. Similarly, ProSs can be found in the IDEAL database [31], and to the best of our knowledge there are no predictors that are specific to these binding IDRs. Finally, four methods predict protein-binding IDRs that are not explicitly limited in their segment length: ANCHOR [55], ANCHOR2 [56], DisoRDPbind [57, 58] and DeepDISObind [59]. ANCHOR and ANCHOR2 are constrained to the prediction of the protein-binding IDRs while DisoRDPbind and DeepDISObind also predict IDRs that interact with nucleic acids. The abovementioned predictors of interacting IDRs rely on a wide range of models that cover various machine learning solutions, such as support vector machines [38, 40, 42, 43], neural networks [44-46, 59], Bayesian learning [40, 41], and regression [58]; regular expressions that are particularly useful to identify SLiMs [49, 52, 54]; and energy estimations based on statistical potentials [55, 56]. Among a subset of these methods that were recently evaluated in the CAID experiment [35], ANCHOR2 [56], DisoRDPbind [58] and MoRFchibi-light [40] secure the best performance when evaluated on prediction of binding IDRs.

Linear interacting peptides (LIPs) are a recently introduced class of binding IDRs that is annotated in the popular MobiDB database [60, 61]. LIPs are segments in a protein sequence that undergo disorder-to-order transitions upon binding to proteins and nucleic acids. LIPs, MoRFs and SLiMs focus on different categories of disordered binding regions where LIPs are a broader class that can be seen as a superset of MoRFs and SLiMs. For instance, LIPs are not limited in the segment length, like MoRFs and majority of SLiMs are, and may bind proteins and nucleic acids, unlike MoRFs that interact with proteins. LIPs can be extracted from the structures of the protein-protein and protein-nucleic acids complexes using FLIPPER [62]. While current predictors of MoRFs and disordered protein-binding regions could be used to identify some LIPs, there are no dedicated methods that predict LIPs from protein sequences. Notably, authors of MobiDB declared that defining LIPs "*may drive the development of a new generation of predictors for functional intrinsic disorder regions*" [60]. Moreover, recent CAID assessment concludes that "*disordered binding regions remain hard to predict*" and suggests that new and more accurate predictors are needed [35]. The development of new predictive methods requires availability of a large amount of experimentally annotated binding IDRs that are required to train and test predictive models. Experimental annotations of LIPs are available in several databases including PDB [63] where LIPs can be extracted using FLIPPER, ELM [47, 48], IDEAL [31], DisProt [64, 65], DIBS [66], and MFIB [67]. MobiDB conveniently combines binding data from these resources, providing access to several thousand LIPs [68]. We note that IDRs interact with other types of molecules beyond proteins and nucleic acids, such as lipids, metal ions and small molecules [36, 64]. However, methods for accurate prediction of disordered lipid-binding IDRs are already available [69, 70] while there are too few annotated interacting IDRs for the other ligand types to properly train and assess predictive models.

Given the availability of experimental data, observations from CAID and the lack of dedicated sequence-based predictors of LIPs, we introduce CLIP, an innovative and accurate predictor of LIPs from protein sequences. CLIP relies on three types of inputs that include: 1) co-evolutionary information which identifies conserved residues that co-evolve together (in our case these residues are collectively involved in binding); 2) sequence-based disorder prediction that allows our model to differentiate between structured and disordered regions; and 3) a collection of several relevant physiochemical properties of amino acids, such as hydrophobicity and free energy, that supports finding binding regions among the disordered residues. The use of the co-evolutionary data is motivated by a recent article that shows importance of evolutionary couplings for the detection of distinct structural states of IDRs [71], which is one of hallmarks of LIPs that fold into distinct conformations upon binding. We empirically demonstrate that the co-evolutionary information is a strong predictive input and that combining these inputs together results in substantial improvements in predictive performance. We empirically test CLIP-generated predictions on an independent test dataset (i.e., dataset that shares low sequence similarity with training and validation proteins) and compare them against the results produced by the best predictors from the CAID experiment and a selection of other modern predictors of MoRFs.

2 Materials and methods

2.1 Datasets

We use MobiDB 3.0 to collect the 2,303 proteins that include manually curated LIPs [61, 68]. Following the annotation protocol from [72], we first cluster these 2,303 sequences using CD-HIT with 100% sequence identity [73]. We set the longest chain in a given cluster as a representative sequence and transfer annotations of LIPs in a given cluster into this chain. Altogether, this procedure introduces 69 additional LIP residues when compared to the annotation without the transfer, i.e., 0.19% increase in the amount of LIP annotations; Supplementary Materials provide further details. Second, we group the resulting 2,285 sequences into 1,440 clusters by applying CD-HIT with 25% sequence identity threshold, and select the longest sequence to represent each cluster. We divide these sequences into a training dataset TR1000 and a test dataset TE440 at random. The TR1000 dataset is composed of 1,000 proteins that have 1,380 LIPs (average of 1.38 LIP per protein) and 24,821 amino acids in the LIP regions. The remaining 440 proteins, which cover 612 LIPs (average of 1.39 LIP per protein) and 11,994 residues in the LIP regions, constitute the TE440 dataset. Based on the second clustering, the training and test proteins share low (<25%) similarity.

Given the large number of MoRF predictors and the fact that MoRFs and LIPs share the disorder-to-order transition aspect, we perform an additional test that focuses on the disordered proteins with MoRF regions. We use the experimentally validated EXP53 dataset, which consist of 53 sequences with 2,432 MoRF residues and was developed by Malhis *et al.* [74]. We apply CD-HIT [73] to cluster sequences from the TR1000 dataset together with sequence from the EXP53 dataset. Next, we remove sequences from EXP53 that share the same clusters with the sequences from the TR1000 training dataset. The remaining collection of 25 MoRF-including sequences, named EXP25, is dissimilar to the training proteins at the 25% similarity cutoff. We use this dataset to compare CLIP against modern MoRF predictors.

2.2 Evaluation metrics

Predictions are done at the amino acid level and include a numeric propensity score (higher values denote higher likelihood that residues are in a LIP region) and a binary label that categorizes each residue as LIP vs. non-LIP. We use several popular metrics to assess binary predictions including Matthews correlation coefficient (MCC) and F1, where F1 is computed from the precision (PRE) and recall (REC) metrics:

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}},$$

$$\text{PRE} = \frac{TP}{TP+FP}, \qquad \text{REC} = \frac{TP}{TP+FN}, \qquad \text{F1} = 2 * \frac{\text{PRE} * \text{REC}}{\text{PRE} + \text{REC}}$$

where TP is the number of correctly predicted LIP residues, FP is the number of native non-LIP residues predicted as LIP residues, TN is the number of correctly predicted non-LIP residues, and FN is the number of native LIP residues predicted as non-LIP residues. We compute binary predictions from the propensities, such that residues with propensities higher than a threshold p are assumed to be located in LIPs; the remaining residues are categorized as non-LIPs. We conduct 5-fold cross validation on the TR1000 dataset to tune the threshold p, which we use to report the MCC and F1 values for CLIP. We select $p = 0.2$ that results in the maximal value of an average MCC computed over the five folds in the cross-validation experiment.

We evaluate propensity scores using the receiver operating characteristic (ROC) curves [75]. Specifically, we use all unique propensity values generated by a given predictor as a set of propensity thresholds p_{th}. For a given threshold p_i from p_{th}, a residue is classified as LIP if its putative propensity > p_i; otherwise, it is classified as the non-LIP residue. Next, false positive rates (TPR) and true positive rates (FPR) are computed using these binary predictions:

$$\text{TPR} = \frac{TP}{TP+FN}, \qquad \text{FPR} = \frac{FP}{TP+FN}$$

Finally, ROC curve is drawn by connecting (FPR, TPR) points that are calculated using all thresholds from p_{th}. We report AUC (area under ROC curve) that quantifies the area under the ROC curve and ranges between 0 and 1, where <0.5, ~0.5

and 1 correspond to a reversed, a random and a perfect prediction, respectively. Higher values of AUC, MCC, and F1 imply better predictive performance.

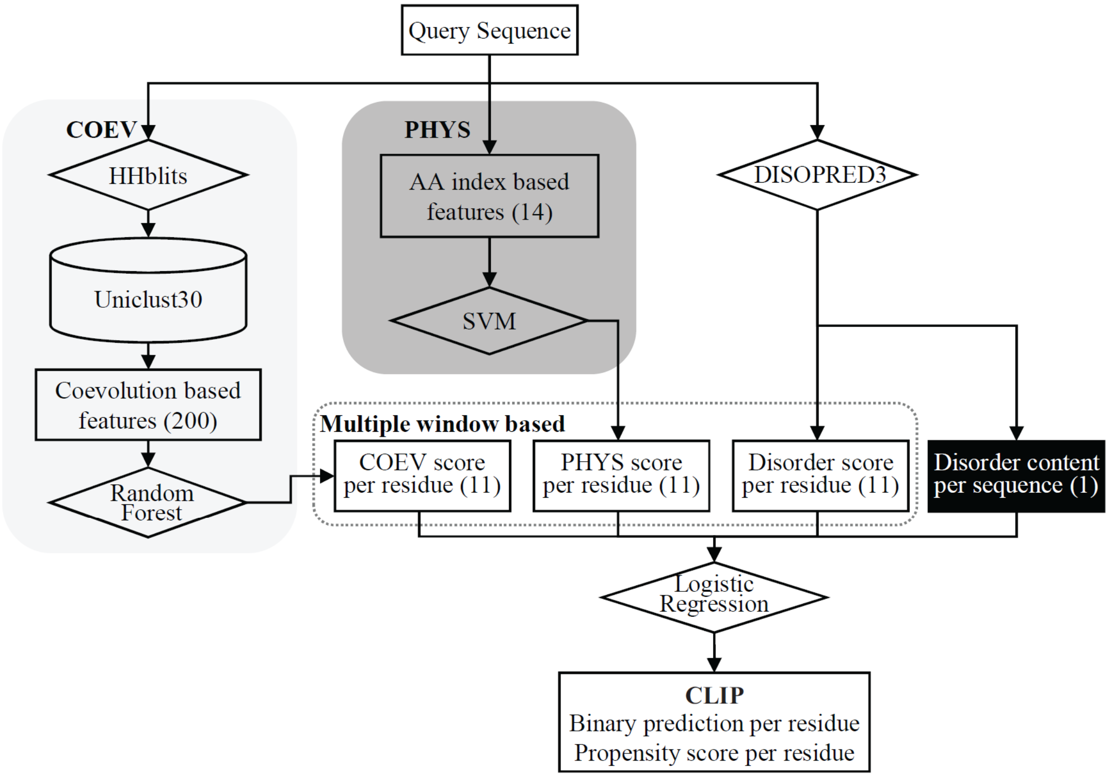
Query Sequence
COEV
HHblits
Uniclust30
Coevolution based features (200)
Random Forest
PHYS
AA index based features (14)
SVM
DISOPRED3
Multiple window based
COEV score per residue (11)
PHYS score per residue (11)
Disorder score per residue (11)
Disorder content per sequence (1)
Logistic Regression
CLIP
Binary prediction per residue
Propensity score per residue

Figure 1. Architecture of the CLIP method. The number of features used by a given predictive input, which is denoted by a rectangle, is shown inside the round brackets.

2.3 Overview of the CLIP predictor

As shown in Figure 1, CLIP has a hierarchical architecture where protein sequences are first used to derive three types of predictive inputs, which are subsequently combined using a logistic regression model. We select regression since this model is fast to train and apply to make predictions. Moreover, given its small number of parameters (i.e., coefficient of the regression function), regression is more robust to overfitting when compared to other popular algorithms, such as neural networks [76]. The latter is crucial because to the low degree of similarity between the training and test sequences. We are also motivated by the successful use of logistic regression in related problems including prediction of disordered linkers [77] and intrinsic disorder [78-81]. The three types of inputs include co-evolutionary information (COEV), physiochemical characteristics of amino acids that are relevant to binding (PHYS), and disorder prediction generated with the popular DISOPRED3 method [82]. We produce these inputs at the residue level, combine them together into a 3-dimensional vector, and process the resulting vector using sliding windows to produce inputs for the logistic regression. We consider multiple sliding window sizes of 5, 7, 9, 11, 13 and 19, use the raw values of the three inputs, and encode several additional features that aggregate these raw values at the window and whole sequence levels. The aggregate features include average COEV/PHYS/disorder scores, avg$_i$, which we use to represent the residue i in the center of the window for which we make the prediction, and difference scores, diff$_i$, that quantify difference between these averages and scores for remote neighbors (residues at both ends of the window):

$$\mathrm{avg}_i = \frac{1}{ws}\sum_{j=i-\frac{ws-1}{2}}^{i+\frac{ws-1}{2}} s_j\,, \qquad \mathrm{diff}_i = \mathrm{avg}_i - \frac{1}{ws-1}\left[\sum_{j=i-ws+1}^{i-\frac{ws+1}{2}} s_j + \sum_{j=i+\frac{ws+1}{2}}^{i+ws-1} s_j\right]$$

where ws is the window size = 5, 7, 9, 11, 13 and 19, s_j is the COEV, PHYS or disorder score for the j^{th} residue in the window. We set $s_j = 0$ for the residues in parts of the windows that are outside of the sequence (i.e., when making predictions near a sequence terminus). Moreover, we compute the sequence-level disorder content (i.e., fraction of disordered residues in a given sequence), since this feature was proven to be useful in the prediction of disordered flexible linkers [83]. We feed the resulting feature set into the logistic regression model that outputs putative propensity for LIPs for the residue in the middle of the window. We note that this predictive architecture is similar to the topology of the MoRFchibi system that also combines three types of inputs using a relatively simple predictive model [40]. However, CLIP relies on co-evolutionary information rather than sequence conservation that is used by MoRFchibi, utilizes a different disorder predictor, and aggregates inputs using a single regression model rather than multiple Bayesian predictors. Next, we describe calculation of the three predictive inputs.

2.4 COEV input

Co-evolutionary computation identifies pairs/groups of residues that change simultaneously, presumably being important to maintain the structure and/or biological functions of proteins. Co-evolutionary data is derived from a multiple sequence alignment (MSA) [84]. This information was successfully used to accurately predict key aspects of protein structure, such as residue-residue contacts [85, 86], and to investigate structural states of IDRs [71]. The latter article reveals that co-evolutionary data can be used to identify formation of distinct structural states (e.g., secondary structures) in IDRs [71]. This is relevant to our prediction since LIPs fold upon binding. Correspondingly, co-evolutionary analysis can be used to detect ability of these regions to attain distinct folded states in order to differentiate them from other IDRs that do not fold into distinct conformations.

We use the fast HHblits program [87] with coverage ≥ 50% and *e*-value < 0.001 to search for the homologues of an input sequence against the Uniclust30 database. We apply the resulting MSA to derive the symmetric covariance (scov) matrix:

$$\text{scov}(i_x, j_y) = \begin{cases} \text{cov}(i_x, j_y), & if x = y \\ \text{cov}(i_x, j_y) + \text{cov}(i_y, j_x), & if x \neq y \end{cases},$$

$$\text{cov}(i_x, j_y) = p(i_x j_y) - p(i_x)p(j_y)$$

where *i* or *j* represents the i^{th} or j^{th} column in the MSA, the i_x is a gap or a given type of 20 standard amino acids, and *p*(.) denotes frequency of a residue or a pair of residues. For each position *i*, we transform the resulting 231 scov values into 21 tcov values that quantify co-evolution:

$$\text{tcov}(i_x, j) = \sum\nolimits_{y>x} \text{scov}(i_x, j_y), x, y = 1, \cdots, 21;\ j = 1, \cdots, L;\ j \neq i$$

Using this approach, each position *i* in the input protein sequence is represented by 21*(*L*-1) tcov values, which quantify the co-evolution between the positions *i* and *j* = 1, 2, …. *L* where *L* is the input sequence length. We select the top *k* tcov values to build the COEV model by using random forest algorithm. We tune parameters for this model using the 5-fold cross validation on TR1000. The COVE model generates a COVE score for each residue in the sequence, which we use to compute inputs for the regression model.

2.5 PHYS input

Our design of the input based on the physicochemical properties of amino acids is directly inspired by a recent MoRFchibi methods that accurately predicts MoRFs [39]. The main difference is that we apply their approach to predict a more generic set of LIP regions. First, we borrow the list of the amino acid indices that quantify physicochemical properties, such as propensity for secondary structure, hydrophobicity and free energy, that were shown to be relevant to the MoRF prediction [39]. We also mimic their approach to encode features based on averaging the index values in a small window centered on the predicted residue (LIP region) in comparison with flanking regions [39]; see Supplementary Materials for details. Consequently, we compute LIP-related averages $\text{avg}_{\text{lip}}{}^{m}$ and flanking regions-based averages $\text{avg}_{\text{flank}}{}^{n}$ as follows:

$$\text{avg}_{\text{lip}}^{m} = \frac{1}{\text{len (LIP)}} \sum\nolimits_{j \in \text{LIP}} AA_j^m,\ \ \text{avg}_{\text{flank}}^{n} = \frac{1}{\text{len(flanks)}} \sum\nolimits_{j \in \text{flanks}} AA_j^n$$

where *m*, *n* = 1, 2, …, 7, len (*) denote the length of a LIP and its two flanking regions, and $\text{AA}_i{}^{m}$ and the $\text{AA}_i{}^{n}$ are the selected amino acid indices. Moreover, since our data is similarly unbalanced to the MoRF data used in ref. [39] (i.e., there are 28 times more non-LIP residues compared to the LIP residues), we reproduce their sampling procedure for our TR1000 training dataset. The resulting 14 features (7 indices, each encoded using $\text{avg}_{\text{lip}}{}^{m}$ and $\text{avg}_{\text{flank}}{}^{n}$) are input into a support vector machine (SVM) algorithm to develop the PHYS model. SVM was also used to implement MoRFchibi [39]. We tune the SVM model (*C* = 50) that relies on the Gaussian kernel (*gamma* = 0.0007) by maximizing the average AUC derived from 5-fold cross validation on the sampled TR1000 dataset. Predictions from the tuned SVM model constitute the PHYS input to the regression model.

2.6 Disorder prediction input

While the PHYS and COEV inputs focus on identifying LIP residues among IDRs, CLIP also needs to separate disordered from structured regions. In other words, we have to ensure that PHYS and COEV inputs are not used by the regression to identify structured binding residues. Correspondingly, we add the third input that identifies putative disorder in an input sequence. We investigate several recent surveys of disorder predictors to rationally select a high-quality method [88-90]. We select the popular DISOPRED3 method [82], which was shown to produce highly accurate predictions in several recent comparative studies [89, 91-93]. We use the author-provided standalone DISOPRED3 software with default parameters to generate the disorder prediction for an input protein sequence. We use the resulting residue-level putative disorder propensities to encode the avg$_i$ and diff$_i$ features and we apply the putative binary disorder predictions to compute the sequence-level disorder content.

2.7 Parametrization of the logistic regression model

We use logistic regression to combine the COEV, PHYS and disorder prediction inputs. We parametrize this model by identifying a favorable value of the ridge parameter based on the 5-fold cross validation on the TR1000 dataset. More specifically, we consider ridge = 10^n where n = -5, -4, ……4, 5, and select the ridge value = 0.1 that produces the highest average AUC that we calculate over the five AUC scores from the five test folds in the cross-validation experiment.

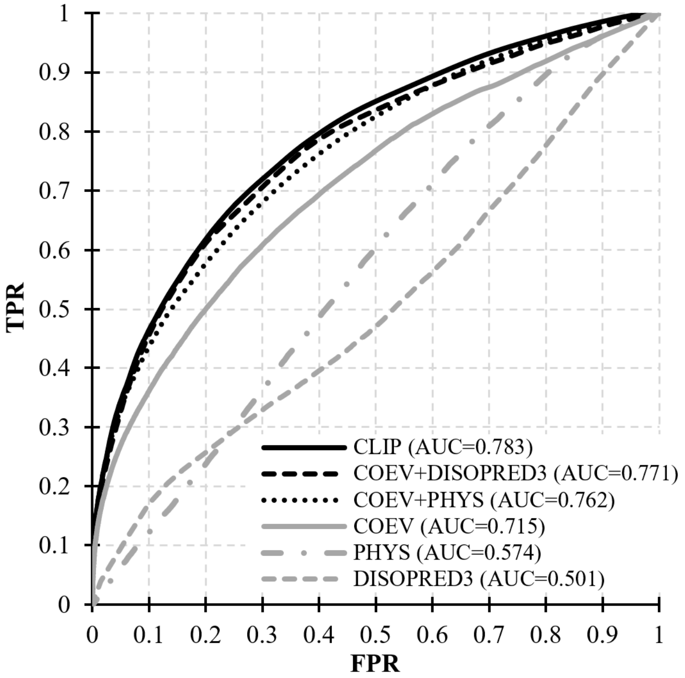

Figure 2. Results for the ablation study computed based on the 5-fold cross validation on the TR1000 dataset. AUC values are listed in the figure legend. The predictions are generated by models that rely on each of the three major input used individually, the best COEV input combined with each of the other two inputs, and the three inputs combined together; the latter option corresponds to the complete CLIP predictor. The evaluation is computed over the combined collection of the five cross-validation test folds.

3 Results

3.1 Ablation study

CLIP relies on the three types of inputs: the co-evolutionary data generated by the COEV model, the physiochemical properties-based data produced by the PHYS model, and the putative disorder output by DISOPRED3. We quantify predictive value of these inputs and investigate whether combining them together leads to improvements when compared to using them individually in the context of the LIP predictions. We quantify predictive performance using AUC values computed for the combined set of the five test folds from the 5-fold cross validation on the training set TR1000 (Figure 2).

Comparison of results produced by the COEV model, the PHYS model, and the putative disorder shows that by far the best results are produced by the co-evolutionary input, with AUC = 0.715 compared to AUC = 0.574 for the second-best PHYS model. This stems from the observations that LIPs fold into unique structures upon binding and that co-evolutionary couplings reflect formation of distinct structural states in IDRs [71]. Interestingly, the putative disorder used alone is not a strong predictor of LIPs (AUC = 0.501). This can be explained by the fact that many IDRs carry out other functions, such as entropic chains, disordered linkers and lipid, ion and small ligand binding. In other words, DISOPRED3's prediction cannot effectively separate LIP residues from the other types of disordered residues.

Next, we investigate predictive performance of designs that combine the best COEV input with one of the other two inputs. The COEV+DISOPRED3 and COEV+ PHYS models improve over the COEV model, which suggests that the disorder prediction and physicochemical-based input complement the co-evolutionary information. Moreover, the COEV+DISOPRED3 model provides slightly better predictive performance compared to the COEV+PHYS model (AUC = 0.771 vs 0.762, respectively). Interestingly, adding the disorder prediction, which by itself does not provide useful predictive input, provides a strong boost (AUC = 0.715 for COEV vs. 0.771 for COEV+DISOPRED3), suggesting that this prediction allows to effectively identify intrinsic disorder among binding residues that are predicted by the COEV model.

Finally, we compare the use of two inputs to the CLIP model that combines three inputs. We observe that adding PHYS input to the COEV+DISOPRED3 model generates a modest increase in predictive performance (AUC = 0.783 vs. 0.771). However, we still combine the three inputs to implement CLIP since the computational cost of adding PHYS model is negligible, particularly when compared to the computation of the co-evolutionary input. Altogether, we find that the co-evolutionary information is useful for the prediction of LIP residues and that combining the three inputs provides substantial improvements when compared to using these inputs individually.

3.2 Comparative assessment on the TE440 dataset

We compare CLIP with a selection of relevant published tools. The recently completed CAID experiment finds that ANCHOR2 [55, 56] and DisoRDPbind [58] are the two best predictors of binding IDRs [35]. Moreover, ANCHOR2 is used to predict protein-binding IDRs in the popular MobiDB database [61, 68, 94]. LIPs can be seen as a superset of MoRFs since they both share certain aspects, such as concomitant binding and folding when interacting with proteins and peptides. This combined with the fact that majority of the current predictors of binding IDRs target prediction of MoRFs [36, 37], motivates us to compare CLIP against a selection of modern MoRF predictors. The evaluation of the prediction of binding IDRs in the CAID experiment finds that the two top-performing MoRF predictors are MoRFchibi-light [40] and MoRFchibi-web [41]; they are ranked 3rd and 4th after ANCHOR2 and DisoRDPbind [35]. These two MoRF predictors are part of the MoRFchibi system that also includes MoRFchibi predictor [40]. Moreover, we consider one of the latest MoRF predictors, SPOT-MoRF [45]. SPOT-MoRF is limited to prediction of proteins that have up to 750 residues. Based on advice from peer reviewers, we split longer test proteins into segments that are 750 residues long and combine these predictions together to derive results for SPOT-MoRF. Altogether, we compare CLIP to six tools, which include ANCHOR2, DisoRDPbind, MoRFchibi, MoRFchibi-light, MoRFchibi-web and SPOT-MoRF, using the independent TE440 dataset (i.e., dataset that shares <25% similarity to the training proteins). We assess robustness of the predictive performance measured over different test sets by quantifying results over 20 disjoint protein sets obtained by randomly splitting the TE440 dataset into 20 subsets of 22 proteins. This experiment allows us to measure statistical significance of the differences in predictive quality between CLIP and the other two predictors over these 20 datasets. For normal measurements (which we verify with the Anderson–Darling test at the 0.05 *p*-value) we apply the *t*-test; otherwise we use the non-parametric Wilcoxon rank sum test. We assume that differences are statistically significant if the resulting *p*-value < 0.05. We apply this analysis to the AUC, MCC and F1 metrics.

Results that we summarize in Figure 3 show that CLIP secures the best predictive performance with AUC = 0.806, MCC = 0.343 and F1 = 0.325. These results are statistically higher than the corresponding AUC, MCC and F1 scores of the other methods (*p*-value < 0.05), with the second best MoRFchibi-light securing AUC = 0.763, MCC = 0.165 and F1 = 0.176. The ROC curve of CLIP (black line in Figure 3A) is consistently better than the curves of the other tools. The part of the curve where MoRFchibi-light, MoRFchibi-web and SPOT-MoRF outperform CLIP is for FPR scores > 0.7, which is where LIPs are significantly over-predicted. We observe that MoRFchibi-web and MoRFchibi-light provide similar predictive quality, which agrees with the results in CAID [35]. The lower predictive performance offered by the other six methods can be explained by the fact that they predict a specific sub-type of LIPs (MoRFs, disordered protein-binding IDRs, etc.) while CLIP provides a more holistic solution that targets interactions with nucleic acids and proteins that undergo coupled binding and folding.

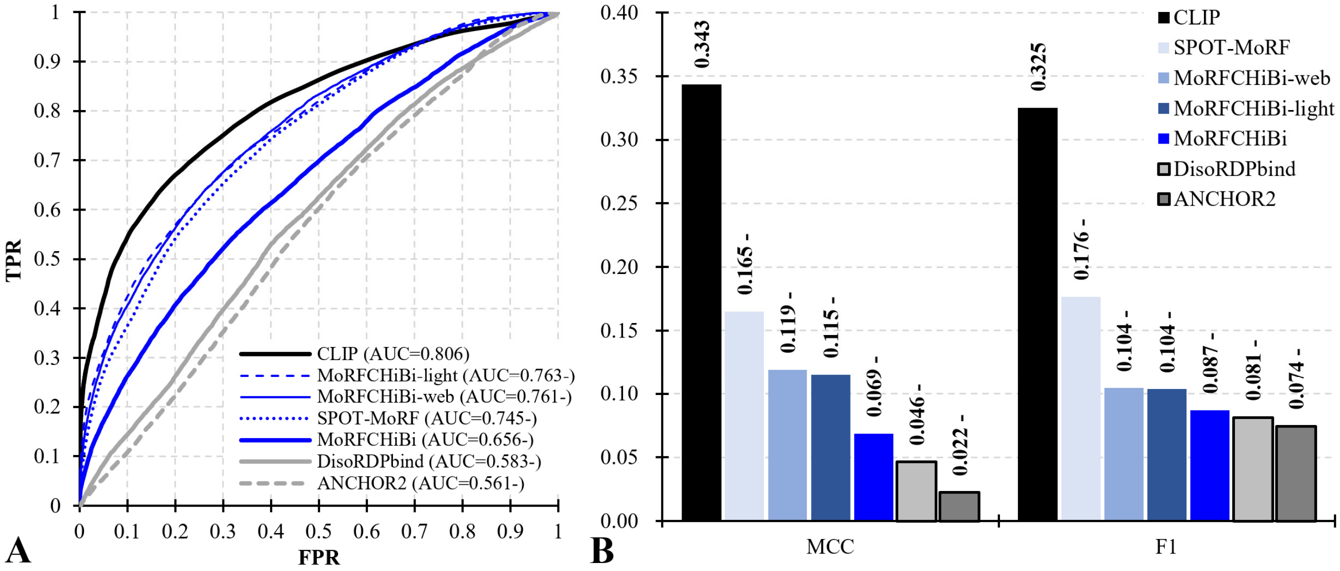

Figure 3. Predictive quality of CLIP, MoRFchibi-light, MoRFchibi-web, MoRFchibi, SPOT-MoRF, ANCHOR2 and DisoRDPbind on the TE440 dataset. The predictors of MoRF regions (MoRFchibi-light, MoRFchibi-web, MoRFchibi, and SPOT-MoRF) are denoted using blue color while the predictors of disordered protein binding regions (ANCHOR2 and DisoRDPbind) are shown using gray. Panel A shows the ROC curves and the corresponding AUC values. Panel B gives the MCC and F1 scores. "-" next to a given AUC, MCC, and F1 value indicates that the corresponding result/method is significantly worse than the result from CLIP (*p*-value < 0.05).

We also investigate whether combining CLIP results with state-of-the-art predictors of intrinsic disorder and protein structure would further improve predictive performance. The underlying idea is to utilize other tools to accurately identify disordered regions, which may help improve CLIP's ability to find the disordered LIPs. We consider the most accurate disorder predictor based on the results from the CAID assessment [35, 95], flDPnn [96], and the well-known AlphaFold2 for the structure prediction [97, 98]. We note that the pLDDT scores output by AlphaFold2 are used to identify disorder [99]. We apply the pLDDT scores from AlphaFold2 and the putative disorder propensities generated by flDPnn to refine the putative propensities generated by CLIP as follows

$$p_i^* = \frac{p_i * p_{1i}}{p_i + p_{1i}}$$

where p_i and p_{1i} are the propensities output by CLIP and AlphaFold2/flDPnn, respectively. This results in new propensity scores p_i^* that are higher than the original CLIP propensities when the scores from either AlphaFold2 or flDPnn are high, i.e., either of these two methods predicts disorder. Since AlphaFold2 takes about 3 hours to predict a single protein structure, we select 10% of proteins from the TE440 dataset at random for this analysis. We compare results from CLIP, flDPnn, AlphaFold2 and the combinations of CLIP with flDPnn, and CLIP with AlphaFold2. We evaluate statistical significance of differences by using the procedure described earlier in this section, except for relying on 20 repetitions of randomized sampling of 20 proteins, given the relatively small size of this dataset. The results, which we summarize in Suppl. Figure S2, reveal that the quality of the CLIP's predictions is consistent with the test on the entire TE440 dataset (AUC = 0.803 vs. 0.806), which suggests that the estimates from this analysis should be robust. We find that AlphaFold2 by itself cannot be used to accurately identify LIPs (AUC = 0.529). Similarly, predictions from flDPnn have statistically lower AUC than the AUC of CLIP (AUC = 0.767; *p*-value < 0.05). Our observation that flDPnn produces better results compared to AlphaFold2 when predicting disordered regions (i.e., LIPs are intrinsically disordered) aligns with a recent study that similarly shows that modern disorder predictors, such as flDPnn, outperform AlphaFold2 in the context of disorder prediction [99]. Moreover, combining CLIP with either AlphaFold2 (AUC = 0.806) or with flDPnn (AUC = 0.804), leads to only marginal gains that are not statistically significant when compared to using CLIP by itself; *p*-value = 0.36 for CLIP+AlphaFold2 and *p*-value = 0.47 for CLIP+flDPnn. This means that CLIP captures characteristics of intrinsic disorder sufficiently well to provide accurate predictions of LIPs.

Altogether, our analysis suggests that CLIP provides relatively accurate predictions of LIPs that are statistically better than the predictions of the currently available methods and that can be used without the need to apply secondary disorder predictors.

3.3 Comparative assessment on proteins with SLiMs in the TE440 dataset

LIPs cover multiple types of binding IDRs, including SLiMs. We map SLiMs from the ELM database [47, 48] into the test proteins from the TE440 dataset. We find that 95.5% of these SLiMs are located in the LIP regions and that the corresponding 3,849 SLiM residues constitute 3,849/11,994 = 32.1% of the LIP residues in the TE440 dataset. This reveals that SLiMs comprise a large fraction of LIPs. The missing 4.5% of SLiMs can be explained by the fact that we collect SLiMs at a later time compared to when the annotations from MobiDB that we use to mark LIPs in TE440 were processed. We re-evaluate CLIP and the other four methods that predict short protein-binding regions (SPOT-MoRF, MoRFCHiBi-light, MoRFCHiBi-web and MoRFCHiBi) on 268 proteins from TE440 that have only SLiMs and no other LIPs. This allows us to directly compare performance of these tools on SLiMs with the results from Figure 3 that consider a broader class of LIPs. We summarize these results in Suppl. Figure S3 and perform statistical analysis that applies the procedure from section 3.2. We find that CLIP generates the highest AUC = 0.760, however, SPOT-MoRF and MoRFchibi-web produce similarly accurate results with AUCs of 0.756 and 0.741, respectively. Moreover, CLIP's AUC for the dataset of 268 test proteins with SLiMs is statistically higher than AUC = 0.714 secured by MoRFchibi-light (p-value < 0.05) and AUC = 0.650 by MoRFchibi (p-value < 0.05). Interestingly, CLIP's ROC curve (Suppl. Figure S3) is better by a substantial margin than the ROC curves of the other methods for FPRs < 0.3, while being comparable to the MoRFchibi predictors and slightly worse to SPOT-MoRF for larger FPRs. We argue that the results for the low FPR range are more practical since SLiMs constitute a small fraction of residues and therefore high FPRs correspond to significant over-predictions. Finally, relative to the results on the TE440 dataset, CLIP, MoRFchibi-light and MoRFchibi-web register a notable drop in predictive quality, with AUC = 0.806 on TE440 vs. 0.760 on the 268 SLiMs-containing proteins for CLIP, 0.763 vs. 0.714 for MoRFchibi-light, and 0.761 vs. 0.741 for MoRFchibi-web. On the other hand, SPOT-MoRF and MoRFchibi provide similar levels of predictive quality, with AUC = 0.745 for TE440 vs. 0.756 for the SLiMs for SPOT-MoRF and AUC = 0.656 vs. 0.650 for MoRFchibi.

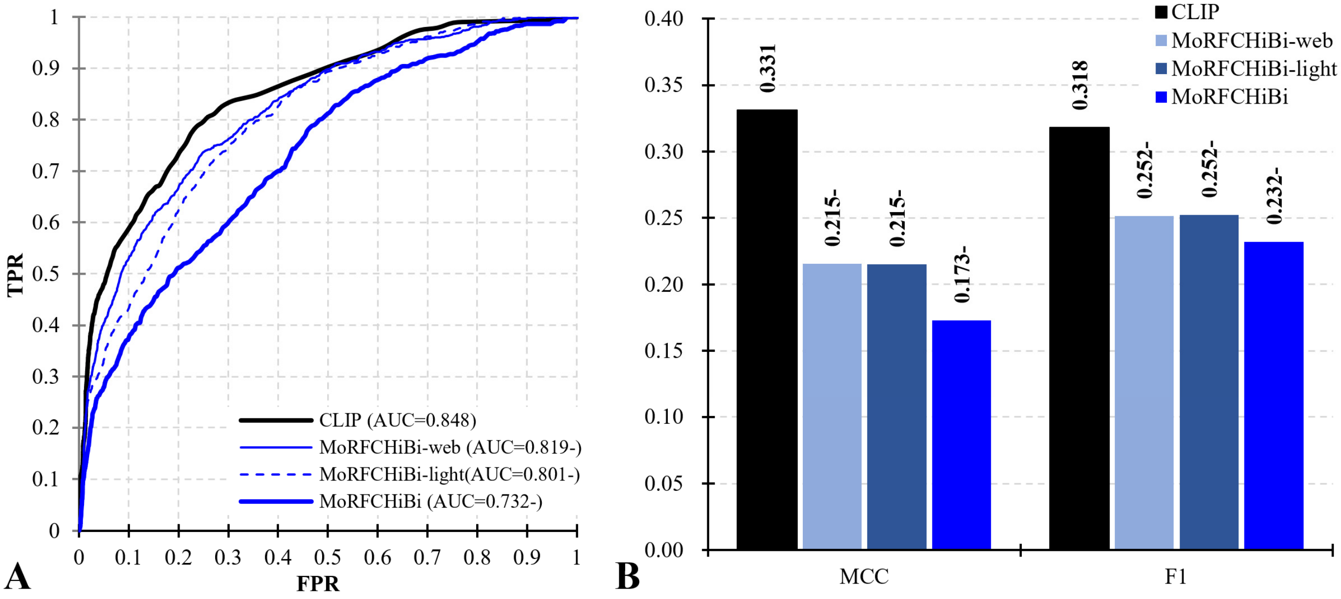

Figure 4. Predictive quality of CLIP, MoRFchibi-light, MoRFchibi-web, and MoRFchibi on the EXP25 dataset. Panel A shows the ROC curves and the corresponding AUC values. Panel B gives the MCC and F1 scores. "-" next to a given AUC, MCC, and F1 value indicates that the corresponding result/method is significantly worse than the result from CLIP (p-value < 0.05).

3.4 Comparative assessment of MoRF predictions using the EXP25 dataset

Besides testing on the datasets of LIPs and SLiMs, we also compare predictions generated by CLIP and a selection of top MoRF predictors on the EXP25 test dataset that exclusively covers MoRFs, which are a subtype of LIP regions. We compare CLIP against MoRFchibi-light and MoRFchibi-web, which are the top-ranked MoRF predictors in CAID [35]. We also include MoRFchibi that is a part of the MoRFchibi system [40]. We note that EXP25 shares low similarity with the training data of CLIP and the MoRFchibi predictors. We could not include SPOT-MoRF in this comparison since its training dataset partly overlaps with EXP25. We summarize results in Figure 4. CLIP obtains AUC = 0.848, MCC = 0.331 and F1 = 0.318, which are consistent with its performance on the TE440 dataset (Figure 3) and suggests that it can be used to accurately identify MoRFs. The best of the considered MoRF predictors, MoRFchibi-web, obtains AUC = 0.819, MCC = 0.215 and F1 = 0.252. Statistical analysis, which follows the procedure from section 3.2 (except for relying on 20 repetitions of randomized sampling of 70% of proteins, which is motivated by the relatively small size of EXP25), reveals that CLIP's results are statistically better (p-value < 0.05). However, we observe a lower magnitude of improvements when compared to the improvements for the LIP predictions on TE440. This is because EXP25 includes binding regions that are directly targeted by the MoRF predictors, resulting in a better performance of these methods on this dataset. Overall, we find that CLIP generates accurate predictions of MoRF at the levels that are consistent with the predictions for a more generic class of LIP regions.

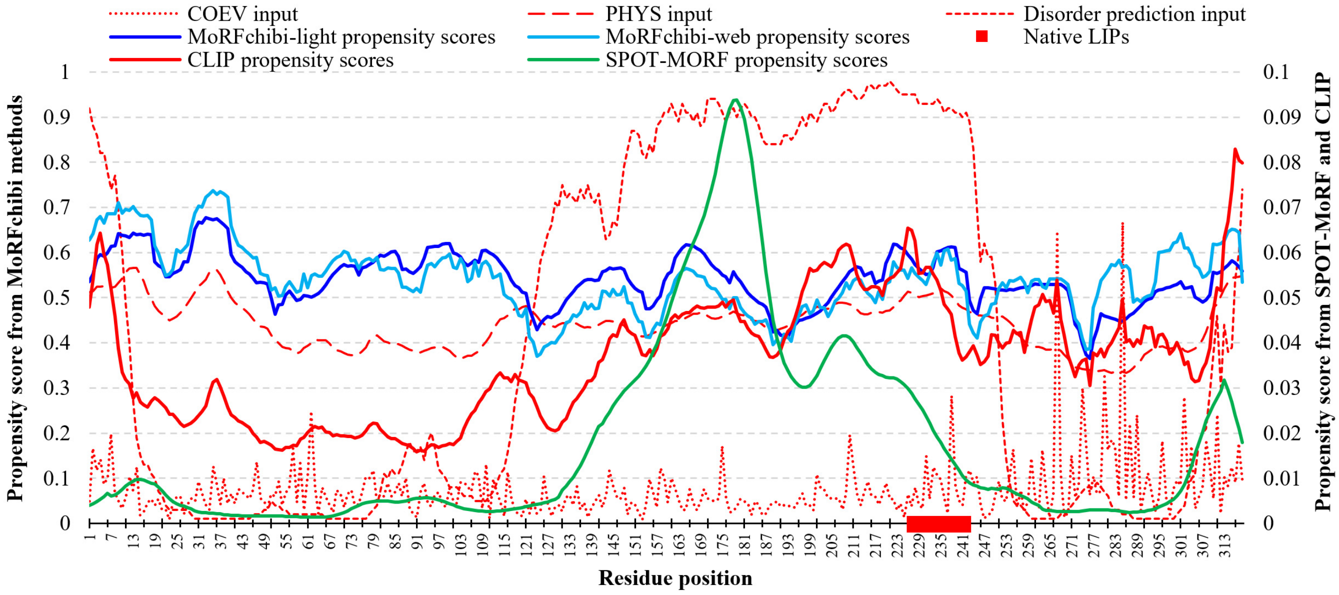

Figure 5. Visualization of the predicted propensity scores and the three main inputs to the CLIP model (i.e., COEV, PHYS, and predicted disorder that are shown using dashed and dotted red lines) for an example protein from the TE440 dataset, the STAM-binding protein from human (MobiDB ID: C9JK83). The putative propensities were produced using CLIP (red solid line) and the top three other methods that we tested on the TE440 dataset, MoRFchibi-light (dark blue solid line), MoRFchibi-web (light blue solid line), and SPOT-MoRF (green solid line). The location of the native LIP region (positions 228 to 241) is shown with the horizontal red bar at the bottom of the plot.

3.5 Case study

We illustrate and explain predictions generated by CLIP and the top three other methods selected based on results on the TE440 dataset (SPOT-MoRF, MoRFchibi-web and MoRFchibi-light) for an example protein from this test dataset, the human STAM-binding protein (MobiDB ID: C9JK83). The STAM-binding protein is 318-residues long and includes LIP in positions 228 to 241. The two versions of MoRFchibi produce relatively similar propensities that register a local spike in values around the native LIP region; see blue lines in Figure 5. However, these tools also predict other regions with similarly high propensities, suggesting that they might be over-predicting MoRF residues for this protein. SPOT-MoRF produces relatively higher propensities between positions 140 and 240, and at the C-terminus; see green line in Figure 5. This prediction partly overlaps with the location of the native LIP, although SPOT-MoRF's propensities suggest that the most likely spot for a putative MoRF region is between positions 166 and 186. CLIP predicts high propensities for positions 197 to 235, and for a couple of short regions at both termini; see solid red line in Figure 5. The prediction in the middle of the sequence substantially overlaps with the native LIP, and arguably could be used to identify this region. We explain CLIP's predictions by showing how the CLIP-generated propensities relate to the inputs for our model. We show the three inputs in Figure 5 using the dotted red line (COEV input), dashed red line (disorder prediction), and long dashed red line (PHYS input). We observe that high propensities produced by CLIP coincide with high values of the three inputs. The native LIP region has relatively high propensities for the evolutionary couplings and intrinsic disorder and favorable values of the physicochemical properties at the amino acid level. Moreover, while other regions in this protein are also predicted as disordered, i.e., dashed red line suggests high likelihood of disorder between positions 124 and 250, positions 124 to 196 have relatively lower COEV and PHYS values, and the 236 to 150 segment has lower COEV values. This suggests that it is crucial to combine these three factors to make correct predictions.

4 Conclusions

LIPs are a recently introduced category of disordered binding regions that generalize previously defined classes, such as MoRFs, SLiMs, and disordered binding domains [60, 61]. Motivated by the lack of dedicated predictors of LIPs, availability of a large quantity of training data, and the documented need to improve predictions of binding IDRs [35], we design, implement and comparatively evaluate CLIP, a new predictor of LIPs from protein sequences. CLIP combines three types of inputs to predict LIPs: co-evolutionary information, putative disorder and literature-inspired approach to quantify relevant physiochemical properties of amino acids [39]. Ablation analysis shows that co-evolutionary information is a strong marker of LIPs. This can be explained by the fact that evolutionary couplings detect structural states of IDRs, which are among the key hallmarks of LIPs that undergo disorder-to-order transitions upon binding [71]. We also empirically demonstrate that combining the three inputs provides a large increase in the predictive quality when compared to using these inputs individually.

Comparative tests based on two independent datasets, TE440 and EXP25, show that CLIP secures favorable predictive performance when compared to a representative selection of closest current methods that predict MoRFs and disordered protein-binding regions. The defining differences between CLIP and the other methods is the use of the co-evolutionary information and the fact that CLIP was designed using a training dataset annotated with a broader class of LIP regions. Altogether, our analysis suggests that CLIP provides accurate predictions of LIPs. We provide a convenient webserver that implements CLIP at http://biomine.cs.vcu.edu/servers/CLIP/. The webserver takes up to 3 amino acid sequences in the FASTA format as the input, with an option to provide email address where a notification of completed prediction is delivered. It takes about 10 minutes to complete prediction for an average length sequence with about 300 residues. The prediction is automated and performed entirely on the server side. Users do not need to install any software. The webserver outputs propensities and binary predictions for each residue in the input sequence and these values are provided in an easily parsable text file. Users can also download standalone code for CLIP at http://yanglab.qd.sdu.edu.cn/download/CLIP/.

The most likely reason why the considered here predictors of binding IDRs provide lower quality results when compared to CLIP is that they were designed for more specific types of interacting IDRs, i.e., MoRFs and protein-binding domains vs. LIPs. We also note that results for the other tools that we present differ from observations in the CAID study. For instance, CAID shows that ANCHOR2 and DisoRDPbind improve over other methods, including the family of MoRFchibi tools, in the prediction of binding IDRs [35]. Our results (Figure 3) show that MoRFchibi tools and SPOT-MoRF provide more accurate results than ANCHOR2 and DisoRDPbind. The most likely reason for this difference is the fact that CAID evaluates predictors on IDRs that bind a variety of ligand types (Suppl. Figure 51 in [35] lists these ligands), while we focus specifically on LIPs. Ultimately, one of the key aspects that affects predictive performance is the fit between the type of interaction that a given predictor addresses and the type of interactions that are considered in a given assessments.

One interesting option to expand this work is to consider context of an interacting partner molecule, i.e., predict whether a given protein sequences (i.e., putative binding IDR in that sequence) interacts with a given sequence of the partner molecule (protein or nucleic acid) at the residue level (i.e., identify binding residues in both sequences). To the best of our knowledge, the residue level predictions of disordered partner-specific interactions were not yet pursued while similar attempts for the partner-specific prediction of protein-RNA interactions have not been successful so far. In fact, performance of RNA partner-specific predictors was shown to be equivalent to the performance of partner-agnostic methods [100]. Moreover, the currently available amount of annotated data for interacting IDRs is insufficient to design and assess partner-specific predictors. A significant majority of binding annotations for IDRs in DisProt and MobiDB do not have information about the sequence of the binding regions for the interacting partner(s). We plan to address this challenging topic in the future work, once the amount of the partner-annotated interaction data becomes sufficient large, by taking advantage of accurate predictions of LIPs that are produced by CLIP.

Funding

Z.P. was supported in part by National Natural Science Foundation of China [grants 61873185 and T2222012]. L.K. and B.Z. were supported in part by Robert J. Mattauch Endowment and National Science Foundation [grant 2146027].

References

1. van der Lee R, Buljan M, Lang B et al. Classification of intrinsically disordered regions and proteins, Chem Rev 2014;114:6589-6631.
2. Oldfield CJ, Uversky VN, Dunker AK et al. Chapter 1 - Introduction to intrinsically disordered proteins and regions. In: Salvi N. (ed) Intrinsically Disordered Proteins. Academic Press, 2019, 1-34.
3. Habchi J, Tompa P, Longhi S et al. Introducing protein intrinsic disorder, Chem Rev 2014;114:6561-6588.
4. Xie H, Vucetic S, Iakoucheva LM et al. Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions, Journal of Proteome Research 2007;6:1882-1898.
5. Peng Z, Yan J, Fan X et al. Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life, Cellular & Molecular Life Sciences 2015;72:137-151.
6. Lieutaud P, Ferron F, Uversky AV et al. How disordered is my protein and what is its disorder for? A guide through the "dark side" of the protein universe, Intrinsically Disord Proteins 2016;4:e1259708.

7. Meng F, Na I, Kurgan L et al. Compartmentalization and Functionality of Nuclear Disorder: Intrinsic Disorder and Protein-Protein Interactions in Intra-Nuclear Compartments, Int J Mol Sci 2015;17.
8. Babu MM. The contribution of intrinsically disordered regions to protein function, cellular complexity, and human disease, Biochem Soc Trans 2016;44:1185-1200.
9. Kjaergaard M, Kragelund BB. Functions of intrinsic disorder in transmembrane proteins, Cellular and Molecular Life Sciences 2017;74:3205-3224.
10. Dunker AK, Silman I, Uversky VN et al. Function and structure of inherently disordered proteins, Curr Opin Struct Biol 2008;18:756-764.
11. Chen J, Kriwacki RW. Intrinsically Disordered Proteins: Structure, Function and Therapeutics, J Mol Biol 2018;430:2275-2277.
12. Zhao B, Katuwawala A, Uversky VN et al. IDPology of the living cell: intrinsic disorder in the subcellular compartments of the human cell, Cell Mol Life Sci 2020.
13. Zhao B, Katuwawala A, Oldfield CJ et al. Intrinsic Disorder in Human RNA-Binding Proteins, J Mol Biol 2021;433:167229.
14. Dyson HJ, Wright PE. Coupling of folding and binding for unstructured proteins, Curr Opin Struct Biol 2002;12:54-60.
15. Dunker AK, Cortese MS, Romero P et al. Flexible nets: The roles of intrinsic disorder in protein interaction networks, FEBS Journal 2005;272:5129-5148.
16. Receveur-Brechot V, Bourhis JM, Uversky VN et al. Assessing protein disorder and induced folding, Proteins 2006;62:24-45.
17. Uversky VN. The multifaceted roles of intrinsic disorder in protein complexes, FEBS Lett 2015;589:2498-2506.
18. Hsu WL, Oldfield C, Meng J et al. Intrinsic protein disorder and protein-protein interactions, Pac Symp Biocomput 2012:116-127.
19. Fuxreiter M, Toth-Petroczy A, Kraut DA et al. Disordered proteinaceous machines, Chem Rev 2014;114:6806-6843.
20. Neduva V, Linding R, Su-Angrand I et al. Systematic discovery of new recognition peptides mediating protein interaction networks, PLoS Biol 2005;3:e405.
21. Oldfield CJ, Meng J, Yang JY et al. Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners, BMC Genomics 2008;9 Suppl 1:S1.
22. Hsu WL, Oldfield CJ, Xue B et al. Exploring the binding diversity of intrinsically disordered proteins involved in one-to-many binding, Protein Sci 2013;22:258-273.
23. Mohan A, Oldfield CJ, Radivojac P et al. Analysis of molecular recognition features (MoRFs), J Mol Biol 2006;362:1043-1059.
24. Vacic V, Oldfield CJ, Mohan A et al. Characterization of molecular recognition features, MoRFs, and their binding partners, J Proteome Res 2007;6:2351-2366.
25. Yan J, Dunker AK, Uversky VN et al. Molecular recognition features (MoRFs) in three domains of life, Mol Biosyst 2016;12:697-710.
26. Neduva V, Russell RB. Linear motifs: evolutionary interaction switches, FEBS Lett 2005;579:3342-3345.
27. Davey NE, Van Roey K, Weatheritt RJ et al. Attributes of short linear motifs, Mol Biosyst 2012;8:268-281.
28. Bhowmick P, Guharoy M, Tompa P. Bioinformatics Approaches for Predicting Disordered Protein Motifs, Adv Exp Med Biol 2015;870:291-318.
29. Dinkel H, Van Roey K, Michael S et al. ELM 2016--data update and new functionality of the eukaryotic linear motif resource, Nucleic Acids Res 2016;44:D294-300.
30. Shaji D, Amemiya T, Koike R et al. Interface property responsible for effective interactions of protean segments: Intrinsically disordered regions that undergo disorder-to-order transitions upon binding, Biochem Biophys Res Commun 2016;478:123-127.
31. Fukuchi S, Amemiya T, Sakamoto S et al. IDEAL in 2014 illustrates interaction networks composed of intrinsically disordered proteins and their binding partners, Nucleic Acids Res 2014;42:D320-325.
32. Tompa P, Fuxreiter M, Oldfield CJ et al. Close encounters of the third kind: disordered domains and the interactions of proteins, Bioessays 2009;31:328-335.
33. Dunker AK, Babu MM, Barbar E et al. What's in a name? Why these proteins are intrinsically disordered, Intrinsically Disordered Proteins 2013;1:e24157.
34. Zhao B, Kurgan L. Compositional Bias of Intrinsically Disordered Proteins and Regions and Their Predictions, Biomolecules 2022;12.
35. Necci M, Piovesan D, Predictors C et al. Critical assessment of protein intrinsic disorder prediction, Nature Methods 2021;18:472-481.
36. Katuwawala A, Ghadermarzi S, Kurgan L. Computational prediction of functions of intrinsically disordered regions, Prog Mol Biol Transl Sci 2019;166:341-369.
37. Katuwawala A, Peng Z, Yang J et al. Computational Prediction of MoRFs, Short Disorder-to-order Transitioning Protein Binding Regions, Comput Struct Biotechnol J 2019;17:454-462.
38. Disfani FM, 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, Bioinformatics 2012;28:i75-83.

39. Malhis N, Gsponer J. Computational identification of MoRFs in protein sequences, Bioinformatics 2015;31:1738-1744.
40. Malhis N, Jacobson M, Gsponer J. MoRFchibi SYSTEM: software tools for the identification of MoRFs in protein sequences, Nucleic Acids Res 2016;44:W488-493.
41. Malhis N, Wong ETC, Nassar R et al. Computational Identification of MoRFs in Protein Sequences Using Hierarchical Application of Bayes Rule, PLoS One 2015;10.
42. Sharma R, Raicar G, Tsunoda T et al. OPAL: prediction of MoRF regions in intrinsically disordered protein sequences, Bioinformatics 2018;34:1850-1858.
43. Sharma R, Sharma A, Raicar G et al. OPAL+: Length-Specific MoRF Prediction in Intrinsically Disordered Protein Sequences, Proteomics 2019;19:e1800058.
44. He H, Zhao J, Sun G. Prediction of MoRFs in Protein Sequences with MLPs Based on Sequence Properties and Evolution Information, Entropy (Basel) 2019;21.
45. Hanson J, Litfin T, Paliwal K et al. Identifying molecular recognition features in intrinsically disordered regions of proteins by transfer learning, Bioinformatics 2020;36:1107-1113.
46. He H, Zhou Y, Chi Y et al. Prediction of MoRFs based on sequence properties and convolutional neural networks, BioData Min 2021;14:39.
47. Dinkel H, Michael S, Weatheritt RJ et al. ELM--the database of eukaryotic linear motifs, Nucleic Acids Res 2012;40:D242-251.
48. Kumar M, Gouw M, Michael S et al. ELM-the eukaryotic linear motif resource in 2020, Nucleic Acids Res 2020;48:D296-D306.
49. Lyon KF, Cai X, Young RJ et al. Minimotif Miner 4: a million peptide minimotifs and counting, Nucleic Acids Res 2018;46:D465-D470.
50. Mi T, Merlin JC, Deverasetty S et al. Minimotif Miner 3.0: database expansion and significantly improved reduction of false-positive predictions from consensus sequences, Nucleic Acids Res 2012;40:D252-260.
51. Balla S, Thapar V, Verma S et al. Minimotif Miner: a tool for investigating protein function, Nat Methods 2006;3:175-177.
52. Krystkowiak I, Davey NE. SLiMSearch: a framework for proteome-wide discovery and annotation of functional modules in intrinsically disordered regions, Nucleic Acids Res 2017;45:W464-W469.
53. Davey NE, Haslam NJ, Shields DC et al. SLiMSearch 2.0: biological context for short linear motifs in proteins, Nucleic Acids Res 2011;39:W56-60.
54. Gutman R, Berezin C, Wollman R et al. QuasiMotiFinder: protein annotation by searching for evolutionarily conserved motif-like patterns, Nucleic Acids Res 2005;33:W255-261.
55. Dosztanyi Z, Meszaros B, Simon I. ANCHOR: web server for predicting protein binding regions in disordered proteins, Bioinformatics 2009;25:2745-2746.
56. Meszaros B, Erdos G, Dosztanyi Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding, Nucleic Acids Res 2018;46:W329-W337.
57. 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.
58. Peng Z, Kurgan L. High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder, Nucleic Acids Res 2015;43:e121.
59. 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;23.
60. Piovesan D, Tosatto SCE. Mobi 2.0: an improved method to define intrinsic disorder, mobility and linear binding regions in protein structures, Bioinformatics 2018;34:122-123.
61. Piovesan D, Tabaro F, Paladin L et al. MobiDB 3.0: more annotations for intrinsic disorder, conformational diversity and interactions in proteins, Nucleic Acids Res 2018;46:D471-D476.
62. Monzon AM, Bonato P, Necci M et al. FLIPPER: Predicting and Characterizing Linear Interacting Peptides in the Protein Data Bank, J Mol Biol 2021;433:166900.
63. wwPDB consortium. Protein Data Bank: the single global archive for 3D macromolecular structure data, Nucleic Acids Res 2019;47:D520-D528.
64. Quaglia F, Meszaros B, Salladini E et al. DisProt in 2022: improved quality and accessibility of protein intrinsic disorder annotation, Nucleic Acids Res 2022;50:D480-D487.
65. Sickmeier M, Hamilton JA, LeGall T et al. DisProt: the Database of Disordered Proteins, Nucleic Acids Res 2007;35:D786-793.
66. Schad E, Ficho E, Pancsa R et al. DIBS: a repository of disordered binding sites mediating interactions with ordered proteins, Bioinformatics 2018;34:535-537.
67. Ficho E, Remenyi I, Simon I et al. MFIB: a repository of protein complexes with mutual folding induced by binding, Bioinformatics 2017;33:3682-3684.
68. Piovesan D, Necci M, Escobedo N et al. MobiDB: intrinsically disordered proteins in 2021, Nucleic Acids Res 2021;49:D361-D367.

69. 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.
70. Dobson L, Tusnady GE. MemDis: Predicting Disordered Regions in Transmembrane Proteins, Int J Mol Sci 2021;22.
71. Toth-Petroczy A, Palmedo P, Ingraham J et al. Structured States of Disordered Proteins from Genomic Sequences, Cell 2016;167:158-+.
72. Meng Q, Peng Z, Yang J. CoABind: a novel algorithm for Coenzyme A (CoA)- and CoA derivatives-binding residues prediction, Bioinformatics 2018;34:2598-2604.
73. Huang Y, Niu B, Gao Y et al. CD-HIT Suite: a web server for clustering and comparing biological sequences, Bioinformatics 2010;26:680-682.
74. Malhis N, Wong ET, Nassar R et al. Computational Identification of MoRFs in Protein Sequences Using Hierarchical Application of Bayes Rule, PLoS One 2015;10:e0141603.
75. Fawcett T. An introduction to ROC analysis, Pattern Recognition Letters 2006;27:861-874.
76. Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review, Journal of Biomedical Informatics 2002;35:352-359.
77. Meng F, Kurgan L. DFLpred: High-throughput prediction of disordered flexible linker regions in protein sequences, Bioinformatics 2016;32:i341-i350.
78. Fan X, Kurgan L. Accurate prediction of disorder in protein chains with a comprehensive and empirically designed consensus, J Biomol Struct Dyn 2014;32:448-464.
79. Peng Z, Mizianty MJ, Kurgan L. Genome-scale prediction of proteins with long intrinsically disordered regions, Proteins 2014;82:145-158.
80. Yan J, Mizianty MJ, Filipow PL et al. RAPID: fast and accurate sequence-based prediction of intrinsic disorder content on proteomic scale, Biochim Biophys Acta 2013;1834:1671-1680.
81. Radivojac P, Obradovic Z, Brown CJ et al. Prediction of boundaries between intrinsically ordered and disordered protein regions, Pac Symp Biocomput 2003:216-227.
82. Jones DT, Cozzetto D. DISOPRED3: precise disordered region predictions with annotated protein-binding activity, Bioinformatics 2015;31:857-863.
83. Peng Z, Xing Q, Kurgan L. APOD: accurate sequence-based predictor of disordered flexible linkers, Bioinformatics 2020;36:i754-i761.
84. Hu G, Kurgan L. Sequence Similarity Searching, Curr Protoc Protein Sci 2019;95:e71.
85. Wu Q, Peng Z, Anishchenko I et al. Protein contact prediction using metagenome sequence data and residual neural networks, Bioinformatics 2020;36:41-48.
86. Wuyun Q, Zheng W, Peng Z et al. A large-scale comparative assessment of methods for residue-residue contact prediction, Brief Bioinform 2018;19:219-230.
87. Remmert M, Biegert A, Hauser A et al. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment, Nat Methods 2011;9:173-175.
88. Meng F, Uversky VN, Kurgan L. Comprehensive review of methods for prediction of intrinsic disorder and its molecular functions, Cell Mol Life Sci 2017;74:3069-3090.
89. Liu Y, Wang X, Liu B. A comprehensive review and comparison of existing computational methods for intrinsically disordered protein and region prediction, Brief Bioinform 2019;20:330-346.
90. Kurgan L, Li M, Li Y. The Methods and Tools for Intrinsic Disorder Prediction and their Application to Systems Medicine. In: Wolkenhauer O. (ed) Systems Medicine. Oxford: Academic Press, 2021, 159-169.
91. Katuwawala A, Oldfield CJ, Kurgan L. Accuracy of protein-level disorder predictions, Brief Bioinform 2019.
92. Monastyrskyy B, Kryshtafovych A, Moult J et al. Assessment of protein disorder region predictions in CASP10, Proteins 2014;82 Suppl 2:127-137.
93. Katuwawala A, Oldfield CJ, Kurgan L. DISOselect: Disorder predictor selection at the protein level, Protein Sci 2020;29:184-200.
94. Di Domenico T, Walsh I, Martin AJM et al. MobiDB: a comprehensive database of intrinsic protein disorder annotations, Bioinformatics 2012;28:2080-2081.
95. Lang B, Babu MM. A community effort to bring structure to disorder, Nature Methods 2021;18:454-455.
96. Hu G, Katuwawala A, Wang K et al. flDPnn: Accurate intrinsic disorder prediction with putative propensities of disorder functions, Nat Commun 2021;12:4438.
97. Jumper J, Evans R, Pritzel A et al. Highly accurate protein structure prediction with AlphaFold, Nature 2021;596:583-589.
98. Kryshtafovych A, Schwede T, Topf M et al. Critical assessment of methods of protein structure prediction (CASP)-Round XIV, Proteins 2021;89:1607-1617.
99. Wilson CJ, Choy WY, Karttunen M. AlphaFold2: A Role for Disordered Protein/Region Prediction?, Int J Mol Sci 2022;23.
100. Jung Y, El-Manzalawy Y, Dobbs D et al. Partner-specific prediction of RNA-binding residues in proteins: A critical assessment, Proteins 2019;87:198-211.