Reliable prediction of short linear motifs in the human proteome

Rita Pancsa[1,†], Erzsébet Fichó[2,†], Zsófia E. Kálmán[3], Csongor Gerdán[1], István Reményi[4], András Zeke[1,5], Gábor E. Tusnády[1,6], Laszlo Dobson[1,6,*]

[1]Institute of Molecular Life Sciences, Research Centre for Natural Sciences, Magyar Tudósok Körútja 2, Budapest 1117, Hungary
[2]Department of Bioinformatics, Cytocast Hungary Kft, Petőfi Sándor utca 5/A, Budapest 1052, Hungary
[3]Faculty of Information Technology and Bionics, Pázmány Péter Catholic University, Práter utca 50/A, Budapest 1083, Hungary
[4]Department of Artificial Intelligence, Faculty of Informatics, ELTE Eötvös Loránd University, Pázmány Péter sétány 1/C, Budapest 1117, Hungary
[5]Department of Biology, University of Padua, Via Ugo Bassi, 58/B, 35131 Padua, Italy
[6]Department of Bioinformatics, Semmelweis University, Tűzoltó u. 7, Budapest 1094, Hungary
[†]These authors have contributed equally to the work.
[*]To whom correspondence should be addressed: Laszlo Dobson: dobson.laszlo@ttk.hu

Abstract

Short linear motifs (SLiMs) are small, often transient interaction modules within intrinsically disordered regions (IDRs) of proteins that interact with particular domains and thereby regulate numerous biological processes. The limited sequence information within these short peptides leads to frequent false positive hits in both computational and experimental SLiM identification methods. This makes the description of novel SLiMs challenging and has limited the number of known cases to a few thousand, even though SLiMs play widespread roles in cellular functions. We present SLiMMine, a deep learning-based method to identify SLiMs in the human proteome. By refining the annotations of known, annotated motif classes, we created a high-quality dataset for model training. Using protein embeddings and neural networks, SLiMMine reliably predicts novel SLiM candidates in known classes, eliminates ~80% of the pattern matching-based motif hits as false-positives, furthermore, it can also be used as a discovery tool to find uncharacterized SLiMs based on optimal sequence environment. In addition, we narrowed the highly general interactor-domain definitions of known SLiM classes to specific human proteins, enabling more precise prediction of a wide range of potential protein-protein interactions (PPIs) in the human interactome. SLiMMine is available in the form of an appealing, user-friendly, multi-purpose web-server at https://slimmine.pbrg.hu/.

Introduction

Short linear motifs (SLiMs) are protein segments typically composed of 3–10 amino acids that can interact with specific protein domain types[1]. Their structural flexibility and limited number of specificity-determining residues allow them to *de novo* emerge within the intrinsically disordered regions (IDRs) of proteins through evolution[2]. SLiMs usually mediate weak, dynamic and transient interactions, often through coupled folding and binding, which enables them to function in cooperative regulatory systems. SLiMs are the characteristic interaction modules of IDRs, and those contribute to approximately half of the solvent-accessible surface of the human proteome[2,3], therefore, millions of SLiMs are expected to regulate the diverse biological processes of human cells[4]. Their importance is reflected in their involvement in various diseases, ranging from disease-associated mutations that abolish the function of the motif to pathogenic effector proteins that mimic these segments to hijack host-cell regulation.

Despite the growing knowledge on SLiMs, their identification remains challenging. Both computational and experimental methods struggle due to the limited sequence information within these short segments, which complicates robust screenings and leads to high false-positive rates[5]. Unlike other structural or functional protein units, SLiM identification often requires a combination of *in silico*, *in vitro*, and *in vivo* validation. Because multiple steps are needed to confirm a motif's functionality, the number of classified examples remains limited to only a few thousand[6].

In accord, current resources for SLiMs are also limited. The Eukaryotic Linear Motif (ELM) resource is richly annotated and its curation process incorporates detailed structural, functional and evolutionary analysis[6]. Other sources of SLiMs (such as LMPID[7] or UniProt annotations[8]) can be relevant starting points, however, they do not cover all of these aspects.

Numerous attempts have been made to enrich the limited information stored in these databases, including high-throughput experimental and computational approaches. The proteomic peptide-phage display (ProP-PD) protocol identifies short linear motifs on the proteome-scale. Initially developed for PDZ domains[9], it has now been extended to other SLiM-binding bait domains, enabling high-throughput discovery of domain–SLiM interactions across the human proteome[10,11]. Complementing these experimental strategies, quantitative fragmentomics studies such as those by Gogl et al.[12] have systematically measured tens of thousands of PDZ-binding motif (PBM)-PDZ domain affinities within the human interactome, providing a detailed landscape of binding specificity and promiscuity that enriches our understanding of PDZ recognition and specificity beyond traditional motif definitions. In parallel, computational approaches can identify protein binding segments in disordered regions based on biophysical properties[13] or protein embeddings[14], find conserved segments[15], apply contextual filters[16] or prioritize those motifs of a protein, whose partner domains are present in the experimentally verified interaction partners of the protein[17].

Here, we propose an accurate deep learning-based method for identifying SLiMs in the human proteome that largely solves the problem of abundant false positive hits. We heavily relied on the manually curated Eukaryotic Linear Motif (ELM) resource and manually refined and extended its annotations to construct a high-quality training dataset that represents a high added value in itself. Using protein embeddings and various neural network architectures, we developed a predictive model that reliably identifies novel SLiM candidates. Based on the new, more detailed annotations and available PPI data, we

pinpoint the likely interaction partners of the predicted motif instances, and thereby also propose a likely interaction mechanism for a large number of human PPIs, as well as predict potential novel PPIs.

Results

Refinement of ELM class annotations to gain an optimal training set for developing a SLiM predictor

For the training of a SLiM predictor and identification of the most likely binding partners of the predicted SLiM instances in PPI data, we needed a large amount of reliable SLiM data annotated along with specific criteria.

Since we needed true motifs precisely mapped onto human proteins, and some annotations in the Eukaryotic Linear Motif (ELM) Resource are quite old, we first mapped motif instances described in orthologous mammalian proteins to the current version of the human proteome, and then also adjusted the motif boundaries of human instances (if the UniProt[8] sequence was updated since the release of the ELM[6] annotation).

To obtain a training dataset that fulfills our pre-defined criteria, we have refined and extended the annotations of the ~320 motif classes of the ELM database that are relevant for human proteins. Relying on the text descriptions of the ELM database, the biological context of the motif classes and the related literature, we determined for each class if the motif-containing protein counterparts 1) are extracellular or intracellular, 2) if intracellular, what compartment they are expected to localize to, and 3) what are their expected Gene Ontology[18] (GO) biological processes and molecular functions. Although the ELM database has its own GO annotations, those would have been problematic to use, since they contained GO terms that were mixed, meaning either relevant to the motif- or domain-containing counterparts of the interactions (or both). In addition, whenever possible, we precisely defined the possible binding partners for the given SLiM class, i.e. the subset of the human proteins carrying the specific binding domain that was described to actually bind to the given SLiM type. This was required because the domain definitions provided in ELM are most often not specific enough. For instance, the docking motif specific for binding to the kinase domains of the three JNK kinases (ELM class: DOC_MAPK_JIP1_4) had the kinase domain as the assigned domain type, which can be found in hundreds of human kinases. In our reannotation procedure, only the UniProt ACs of the three human JNK kinases were listed as possible partners for this motif class, thereby precisely defining the real sequence space of the given interaction type. In other cases, broad definitions of the interacting domains, usually provided as PFAM domain identifiers, could be changed to more specific definitions from InterPro[19]. Precise definition of the possible partners for each ELM class enabled us to determine the most likely binding partners of the predicted motif instances in later steps: for a predicted motif instance in any given protein, the overlap between the possible partners of the respective ELM class and the known binding partners of the protein (in the human interactome) provides the most likely binding partners of the given predicted motif instance. From now on, we refer to this modified version of ELM data as ELM_refined (Figure 1/A).

Assembling a reliable benchmark set for the training of SLiMMine

Classified ELM instances with the adjusted motif boundaries were used as positive data. Constructing a negative dataset is more challenging, as we cannot genuinely rule out regions without negative evidence - except motifs labeled as false positives in ELM. As the latter dataset is very small, we defined a negative dataset for each motif class. To this end, we defined a set of criteria that could indicate that a given protein or protein region is not favourable/likely to carry a true functional motif belonging to a given motif class. We scanned the human proteome for regular expressions defined in ELM and flagged all motif hits using the following criteria (these definitions may overlap). I) The protein localization is not appropriate for the motif class. II) The motif hit is not conserved in mammals. III) The motif hit is buried in a monomeric PDB[20] structure. IV) The motif hit was not predicted to be disordered using AIUPred[21]. V) The motif hit, although tested, did not bind to the annotated partner domain in ProP-PD[11] experiments. VI) False positive annotation in ELM. Each motif class has a different number of annotated instances (with experimental evidence) and motif hits in the proteome (based on how permissive the regular expression is). Using the flagged (likely false-positive) motif hits, five times more negative data was selected for each class than the associated positive data (Figure 1/B).

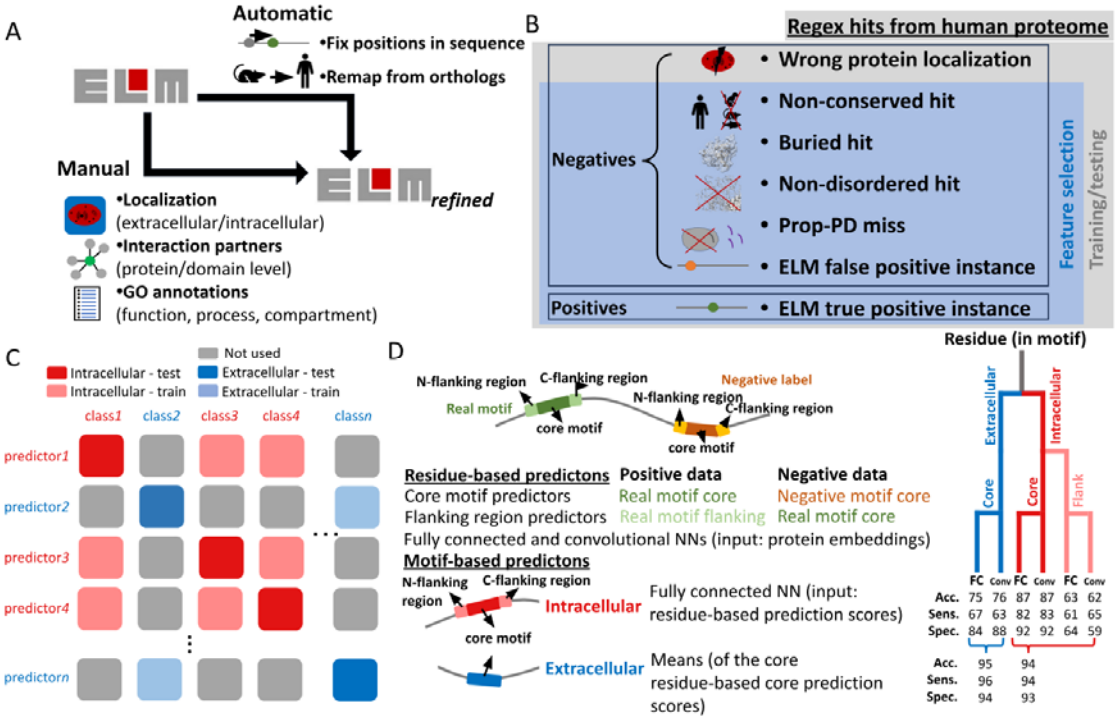

Figure 1. A) Automatic remapping of motif instances and manual refinement of ELM class annotations. B) Definition of labels used to train (and test) SLiMMine. C) Definition of predictors, training set and testing sets considering the motif class and the broad-level localization definition of the motif class. D) Prediction of SLiMs: Left: Schematic view of how different motif regions are defined for the predictors (including residue-level and the final motif-level predictors); Right: The accuracy of different neural networks.

Accurate prediction of Short Linear Motifs

Compared to the estimated number of SLiMs involved in cell regulation[4], the number of experimentally verified motifs is very low, complicating the training and objective benchmarking of prediction methods. To overcome this problem, we created a separate prediction for each motif class. The training data includes all (positive and negative) motif instances from the same cellular milieu (intracellular or extracellular) that do not belong to the selected class, and then the predictor was benchmarked on instances of the particular class (Figure 1/C), ensuring that the benchmark set is entirely independent from the training set. We used a Gaussian mixture model for data augmentation, thereby increasing the number of positive instances to match the number of negative examples.

For all predictions, the input specifies the start and end positions of the candidate motif within a protein. For intracellular motifs, we first built residue-level predictors for core motifs (using the positive and negative motifs as input) and flanking regions (where the positive data consisted of the flanking regions of experimentally verified motifs, and the negative data consisted of the core motif regions) (Figure 1/D). In each case, a fully connected and convolutional Neural Network (NN) was used: while the core motif predictors have 87% accuracy (Figure 1/D), the flanking region predictors have a somewhat limited usability (62-63% accuracy). The lower accuracy is not unexpected, as motifs often stack on each other (therefore, the flanking and core regions may overlap), and different motif classes rely on flanking regions to a different extent for specific binding. By incorporating all residue-level prediction results into a fully connected NN, the prediction accuracy reached 94% for intracellular motifs (Figure 1/D). For extracellular motifs, we used only core motif predictions (Figure 1/C), and the final score is the mean of the residue-level prediction results, with 95% accuracy (Figure 1/D).

We performed further tests to objectively evaluate the performance of SLiMMine. Using the benchmark set, we assessed the performance of a classical approach by incorporating protein disorder (from AIUPred) and conservation (for more details, see methods). We also benchmarked AIUpred-binding[13], a successor of Anchor[22]. Although it was not specifically developed for motif analysis, it is widely used, available both as a web server and a standalone program, and its authors have demonstrated its capability to identify functional motifs as well. SLiMMine outperforms both traditional conservation and protein disorder-based analysis and AIUPred-binding on our benchmark set (Figure 2/A). While the training/benchmark sets may be selected arbitrarily, to demonstrate the power of our method, we tested SLiMMine on further independent datasets: motif regular expression (regex) hits on monomeric structures (decoy set) and non-conserved disordered linker and terminal regions (neutral set) (Figure 2/B). The distribution of scores for monomeric structures shows that SLiMMine does not falsely identify motifs within these folded regions. The distribution of the SLiMMine prediction scores on disordered linker and terminal regions is uniform, in line with the regex hits being either true motifs or falsely identified ones in these regions. In contrast, scores on conserved LMPID[7] motif entries and ProP-PD[11] motif hits (positive sets) are high (Figure 2/B), showing that SLiMMine can also identify motifs that are independent from ELM instances.

Prediction of Short Linear Motifs across the human proteome

Considering all regex hits in the proteome, SLiMMine scores show a bimodal distribution (Figure 2/C, reference distribution on the benchmark set is visible on SFigure 1). While there are hundreds of thousands of motif regex hits in a grey zone where specificity is lower, the proportion of hits is higher above the 0.9 threshold, where the specificity is 0.98. Altogether, we predicted 304,238 SLiM instances with 0.9 or higher score, and 696,329 SLiM instances with 0.5 or higher score, among known motif classes. This methodology could filter out 2,718,778 regex hits as likely false-positives.

The power of SLiMMine prediction is demonstrated on *Fibronectin* which contains multiple integrin-binding motifs. These functional motifs are often located in loops connecting secondary structure elements of ordered domains, thus they are not necessarily disordered. In addition, the residue-level conservation value (using Shannon entropy) fluctuates along the Fibronectin sequence, making motif identification harder. Despite these difficulties, SLiMMine successfully identifies all experimental motifs. In contrast, decoy motif regex hits (meaning regex hits for intracellular motifs in extracellular or buried regions) were assigned with lower prediction scores (Figure 3/A).

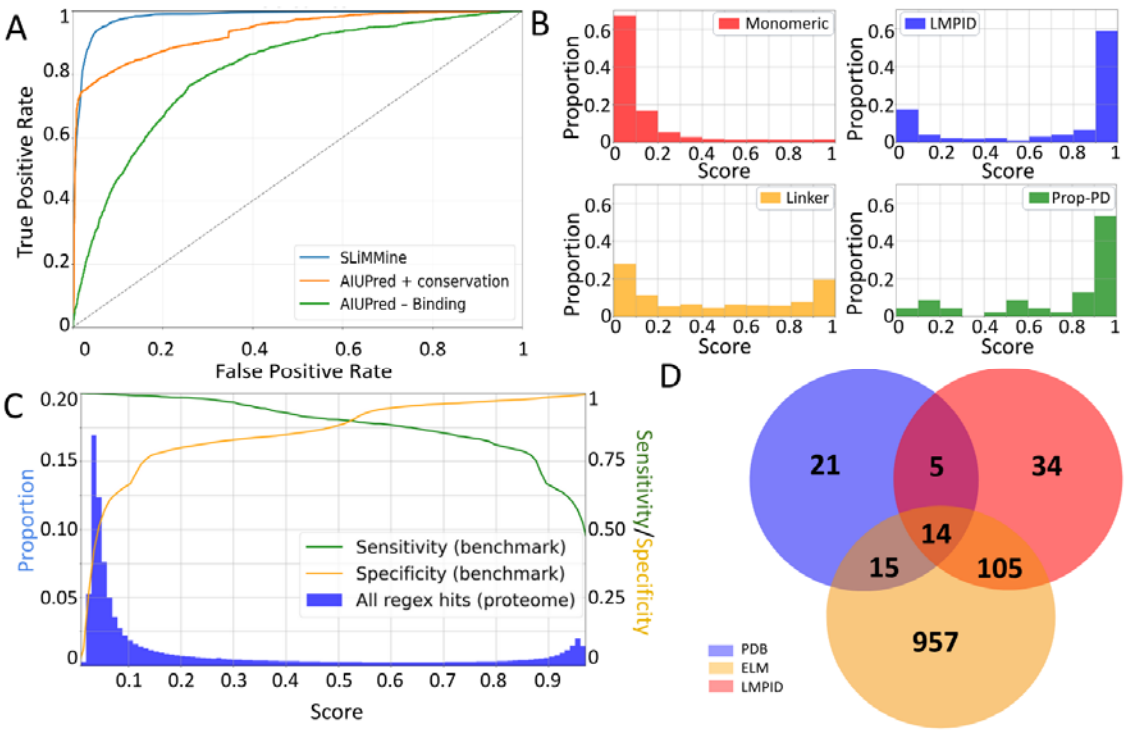

Figure 2. A) Receiver Operating Characteristics of SLiMMine, AIUPred-binding and a traditional prediction using conservation and protein disorder (AIUPred). B) Prediction score distribution on motif regex hits in monomeric structures (decoy set; top left), disordered linkers and termini (neutral set; bottom left), LMPID motifs (positive set; top right) and ProP-PD peptides (positive set; bottom right, for more details see Methods). C) Distribution of prediction scores for all motif regex matches in the human proteome (bars). Sensitivity and specificity at varying score thresholds are indicated by lines. D) De novo SLiM prediction hits on experimentally verified PDB, ELM and LMPID motifs.

To also showcase an intracellular example, in the SMAUG1 protein (UniProt AC: Q9UPU9), SLiMMine correctly identifies three of the four experimentally validated 14-3-3 binding motifs (centered around phosphosites S168, S254, and S665[23]), the nuclear export signal[24] (residues 691-705) and the C-terminal PDZ-binding motif (Dean group, unpublished results) with very high prediction scores (>0.9), although none of these motifs are included in ELM, so they were not present in the training set.

These results underscore that SLiMMine can reliably identify functional motifs of the already defined motif classes, while efficiently reducing the associated false positive hits.

SLiMMine enables the discovery of novel motifs that may not adhere to existing class definitions

Besides identifying novel instances for the already defined motif classes, SLiMMine also enables the discovery of *de novo* SLiMs that do not necessarily belong to the already defined classes, but are promising SLiM candidates based on sequence context. The reason behind is that SLiMMine operates on protein sequence embeddings rather than raw sequence patterns, so it does not memorize specific known motifs. Instead, it learns general properties characteristic of SLiMs as encoded in the embedding space. Notably, as the flanking region predictors could not be reliably calibrated for extracellular motifs, this functionality is only available for intracellular motifs.

While the core motif predictor of SLiMMine identifies longer regions as favourable for motifs, the flanking region prediction opens prospects to define novel motifs without defining the start and end positions. We rationalized that if the core motif scores are high, and the flanking region scores are low in the very same region, and in addition, the flanking region scores show an increase before and after this segment, the analysed region is probably a SLiM. Considering the constructed benchmark set, SLiMMine predicted *de novo* motifs (i.e. when the start and end positions were not given) with 0.98 specificity (and a lower 0.56 sensitivity) on motifs of 3-9 residues in length, with a standard error of 1.91 residues on motif boundaries (considering the start and end positions as defined by the regular expression). Applying SLiMMine with these thresholds on the full human proteome using all possible lengths, identified 32,501 protein regions. These short protein regions precisely map to over a thousand experimentally validated motifs from ELM, and several motifs from PDB and LMPID that were not part of the training set (Figure 2/D). For example, in *Stonin-2* (UniProt AC: Q8WXE9) SLiMMine could correctly identify the experimentally verified alpha-ear domain-binding motif[25] also stored in LMPID, even though it is not included in ELM and therefore was not part of the training set (Figure 3/B).

A powerful feature of our *de novo* method is that it can also identify atypical motif variants not covered by standard regular expressions. For instance, it correctly identified an experimentally validated CD2AP-binding motif within the TKS4 protein[26] as a *de novo* motif (UniProt AC: A1X283; residues 754-RPVVPPRR-761), even though it is not present in ELM and it does not completely fit the strict regular expression of the corresponding ELM motif class (class: LIG_SH3_CIN85_PxpxPR_1, regular expression: Px[PA]xPR). Other studies also support that the three SH3 domains of CD2AP have slightly different recognition specificities and some of them also accept valine as the middle residue of the recognized motif[27]. As another example, the N-terminus of Dynein 2 intermediate chain 2 (DYNC2I2; Q96EX3) can engage no less than 3 dynein light chain molecules in the dynein tail, forming 3 successive beta-sheets, according to the structure pdb:6RLB[28]). Here, our method successfully identified the 3rd segment as a novel motif, although its sequence

“HVDAQVQT” does not respect the canonical [KR]xTQT pattern. SLiMMine also finds the conserved, unconventional 14-3-3 binding motif in CREB-regulated transcription coactivator 1 (CRTC1; Q6UUV9), as evidenced by the structure pdb:7D8H[29], despite its complete lack of arginine/lysine residues (QYYGGSLP).

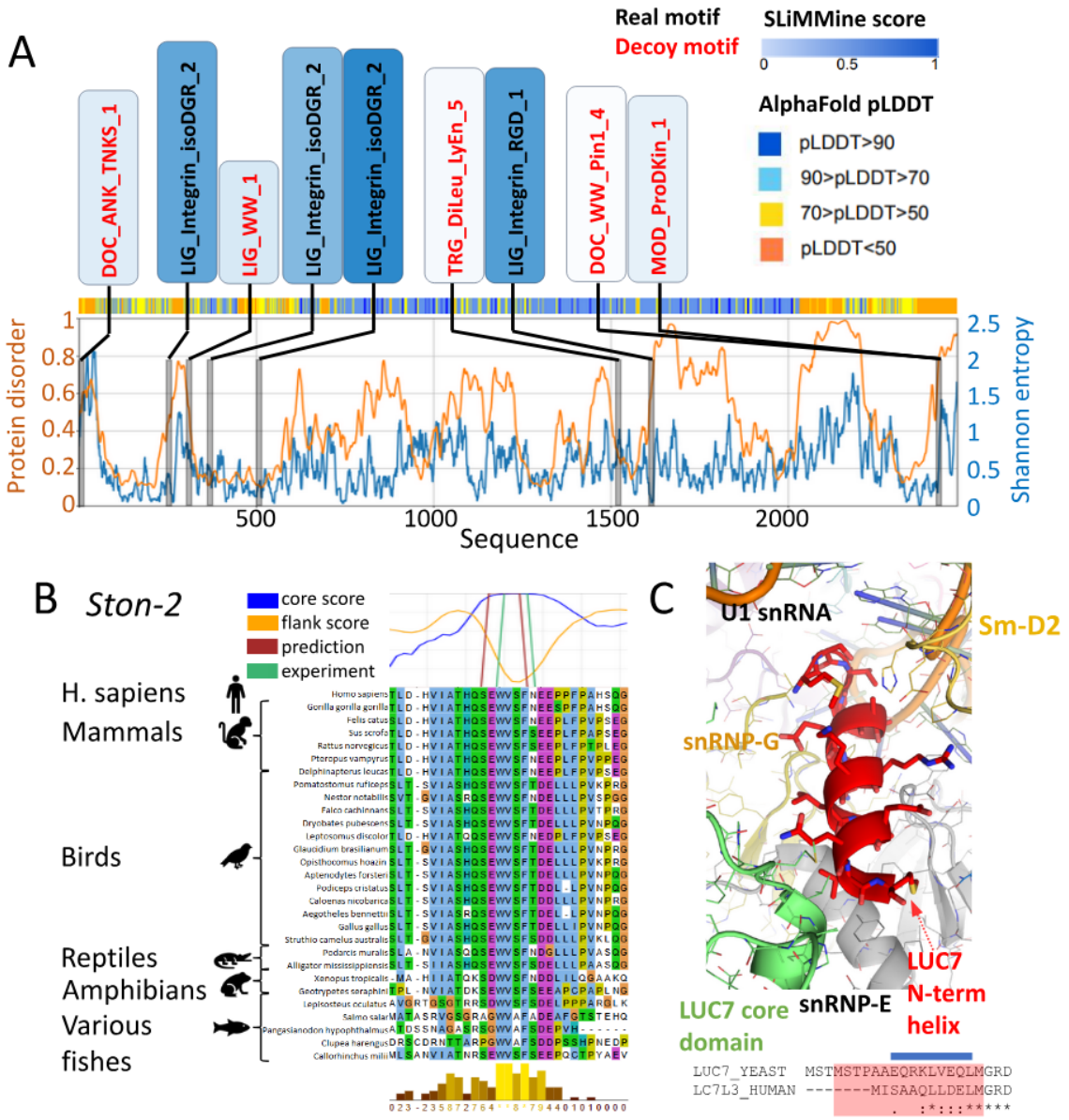

Figure 3: A) SLiMMine prediction results of known motif classes on Fibronectin using real and decoy motifs. AlphaFold pLDDT values, AIUPred output and residue conservation are also displayed. B) De novo prediction, alignment and protein conservation for a SLiM in the Stonin-2 protein, mediating the binding to the alpha-ear domain of AP-2 complex subunit alpha-1 or Alpha1-adaptin. C) Structural snapshot of the N-terminus of yeast LUC7 protein in the spliceosomal E complex (pdb:6N7R, in red), contacting multiple spliceosomal proteins (colored variously) as well as the U1 spliceosomal RNA (orange). The alignment with the N-terminus of the human LUC7L3 protein is shown below, with the red box highlighting the ordered amino acids observed in the crystal structure. The conserved segment, predicted as a de novo motif by our algorithm, is indicated by a blue bar.

Some *de novo* identified motifs are located in proteins that bind to large complexes (proteasome, ribosome or spliceosome) with decent structural evidence. Large ribosomal subunit protein eL39 (P62891) is a constitutive component of the large ribosomal subunit, and its predicted motif directly engages RNA strands in the structure pdb:8FLA[30]. We have also observed that the flexible N-terminal region of the human LUC7L3 protein (O95232), predicted as a novel motif by our method, is also preserved in its yeast homolog LUC7. In the crystal structure of the Saccharomyces cerevisiae spliceosomal E complex (pdb:6N7R[31]), this segment forms a short helix, binding simultaneously to U1 spliceosomal RNA, as well as proteins snRNP-E, snRNP-G and Sm-D2 (Figure 3/C). However, the precise arrangement of this peptide segment might differ slightly between human and yeast spliceosomes, due to their different lengths and limited conservation.

In all, SLiMMine can successfully, and, regarding sequence boundaries, very precisely identify SLiMs that were not present in its training set and may not even adhere to the SLiM definitions it has been trained on.

SLiMMine predictions can be linked to protein-protein interaction (PPI) data and explain underlying binding mechanisms

After predicting novel instances for the ELM-derived SLiM classes in the human proteome, we also created a network of potential interactors using the newly defined sets of partner proteins created for the classes (refined based on the ELM database's descriptions and associated literature as described in the the 1st section of the Results). For each predicted motif, experimentally verified PPIs of the given protein (from PDB (through 3did[32]), IntAct[33] and BIOGRID[34]) were screened to identify partners that were among the predefined partners of the respective motif class. This procedure yielded a list of likely SLiM-mediated interaction partners for each protein with predicted motifs and, at the same time, allowed us to suggest underlying interaction mechanisms and interacting regions for a vast number of human PPIs, for which no such data were previously available.

SLiMMine comes with a user-friendly web server for easy motif discovery

SLiMMine is shared with the research community through a modern, easy-to-use web resource. To cover all possible use cases, the SLiMMine web server provides three different options for motif search. I) Users can search for any protein in the human reference proteome; II) Users can also search for predefined ELM classes. This provides a list of proteins that contain sequence match(es) to the regular expression of the selected motif class along with the associated SLiMMine prediction probabilities; III) Custom regular expressions can also be searched, although the context (intracellular/extracellular) needs to be defined by the user. The server will list all human proteins that match the given regular expression along with predicted probabilities. For class and regular expression searches, results can be downloaded as a TSV file.

All search results can be further inspected on the entry page, which has five main blocks: General, Motif (viewer), motif details, protein interactions (viewer) and protein interaction details. The General section displays general information about the protein, including subcellular localisation, domains, and statistics.

The motif viewer (Figure 4/A) was built in a way to summarize all the relevant information about the predicted motifs along the sequence, also displaying evidence-based filters that help the user decide whether the motifs are functional. At the top of the panel, measures related to protein disorder are shown. Each motif class is shown below in a separate line. Classes that could be linked to experimental PPIs are highlighted. The boxes representing the motifs are coloured and have frames that correspond to different evidence supporting the actual motif. For example, motifs with contradicting evidence (e.g., a phosphodependent motif without experimental phosphorylation evidence[35]) are crossed. Notably, when we defined these contextual filters, we found that they are not likely to rule out functional motifs (SFigure 2). Below the viewer, there is a table that details supporting (or contradicting) evidence on the motifs.

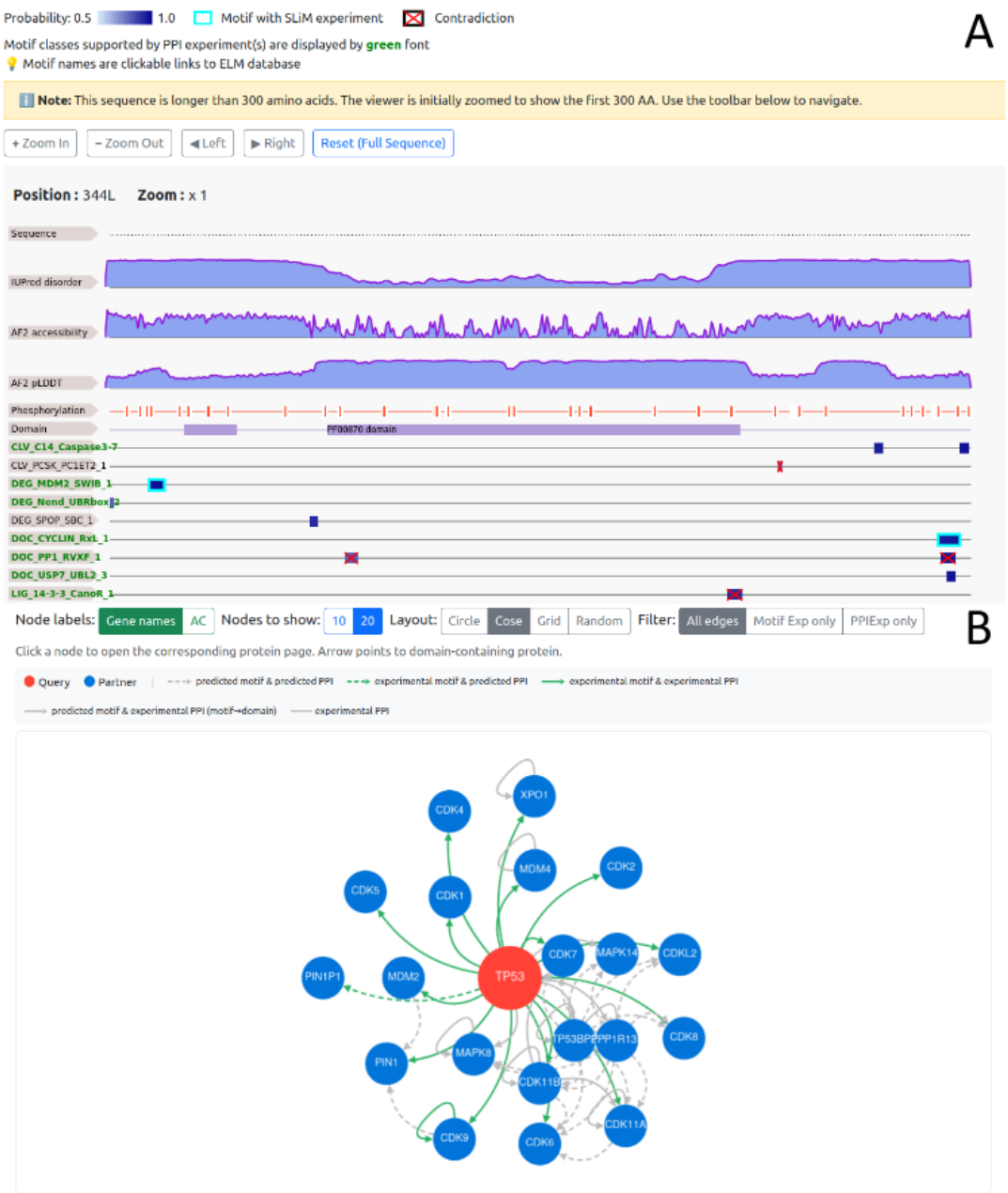

Figure 4: A) *Layout of the sequence viewer and B) the Protein interaction viewer.*

The protein interaction network (Figure 4/B) is largely based on the partner lists for the SLiM classes and displays the best annotated SLiM-mediated interactions for the entry protein. Predicted and experimentally supported motifs, as well as protein interactions can be viewed and filtered. The table below details all the possible SLiM-mediated interactions of the entry protein (all predicted SLiM classes with all their possible partners listed in the refined SLiM class annotations), with supporting evidence also displayed where available.

Discussion

SLiMMine is an alignment-free approach for predicting SLiMs in protein sequences. It is optimized for scenarios in which (approximate) motif boundaries are known, for example, after scanning sequences with regular expressions derived from motif databases such as ELM. Under these conditions, the model can focus on evaluating whether a given candidate region exhibits properties consistent with functional SLiMs. Although the method was trained and optimized using human protein data, we expect that it will perform similarly well on other mammalian proteomes. These sequences share similar characteristics, which are captured by protein embeddings. For more distant taxonomic groups, such as plants, fungi, or bacteria, further benchmarking will be required.

A major obstacle during the development was the limited availability of experimentally validated motif data. Positive samples could be retrieved from ELM[6,36], however, in the case of other resources, the annotation process was not strict enough to grant that only true, functional motifs get included. Defining true negatives was more challenging, as the absence of a motif annotation does not imply the absence of function (discussed in more detail in the SMaterial). This challenge is particularly acute for non-mammalian species, where curated motif annotations and interaction datasets remain sparse. However, the growing number of high-throughput experimental approaches is expected to eliminate this limitation in the long term. Techniques such as proteomic ProP-PD[10] have already expanded the known interaction space for several SLiM-binding domains and have demonstrated their ability to uncover motifs across viral[37] and host proteomes[11]. As coverage of such experiments increases - especially for viruses and other underrepresented taxa - these data will provide valuable benchmarks for evaluating and retraining models such as SLiMMine, thereby improving prediction accuracy across the tree of life. Notably, SLiMMine was trained and tested on conserved motifs, however pathogenic (especially viral) motifs hijacking the host cell can appear across evolution[38], hence they are not always conserved, therefore their prediction may need further considerations.[6]

SLiMMine classifies more than 80% of the proteome-scale sequence matches of ELM-derived regular expressions as false positives (with prediction scores below 0.5). However, it is important to note that even among the predicted positives, erroneous hits may still be present. One contributing factor to false predictions is the heterogeneity of motif definitions even within the ELM resource. Some motif classes, particularly those with highly permissive regular expressions, have high pattern probabilities and therefore occur thousands of times within the human proteome (in almost every protein). In such cases, sequence-based prediction alone is insufficient to distinguish functional instances from coincidental matches. To mitigate this issue, we introduced refined and extended ELM annotations that incorporate contextual constraints, such as phosphorylation dependence, subcellular localization, or restriction to transmembrane proteins. For example, motifs that are known to function only when phosphorylated, or that are restricted to membrane

proteins, are explicitly annotated as such. Within the SLiMMine web resource, predicted motifs that contradict with these contextual annotations are highlighted on the results webpage and the contradicting features are also displayed. Notably, the human proteome is not uniformly annotated, therefore, it is up to the users to decide whether they accept the neural network score or rely on these annotations that can support or contradict the motif.

Beyond identifying candidate SLiMs, SLiMMine aims to provide additional biological insight by proposing likely interaction partners for predicted motifs. By linking motifs to their potential interaction partners, SLiMMine enables functional hypothesis generation. The combined information can serve as a powerful input for both computational and experimental studies. From a computational perspective, this information can also substantially reduce the search space in large-scale interaction modeling attempts. For example, in high-throughput AlphaFold-based domain–motif screening approaches[39], prior knowledge of plausible motif–domain pairs can dramatically lower computational cost and increase interpretability. Rather than modeling all possible domain–peptide combinations, researchers can focus on a curated subset supported by SLiMMine predictions.

From an experimental standpoint, motif–partner predictions can guide the design of targeted validation assays. Peptide pulldowns, mutational analyses, or cell-based interaction assays can be optimized by prioritizing motifs with high prediction confidence. By narrowing down candidate interactions, SLiMMine can help bridge the gap between large-scale in silico prediction and focused experimental validation.

An important extension of SLiMMine is the *de novo* motif prediction module, which is currently released as a beta feature. Unlike the primary workflow, which evaluates candidate motifs defined by regular expressions, the *de novo* approach seeks to identify SLiM-like regions without predefined boundaries. This idea is motivated by the observation that many SLiMs exhibit “island-like” conservation patterns[40]: a short core region shows strong functional constraint, while flanking residues are more variable but still contribute to binding specificity or regulation. We assume that our neural network scores can capture this behaviour. To further support this concept, we introduced a complementary prediction strategy that explicitly evaluates the relationship between the predicted motif core and its surrounding residues. By combining core-centric and flank-centric predictions, the method can highlight regions without defining the start and end position of a segment. This approach can be valuable for identifying novel motif classes and atypical binding modes that are currently underrepresented in curated databases. Notably, in proteins where multiple motifs overlap, the interpretation of flanking regions becomes ambiguous.

SLiMMine can serve as a starting point for addressing biologically relevant questions. Many regulatory proteins contain multiple SLiMs that engage several binding partners simultaneously or sequentially, leading to cooperative or competitive interaction networks[41]. Multivalency can increase effective binding affinity and enable dynamic assembly of signaling complexes. In such systems, individual SLiMs may exhibit weak affinities, yet collectively, they drive strong and specific interactions[1]. In addition, clustering of multiple SLiMs within intrinsically disordered regions may contribute to liquid–liquid phase separation (LLPS), in which networks of weak, multivalent interactions drive the formation of dynamic biomolecular condensates that spatially organise signaling and regulatory processes[42–45]. SLiMMine is well-suited to exploring this complexity, as it can identify multiple candidate motifs within the same protein and link them to distinct or shared binding domains.

SLiM prediction has important implications for understanding human diseases. Although disease-causing mutations are less frequent in disordered regions[46,47], many disease-associated mutations disrupt SLiMs. Loss (or gain) of a single motif can rewire

interaction networks, leading to aberrant signaling, mislocalization, altered protein turnover or stability[48]. By highlighting predicted motifs and their potential interaction partners, SLiMMine can aid in the interpretation of missense variants that fall outside classical functional domains. In addition, pathogens often utilize SLiM mimicry to evade immune responses by modulating cell regulation and suppressing defense mechanisms, resulting in modified physiology. Pathogens can interfere with a diverse range of cellular functions, including entry to the host cell[49], transport to their final destinations, manipulation of the cell cycle to replicate their DNA, and other processes[50,51]. Therefore, analysing the predicted SLiMs can also pinpoint vulnerable regions within human proteins.

In all, with its accurate predictions, SLiMMine represents a large step forward in identifying and understanding true short linear motifs in human proteins. Furthermore, it represents a flexible framework that can be (and will be) updated by new data (data from novel SLiM databases, new motif class definitions, large-scale experimental datasets providing both true positive and false positive motif data) and extended, e.g. towards the proteins of other organisms. Besides predicting novel instances for already defined SLiM classes, SLiMMine could also identify ~32,500 de novo SLiMs in the human proteome. Among these, we could identify several experimentally validated motifs that were not part of the training data and may not fit existing motif definitions used for training. Furthermore, the reliable motif predictions of SLiMMine and refined partner annotations created for hitherto known motif classes enabled rendering likely binding mechanisms and interacting regions to a large number of human PPIs, as well as highlighting novel motif-mediated PPIs with high probability. The associated user-friendly, multi-purpose web server represents a powerful tool with a potential for taking motif discovery endeavours to the next level.

Methods

Manual curation of data from the ELM Resource

We have refined and extended the annotations of the 319 motif classes of the ELM database that are relevant for human proteins. For each class, we determined if the motif-containing protein counterparts 1) are extracellular or intracellular, 2) if intracellular, what compartment they are expected to localize to, and 3) what are their expected GO[18] biological processes and molecular functions. In addition, whenever possible, we precisely determined the possible binding partners for the given SLiM class, i.e. the subset of the proteins carrying the specific binding domain that was described to actually bind to the given SLiM type. For this, either the same (usually PFAM) domain identifier was used that was provided in ELM (if further specification of the partners was not required), or a more specific domain identifier from InterPro[19], or the specific list of UniProt ACs of the possible interaction partners within the human canonical proteome. We also defined if a motif is only present in proteins with a particular domain, or in transmembrane proteins. Last but not least, we catalogued all motif classes that 1) require phosphorylation for binding or 2) result in phosphorylation (as a product of binding), so that available proteome-scale phosphorylation data could be used to support or contradict the functionality of the associated predicted instances (STable 1).

Automatic processing of ELM data

Motif instances and classes were downloaded from the Eukaryotic Linear Motif Resource in November 2024[6]. We checked if instance boundaries in *H. sapiens, M. musculus,* and *R. norvegicus* proteins were altered by mapping them to the respective canonical sequences in UniProt[8] (release: 2025 January). If the motif could not be found at the given position, but the sequence still carried the indicated peptide in a single copy within a disordered region (AIUPred[21] > 0.4), we modified the region boundaries accordingly. In addition, we aligned *M. musculus* and *R. norvegicus* proteins with their *H. sapiens* orthologs (sharing the same gene name) and transferred the motif boundaries from orthologous sequences onto the human counterparts (STable 2).

Defining motif labels

The positive set was defined using all true positive human instances in ELM, extended with those mapped from orthologous species. The negative set contains false positive human instances from ELM. As this set is minimal, we scanned the human proteome for the regular expressions stored in ELM, and defined the following labels. I) Residues of the putative motif are also part of a PDB monomeric structure (defined by SIFTS[52]), and their mean relative solvent accessible area is lower than 0.09. II) The regular expression defining the putative motif had lower than 0.3 conservation within the multiple sequence alignment (MSA) in a set of Chordata species (STable 3) (see Conservation) III) If the localization of the protein rules out the functionality of the putative motif (considering cytoplasmic, nuclear and extracellular motifs). IV) The putative motif is in an ordered region (average AIUPred score across the putative motif residues was below 0.3). V) The respective SLiM-binding domain was used as bait in ProP-PD experiments, but the peptides carrying the putative motif were not found as positive hits[11].

The ELM Resource categorizes motifs into classes (based on interacting with the same binding site of a specific partner domain and the function), and each class contains multiple experimentally validated motif instances. For each class, the positive set included all positive instances from that class. The negative set, on the other hand, included five times more regex hits associated with the same class than the positive data. The negative set was selected to ensure a roughly equal representation of the following cases: I) ordered putative motifs, II) non-conserved putative motifs, and III) putative motifs from proteins with an inappropriate subcellular localization. In addition, less populated negative cases, including false-positive ELM instances, motifs buried within monomeric structures and ProP-PD[11] negative results (where the peptide carried the motif pattern but it was not bound by the corresponding SLiM-binding domain), were included (STable 4). Notably, this ratio may differ depending on the regular expression definition: if a motif pattern only occurred in a limited number of proteins, fewer negative samples were used.

Conservation

We defined motif conservation in two ways. If a regular expression was available, we used a set of Chordata species proteomes (STable 3) and searched orthologs for human proteins using BLAST[53] (the sequence with the highest sequence identity (above 20%) was selected for each species, if it covered at least 60% of the human sequence), and then aligned the selected sequences with ClustalΩ[54]. We searched for the regular expression in the orthologous proteins. Since the alignment of disordered regions is challenging, we

allowed 50 residues deviation in the alignment for regular expression hits. Non-conserved motifs were defined with <0.3 conservation, while conserved motifs were defined with >0.7 conservation across orthologs. When the regular expression was not available, we simply calculated the Shannon entropy for each position in the alignment. Residue positions with <1 Shannon entropy were considered as conserved.

Train and test sets

For each motif class, we defined a broad-level localization definition (extracellular/intracellular), see (STable 1). We created a separate predictor for each class (STable 5), where the test set included all instances from a given class. In contrast, the training set contained instances from all the other classes with the same broad-level localization. This way, it was possible to perform independent tests for each 'class' predictor.

Architectures and training

The goal of the method is to take a short protein segment (usually defined by a regular expression) and predict whether that given segment shows the characteristics of SLiMs. We created multiple predictors for each class using different regions of the motif and different neural network (NN) architectures. First, we separated extracellular and intracellular motifs. In both cases, we defined a core motif predictor (prediction on the actual motif, relying on the previously defined dataset). In addition, for intracellular motifs, we developed a flanking region predictor (for the extracellular flanking regions the NN did not converge), which was defined as five neighbouring residues on both sides of the actual motif. For the flanking predictor, the negative data consisted of the actual motif (see Figure 1/D). The training data is based on the prottrans_t5_xl_u50 embeddings[55] of residues in the motif, and these predictions are performed on the individual residues. For each predictor, we randomly selected 10% of the positive and negative data and used recursive feature elimination to select the top 100 features. Notably, for this step, we omitted all negative instances where the negative annotation was derived from the incorrect localization of the protein, as it seemed that, in such cases, the embedding-based predictors learnt the localization of the motif and did not consider any other biophysical or biological features. Nevertheless, the feature selection was performed 50 times, and features appearing in 50% of the cases were selected for all prediction types (i.e. intracellular core, intracellular flanking and extracellular). As the training set contained five times more negative data, we augmented positive data by adding Gaussian noise to it (on the level of embedding vectors) with a 5% standard deviation. Then, we used both fully connected and convolutional NNs for core and flanking region predictions, as well. In the case of intracellular motifs, a fully connected NN was used to predict the final result for a given segment, using the residue-level core and flanking NN predictions as input. For extracellular motifs, the final prediction is the mean of the residue-level fully connected and convolutional NNs. The architecture of these predictors is visible on SFigure 3. The results of the residue-level predictions on the training and test sets are available in STables 6-13. Final prediction results on the training and test sets are available in STables 14-17. To verify that our method is not overfitting, we used the same procedure with mixed labels too. As expected, the NNs could not converge and produced ~45-55% prediction accuracy.

Independent test sets

The constructed train/test sets were also used to benchmark traditional conservation/disordered-based motif prediction. We used AIUPred[21] results (calculating the mean over the residues within the motif) and a conservation score from Chordata species (see “Defining motif labels”). The mean of these two values was used. We also compared SLiMMine with AIUPred-binding[13] (STable 18, Figure 1/A).

We also used other data sources on motifs to test the robustness of our method on intracellular proteins (“cytosol”, “cytoplasm”, and “nucleus” annotations in GO[18]): 1) While ELM[6] has a very rigorous annotation process, we could not verify this for LMPID[7]; therefore, to ensure that only functional motifs are used, we further filtered LMPID segments for conservation. Files available for download in LMPID do not contain the regular expression, therefore, we used Shannon entropy to define conservation. 2) We also selected high-confidence peptide hits from Prop-PD[11] (HD2 confidence). These 16 amino acids long peptides likely include long flanking regions that might confuse the prediction. Therefore, we searched for the regular expression within the prey peptides that is known to bind to the SLiM-binding domain used as bait (STable 19). Since, again, we could not be sure that these are true functional hits, we filtered them for conservation. In this case, the identified regular expression was subjected to conservation analysis (see Conservation section above).

We defined a decoy set by searching motif candidates with regular expressions in the ordered monomer test set also used for benchmarking Anchor[13]. In addition, we analyzed disordered linker and terminal regions from DisProt as a neutral set (covered by the functional annotation categories: “flexible C-terminal tail”, ”flexible N-terminal tail”, ”entropic chains”, and “flexible linker/spacer”) by pattern matching of ELM class regular expressions for motif candidates. However, to reduce the inclusion of potentially functional motifs in the DisProt set, we retained only motifs with a conservation score below 0.3.

In the above described LMPID, Prop-PD, DisProt[56] and ordered monomers datasets, we only considered hits that did not overlap with any known ELM instance to ensure independence from the training/test sets. Prediction results on these sets are available in STable 20.

De novo motif prediction

Using the constructed benchmark set, we tried multiple score cutoffs to define the most reliable *de novo* motif prediction. Using intracellular motifs from the training and test sets, we benchmarked how different flanking regions and motif scores perform on motifs between 3 and 9 residues. STable 21 contains the sensitivity, specificity, and extent to which the region boundaries of the observed and predicted motifs differ for these thresholds. We achieved the best overall result when the score from the flanking region predictors on the core motif divided by the score from the flanking region predictors on the flanking region is below 0.5, and the final motif prediction score is above 0.9 (STable 22).

Protein-protein interaction data

For all predicted motif instances, we considered the list of annotated, possible partners of the respective SLiM class as possible interactors [STable1]. If an interaction was validated in BIOGRID[34], IntAct[33] (with physical association or stronger annotation (MI:0915)), or it was included in the PDB (accessed via 3did[32]), ELM or LMPID, we accepted it as an

experimentally verified protein interaction. The resulting subnetworks are displayed on the SLiMMine web server entry pages.

Web server for SLiMMine

SLiMMine is implemented as a web application using Laravel 12, PHP 8.4, and MySQL 8.0. The platform integrates multiple predictive models, combining machine learning outputs with AlphaFold2-derived pLDDT confidence scores[57], AIUPred disorder probabilities[21], and UniTmp[58] transmembrane topology annotations. Query performance is optimized through database-level pagination and pre-indexed tables, enabling efficient access to large datasets.

The interactive visualization module uses the FeatureViewer JavaScript library with jQuery 3.6 and Bootstrap 5.3, ensuring responsive behaviour across devices. Batch data export is supported in tab-separated format, including the underlying prediction probabilities. Benchmarking across the complete human proteome (>20,000 entries) demonstrated sub-second response times, achieved through optimized SQL statements and server-side caching.

Acknowledgment

This work builds on two decades of effort behind the Eukaryotic Linear Motif (ELM) resource. We thank Prof. Toby Gibson for his leadership and all the ELM annotators for their invaluable contributions.

Funding

The project was implemented with the support from the National Research, Development and Innovation Fund of the Ministry of Culture and Innovation, financed under the K-146314, FK-142285 and PD-146564 funding schemes granted to G.E.T, R.P. and L.D, respectively. R.P is a holder of the János Bolyai Research Fellowship of the Hungarian Academy of Sciences (BO/00174/22).

Declaration of interest

E.F. is an employee of Cytocast Hungary Kft. Furthermore, E.F. owns equities or stocks of the company.

Supplementary material

STable 1: Refined manual annotations of ELM resource motif classes (localization, binding partners and gene ontology)
STable 2: ELM motif instances subjected to automatic processing (transfer from ortholog, fixing region boundaries)
STable 3: List of species used for creating alignments and calculating conservation
STable 4: Train and test sets of SLiMMine

STable 5: Motif classes used to benchmark different predictors
STable 6: Prediction results on the core ‘inside’ motifs (convolutional NNs, residue level prediction) available at https://zenodo.org/records/18785791
STable 7: Prediction results on the core ‘inside’ motifs (fully connected NNs, residue level prediction) available at https://zenodo.org/records/18785791
STable 8: Prediction results on the flanking regions of ‘inside’ motifs (convolutional NNs, residue level prediction) available at https://zenodo.org/records/18785791
STable 9: Prediction results on the flanking regions of ‘inside’ motifs (fully connected NNs, residue level prediction) available at https://zenodo.org/records/18785791
STable 10: Prediction results on the core ‘outside’ motifs (convolutional NNs, residue level prediction) available at https://zenodo.org/records/18785791
STable 11: Prediction results on the core ‘outside’ motifs (fully connected NNs, residue level prediction) available at https://zenodo.org/records/18785791
STable 12: Summary statistics for residue level predictors (separated based on SLiM class used for testing)
STable 13: Summary statistics for residue level predictors
STable 14: Prediction results on ‘inside’ motifs (motif level prediction) available at https://zenodo.org/records/18785791
STable 15: Prediction results on ‘outside’ motifs (motif level prediction) available at https://zenodo.org/records/18785791
STable 16: Summary statistics for motif level predictors (separated based on SLiM class used for testing)
STable 17: Summary statistics for motif level predictors
STable 18: Comparison of conservation/disordered based prediction and SLiMMine
STable 19: Correspondence between Prop-PD domains and ELM classes
STable 20: Further independent test sets (monomer structure, DisProt linkers, LMPID and Prop-PD)
STable 21: Sensitivity, specificity and motif boundary statistics for *de novo* predicted motifs at different thresholds
STable 22: *de novo* motif prediction results on the benchkark set with selected thresholds

SFigure 1: Distribution of scores on the benchmark set, with performance metrics at different thresholds.
SFigure 2: Contradictions on experimental motifs
SFigure 3: Neural Network architectures.

SMaterial: Extended discussion on constructing the training set

Codes for slimmine are available at: https://zenodo.org/records/18786455

References

1. Van Roey, K. *et al.* Short linear motifs: ubiquitous and functionally diverse protein interaction modules directing cell regulation. *Chem. Rev.* **114**, 6733–6778 (2014).
2. Davey, N. E., Cyert, M. S. & Moses, A. M. Short linear motifs - ex nihilo evolution of protein regulation. *Cell Commun Signal* **13**, 43 (2015).
3. Davey, N. E. The functional importance of structure in unstructured protein regions. *Curr Opin Struct Biol* **56**, 155–163 (2019).
4. Tompa, P., Davey, N. E., Gibson, T. J. & Babu, M. M. A million peptide motifs for the molecular biologist. *Mol Cell* **55**, 161–169 (2014).
5. Davey, N. E. *et al.* Attributes of short linear motifs. *Mol. Biosyst.* **8**, 268–281 (2012).
6. Kumar, M. *et al.* ELM-the Eukaryotic Linear Motif resource-2024 update. *Nucleic Acids Res* **52**, D442–D455 (2024).
7. Sarkar, D., Jana, T. & Saha, S. LMPID: a manually curated database of linear motifs mediating protein-protein interactions. *Database (Oxford)* **2015**, (2015).
8. UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Res* **53**, D609–D617 (2025).
9. Ivarsson, Y. *et al.* Large-scale interaction profiling of PDZ domains through proteomic peptide-phage display using human and viral phage peptidomes. *Proc Natl Acad Sci U S A* **111**, 2542–2547 (2014).
10. Davey, N. E., Simonetti, L. & Ivarsson, Y. ProP-PD for proteome-wide motif-mediated interaction discovery. *Trends Biochem Sci* **47**, 547–548 (2022).
11. Benz, C. *et al.* Proteome-scale mapping of binding sites in the unstructured regions of the human proteome. *Mol Syst Biol* **18**, e10584 (2022).
12. Gogl, G. *et al.* Quantitative fragmentomics allow affinity mapping of interactomes. *Nat Commun* **13**, 5472 (2022).
13. Erdős, G., Deutsch, N. & Dosztányi, Z. AIUPred - Binding: Energy Embedding

to Identify Disordered Binding Regions. *J Mol Biol* **437**, 169071 (2025).

14. Jahn, L. R., Marquet, C., Heinzinger, M. & Rost, B. Protein embeddings predict binding residues in disordered regions. *Sci Rep* **14**, 13566 (2024).

15. Chow, C. F. W. *et al.* SHARK-capture identifies functional motifs in intrinsically disordered protein regions. *Protein Sci* **34**, e70091 (2025).

16. Krystkowiak, I. & Davey, N. E. SLiMSearch: a framework for proteome-wide discovery and annotation of functional modules in intrinsically disordered regions. *Nucleic Acids Res* **45**, W464–W469 (2017).

17. Reys, V., Pons, J.-L. & Labesse, G. SLiMAn 2.0: meaningful navigation through peptide-protein interaction networks. *Nucleic Acids Res* **52**, W313–W317 (2024).

18. Gene Ontology Consortium *et al.* The Gene Ontology knowledgebase in 2023. *Genetics* **224**, (2023).

19. Blum, M. *et al.* InterPro: the protein sequence classification resource in 2025. *Nucleic Acids Res* **53**, D444–D456 (2025).

20. Burley, S. K. *et al.* Updated resources for exploring experimentally-determined PDB structures and Computed Structure Models at the RCSB Protein Data Bank. *Nucleic Acids Res* **53**, D564–D574 (2025).

21. Erdős, G. & Dosztányi, Z. AIUPred: combining energy estimation with deep learning for the enhanced prediction of protein disorder. *Nucleic Acids Res* **52**, W176–W181 (2024).

22. Dosztányi, Z., Mészáros, B. & Simon, I. ANCHOR: web server for predicting protein binding regions in disordered proteins. *Bioinformatics* **25**, 2745–2746 (2009).

23. Carey, O. *et al.* Condensate Formation of the Human RNA-binding Protein SMAUG1 is Controlled by its Intrinsically Disordered Regions and Interactions with 14-3-3 Proteins. *J Mol Biol* **437**, 169314 (2025).

24. Kırlı, K. *et al.* A deep proteomics perspective on CRM1-mediated nuclear export and nucleocytoplasmic partitioning. *Elife* **4**, (2015).

25. Jha, A. *et al.* A novel AP-2 adaptor interaction motif initially identified in the long-splice isoform of synaptojanin 1, SJ170. *J Biol Chem* **279**, 2281–2290 (2004).

26. Kurilla, A. *et al.* Studying the Association of TKS4 and CD2AP Scaffold Proteins and Their Implications in the Partial Epithelial-Mesenchymal Transition (EMT) Process. *Int J Mol Sci* **24**, (2023).

27. Rouka, E. *et al.* Differential Recognition Preferences of the Three Src Homology 3 (SH3) Domains from the Adaptor CD2-associated Protein (CD2AP) and Direct Association with Ras and Rab Interactor 3 (RIN3). *J Biol Chem* **290**, 25275–25292 (2015).

28. Toropova, K. *et al.* Structure of the dynein-2 complex and its assembly with intraflagellar transport trains. *Nat Struct Mol Biol* **26**, 823–829 (2019).

29. Chen, H., Zhang, H., Chen, P. & Xiang, S. Structural Insights into the Interaction Between CRTCs and 14-3-3. *J Mol Biol* **433**, 166874 (2021).

30. Vanden Broeck, A. & Klinge, S. Principles of human pre-60 biogenesis. *Science* **381**, eadh3892 (2023).

31. Li, X. *et al.* A unified mechanism for intron and exon definition and back-splicing. *Nature* **573**, 375–380 (2019).

32. Mosca, R., Céol, A., Stein, A., Olivella, R. & Aloy, P. 3did: a catalog of domain-based interactions of known three-dimensional structure. *Nucleic Acids Res* **42**, D374–9 (2014).

33. Del Toro, N. *et al.* The IntAct database: efficient access to fine-grained molecular interaction data. *Nucleic Acids Res* **50**, D648–D653 (2022).

34. Oughtred, R. *et al.* The BioGRID interaction database: 2019 update. *Nucleic Acids Res* **47**, D529–D541 (2019).

35. Ochoa, D. *et al.* The functional landscape of the human phosphoproteome. *Nat Biotechnol* **38**, 365–373 (2020).

36. Kumar, M. *et al.* The Eukaryotic Linear Motif resource: 2022 release. *Nucleic Acids Res* **50**, D497–D508 (2022).

37. Mihalič, F. *et al.* Large-scale phage-based screening reveals extensive pan-viral mimicry of host short linear motifs. *Nat Commun* **14**, 2409 (2023).

38. Glavina, J., Palopoli, N. & Chemes, L. B. Evolution of SLiM-mediated hijack functions in intrinsically disordered viral proteins. *Essays Biochem* **66**, 945–958 (2022).

39. Lee, C. Y. *et al.* Systematic discovery of protein interaction interfaces using AlphaFold and experimental validation. *Mol Syst Biol* **20**, 75–97 (2024).

40. Davey, N. E. *et al.* SLiMPrints: conservation-based discovery of functional motif fingerprints in intrinsically disordered protein regions. *Nucleic Acids Res* **40**, 10628–10641 (2012).

41. Van Roey, K. & Davey, N. E. Motif co-regulation and co-operativity are common mechanisms in transcriptional, post-transcriptional and post-translational regulation. *Cell Commun Signal* **13**, 45 (2015).

42. Borcherds, W., Bremer, A., Borgia, M. B. & Mittag, T. How do intrinsically disordered protein regions encode a driving force for liquid-liquid phase separation? *Curr Opin Struct Biol* **67**, 41–50 (2021).

43. Case, L. B., Ditlev, J. A. & Rosen, M. K. Regulation of Transmembrane Signaling by Phase Separation. *Annu Rev Biophys* **48**, 465–494 (2019).

44. Banani, S. F., Lee, H. O., Hyman, A. A. & Rosen, M. K. Biomolecular condensates: organizers of cellular biochemistry. *Nat Rev Mol Cell Biol* **18**, 285–298 (2017).

45. Zeng, M. *et al.* Reconstituted Postsynaptic Density as a Molecular Platform for Understanding Synapse Formation and Plasticity. *Cell* **174**, 1172–1187.e16 (2018).

46. Pajkos, M., Mészáros, B., Simon, I. & Dosztányi, Z. Is there a biological cost of protein disorder? Analysis of cancer-associated mutations. *Mol Biosyst* **8**, 296–307 (2012).

47. Dobson, L., Mészáros, B. & Tusnády, G. E. Structural Principles Governing Disease-Causing Germline Mutations. *J Mol Biol* **430**, 4955–4970 (2018).

48. Kliche, J. *et al.* Proteome-scale characterisation of motif-based interactome

rewiring by disease mutations. *Mol Syst Biol* **20**, 1025–1048 (2024).

49. Goswami, S., Samanta, D. & Duraivelan, K. Molecular mimicry of host short linear motif-mediated interactions utilised by viruses for entry. *Mol Biol Rep* **50**, 4665–4673 (2023).

50. Davey, N. E., Travé, G. & Gibson, T. J. How viruses hijack cell regulation. *Trends Biochem Sci* **36**, 159–169 (2011).

51. Via, A., Uyar, B., Brun, C. & Zanzoni, A. How pathogens use linear motifs to perturb host cell networks. *Trends Biochem Sci* **40**, 36–48 (2015).

52. Dana, J. M. *et al.* SIFTS: updated Structure Integration with Function, Taxonomy and Sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. *Nucleic Acids Res* **47**, D482–D489 (2019).

53. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403–410 (1990).

54. Sievers, F. & Higgins, D. G. Clustal Omega for making accurate alignments of many protein sequences. *Protein Sci* **27**, 135–145 (2018).

55. Elnaggar, A. *et al.* ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Trans Pattern Anal Mach Intell* **44**, 7112–7127 (2022).

56. Quaglia, F. *et al.* DisProt in 2022: improved quality and accessibility of protein intrinsic disorder annotation. *Nucleic Acids Res* **50**, D480–D487 (2022).

57. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

58. Dobson, L. *et al.* UniTmp: unified resources for transmembrane proteins. *Nucleic Acids Res* **52**, D572–D578 (2024).