Reliable prediction of short linear motifs in the human proteome

Rita Pancsa[1,†,*], Erzsébet Fichó[2,†], Zsofia E. Kalman[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: Rita Pancsa: pancsa.rita@ttk.hu; Laszlo Dobson: dobson.laszlo@ttk.hu

Abstract

Short linear motifs (SLiMs) are small interaction modules within intrinsically disordered regions of proteins that interact with specific domains, and thereby regulate numerous biological processes. Their limited sequence information leads to frequent false positive hits in computational and experimental SLiM identification methods. We present SLiMMine, a deep learning-based method to identify SLiMs in the human proteome. By refining the annotations of known motif classes, we created a high-quality training dataset. Using protein embeddings and neural networks, SLiMMine reliably predicts novel SLiM candidates in known classes, eliminates ~80% of the pattern matching-based hits as false-positives, furthermore, it also functions as a discovery tool to find uncharacterized SLiMs based on optimal sequence environment. Finally, narrowing the broad interactor-domain definitions of known SLiM classes to specific human proteins enables more precise linking of predicted SLiMs to known protein-protein interactions. SLiMMine is available as a 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 domains[1]. Their structural flexibility and limited number of specificity-determining residues allow them to *de novo* emerge in 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 intrinsically disordered regions (IDRs), which provide 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]. They are implicated in various diseases, ranging from disease-associated mutations that abolish the motif function[5] to pathogenic effector proteins that mimic SLiMs to hijack host-cell regulation[6].

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[7]. Because multiple steps are needed to confirm a motif's functionality, the number of classified examples remains limited to only a few thousand[8].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 analyses[8], but other SLiM sources (such as LMPID[9] or UniProt annotations[10]) do not cover all 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 at the proteome scale. Initially developed for PDZ domains[11], it has now been extended to other SLiM-binding bait domains, enabling high-throughput discovery of domain–SLiM interactions across the human proteome[12]. Complementing these experimental strategies, quantitative fragmentomics studies such as those by Gogl *et al.*[13] have systematically measured tens of thousands of PDZ-binding motif (PBM)-PDZ domain affinities within the human interactome, providing a detailed landscape of PDZ binding specificity and promiscuity. In parallel, computational approaches can identify protein binding segments in disordered regions based on biophysical properties[14] or protein embeddings[15], find conserved segments[16], apply contextual filters[17], detect motifs based on AlphaMissense pathogenicity scores[18] or prioritize those motifs of a protein, whose partner domains are present in the experimentally verified interaction partners of the protein[19].

Here, we propose an accurate deep learning-based method for identifying SLiMs in the human proteome that largely eliminates the problem of abundant false-positive hits. We relied heavily on the manually curated ELM resource and manually refined and extended its annotations to construct a high-quality training dataset that represents a high added value. Using protein embeddings and various neural network architectures, we developed a predictive model that reliably identifies novel SLiM candidates. The accuracy of the method is demonstrated on diverse independent test sets and case studies. Furthermore, based on the new, more detailed annotations and available protein-protein interaction (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 novel PPIs.

Results

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

To train a SLiM predictor and to identify the most likely binding partners of the predicted SLiMs among known PPIs, we needed reliable SLiM data annotated along specific criteria. As a first step, we mapped mammalian motif instances to their human orthologs and adjusted the motif boundaries of human instances to comply with current UniProt[10] sequences.

To obtain a training dataset that meets our pre-defined criteria, we have refined and extended the annotations for the ~320 ELM motif classes relevant to human proteins. Relying on ELM class text descriptions, the biological context of the motif classes and 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 their expected Gene Ontology[20] (GO) biological processes and molecular functions are. In addition, whenever possible, we precisely defined the possible binding partners, i.e. the subset of human proteins carrying the specific binding domain that actually binds to the given SLiM class. This was required because ELM domain definitions are often not specific enough. For instance, the docking motif specific for JNK kinases (ELM class: DOC_MAPK_JIP1_4) was assigned to the generic kinase domain, which can be found in hundreds of human kinases. In our refined dataset, only the three human JNK kinases are listed as possible partners for this motif class. In other cases, broad definitions of the interacting domains could be changed to more specific definitions from InterPro[21]. 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. 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 for 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 if a given protein or region is unlikely to carry a true functional motif belonging to a given class. We scanned the human proteome for regular expressions (regex) 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[22] structure. IV) The motif hit was not predicted to be disordered using AIUPred[23]. V) The motif hit, although tested, did not bind to the annotated partner domain in ProP-PD[12] 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 regex is). Using the flagged (likely false-positive)

motif hits, five times more negative data were selected for each class than the associated positive data (Figure 1/B).

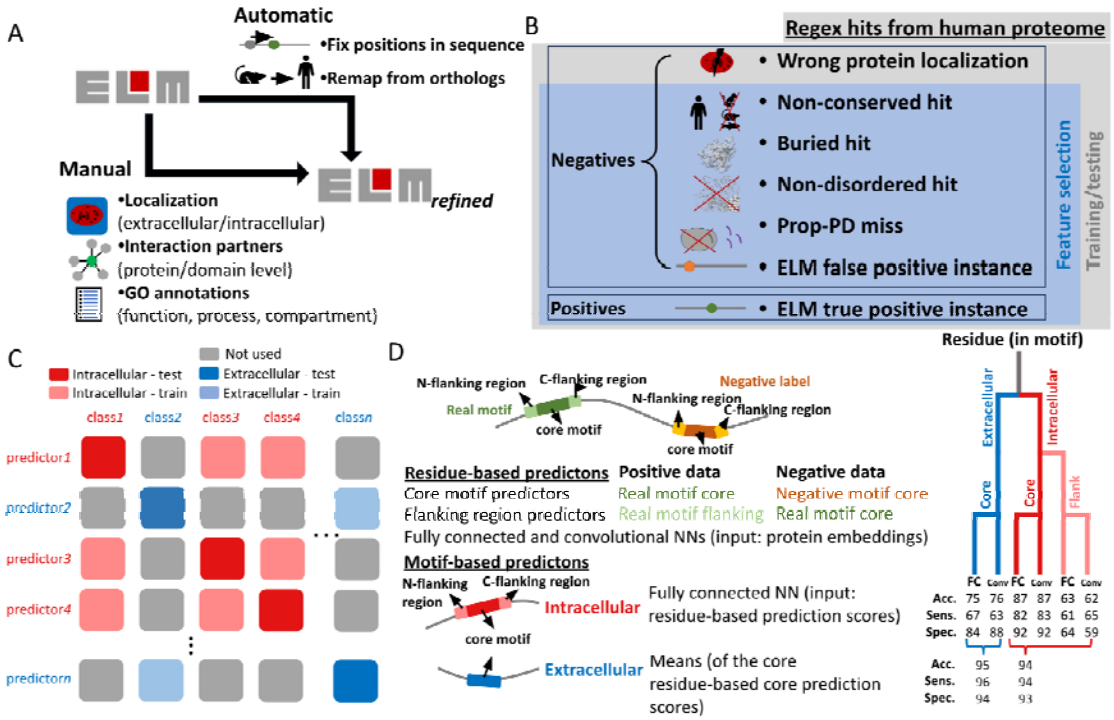

Figure 1. The training data, architecture and accuracy of SLiMMine. *A) Automatic remapping of motif instances and manual refinement of ELM class annotations. B) Definition of labels used to train (and test) SLiMMine. C) Definition of predictors, training set and testing sets considering the motif class 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 (Acc: accuracy, Sens: sensitivity, Spec: specificity).*

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 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 exhibit 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 conducted additional tests to objectively evaluate the performance of SLiMMine. Using the benchmark set, we assessed the performance of a classical approach that incorporates protein disorder (from AIUPred) and conservation (for more details, see methods). We also benchmarked AIUpred-binding[14], a successor of Anchor[24]. Although it was not specifically developed for motif analysis, it is widely used, available as both a web server and a standalone program, and its authors have demonstrated its ability to identify functional motifs. SLiMMine outperforms both traditional conservation and protein disorder-based analyses, as well as 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: phosphomodulated motifs[25], cyclin-docking motifs[26], 14-3-3 binding motifs[27], essential motifs required for normal cell proliferation[28] and LMPID[9] motif entries (Figure 2/B), and it confidently identified the vast majority of the associated motifs (73.6% of the 1177 tested motifs received a score of 0.5 or higher, and 40% of them 0.9 or higher). Notably, SLiMMine scores are not exceptionally high for Prop-PD hits[12], but when filtered to retain only conserved motifs, the resulting distribution becomes comparable to previously defined positive sets (SMaterial, SFigure 1). In contrast, segments obtained by pattern matching of ELM regular expressions on monomeric structures, coiled coils (decoy sets) and disordered linker and terminal regions (neutral set) received low scores (Figure 2/B). The distributions of scores for monomeric structures and coiled coils show that SLiMMine does not falsely identify motifs within these folded regions, even though coiled-coils are disordered in their monomeric form and are therefore often cross-predicted with disordered regions[29]. The distribution of SLiMMine scores for disordered linker and terminal regions is uniform, consistent with regex hits being either true motifs or falsely identified ones in these regions.

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, the reference distribution on the benchmark set is visible in SFigure 2). 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 a score of 0.9 or higher, and 696,329 SLiM instances with a score of 0.5 or higher, among known motif classes. This methodology could filter out 2,718,778 regex hits as likely false-positives (score < 0.5).

The power of SLiMMine prediction is demonstrated on Fibronectin (*FN1*; UniProt: P02751), which contains multiple integrin-binding motifs. These are often located in loops 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. Still, SLiMMine successfully identifies all experimental

motifs, whereas decoy motif regex hits (meaning regex hits for intracellular motifs in extracellular or buried regions) were assigned with lower prediction scores (Figure 3/A).

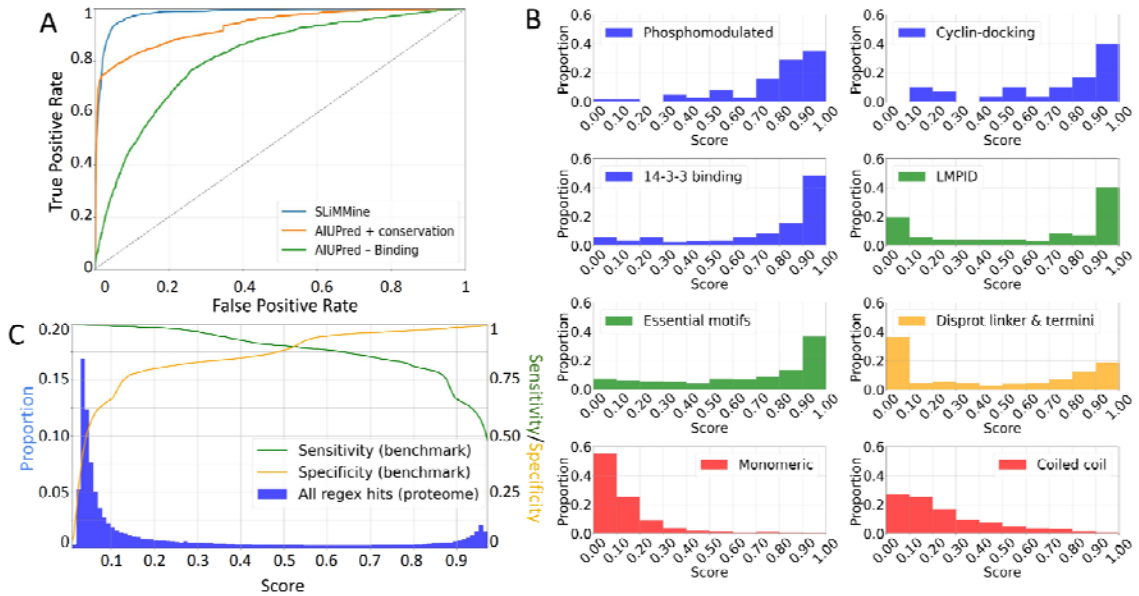

Figure 2. SLiMMine accurately predicts SLiM-mediated interactions across diverse datasets and in the human proteome. *A) Receiver Operating Characteristics of SLiMMine, AIUPred-binding and a traditional prediction using conservation and protein disorder (AIUPred). B) Prediction score distribution on specific motif sets (blue: phosphomodulated motifs (n=83), cyclin docking motifs (n=30) and 14-3-3 binding motifs (n=263)), general motif sets (green: LMPID motif entries (n=276) and essential motifs (n=525)), a neutral set of disordered regions (yellow: ELM class regex hits within disordered linkers and termini (n=2032)) and decoy sets (red: ELM class regex hits within monomeric structures (n=1374) and coiled-coils (n=1092)). These sets are independent from training data, since motifs overlapping with ELM instances have been removed. C) Distribution of prediction scores for all motif regex matches in the human proteome (bars). Sensitivity and specificity values calculated on the benchmark set at varying score thresholds are indicated by lines.*

To also showcase an intracellular example, in the Smaug1 RNA-binding protein (*SAMD4A*; UniProt: Q9UPU9), SLiMMine correctly identifies three of the four experimentally validated 14-3-3 binding motifs[30], the nuclear export signal[31] (NES) and the C-terminal PDZ-binding motif (Dean group, unpublished results) with very high prediction scores (>0.9), although none of these motifs were present in the training set (SFigure 3).

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

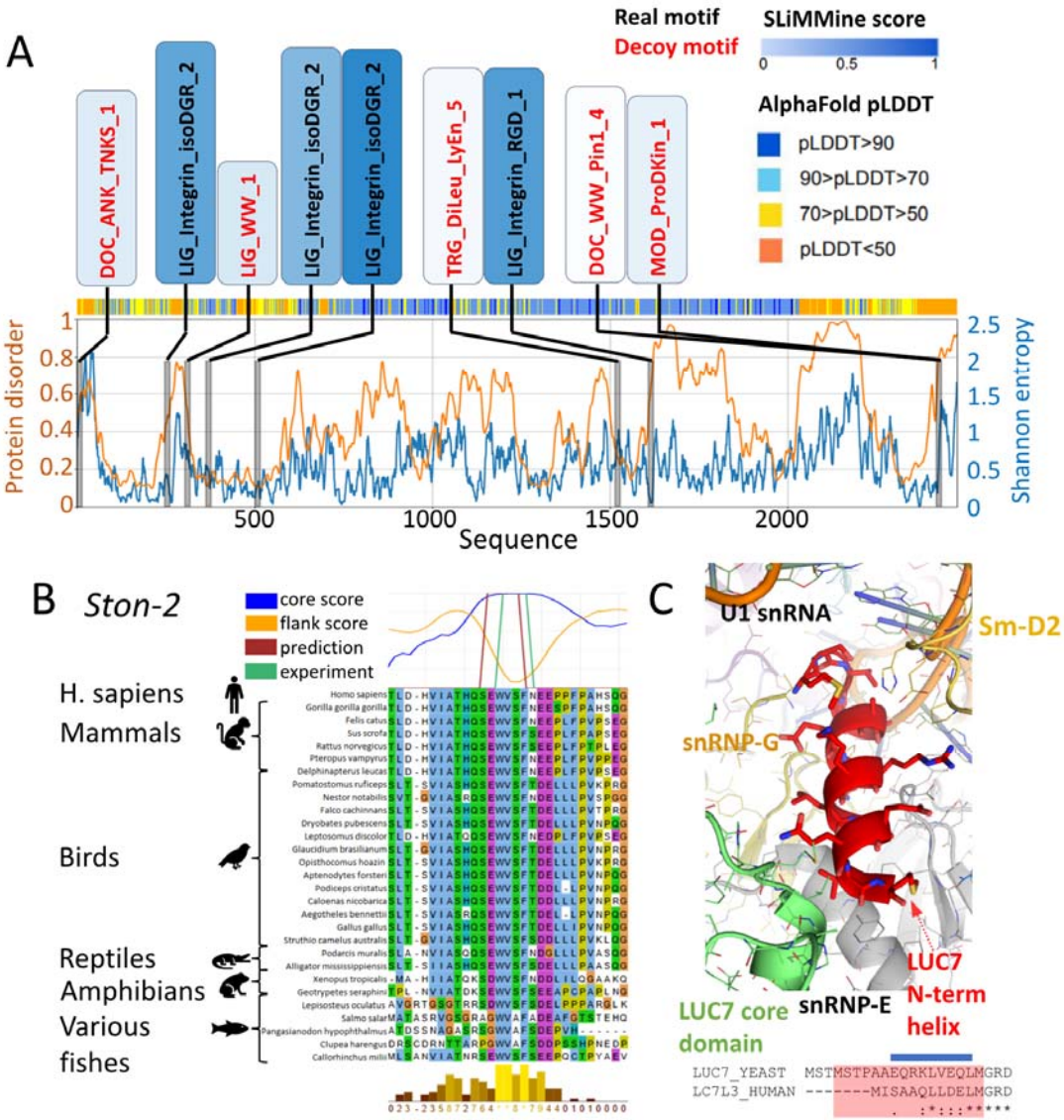

Figure 3: SLiMMine successfully identifies novel instances of known motif classes as well as yet undefined motifs. *A) SLiMMine prediction results of known motif classes on Fibronectin using real and decoy motifs. AlphaFold pLDDT values, AIUPred output and residue conservation are also displayed. B) De novo prediction, alignment and protein conservation for a SLiM in the Stonin-2 protein, which binds to the alpha-ear domain of AP-2 complex subunit alpha-1. C) Structural snapshot of the N-terminus of yeast Protein LUC7 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 Luc7-like protein 3 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.*

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

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 this 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 in SLiMMine identifies longer regions as favourable for motifs, flanking region prediction opens prospects for defining novel motifs without specifying their 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 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 regex). Applying SLiMMine with these thresholds on the full human proteome across all possible lengths, identified 32,501 short protein regions. These 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 (SFigure 4). For example, in Stonin-2 (*STON2*; UniProt: Q8WXE9), SLiMMine correctly identified the experimentally verified alpha-ear domain-binding motif[32] stored in LMPID, even though it 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. As an example, the N-terminus of Dynein 2 intermediate chain 2 (*DYNC2I2*; UniProt: 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[33]. 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*; UniProt: Q6UUV9), evidenced by the structure pdb:7D8H[34], despite its complete lack of arginine/lysine residues (QYYGGSLP).

The flexible N-terminal region of the human Luc7-like protein 3 (*LUC7L3*; UniProt: O95232) was predicted to contain a novel motif that is conserved in its yeast homolog Protein LUC7. In the crystal structure of the *S. cerevisiae* spliceosomal E complex (pdb:6N7R[35]), 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).

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 was trained on.

Case studies

The accuracy of SLiMMine is also demonstrated on some full-length proteins that have multiple validated SLiM-mediated interactions covering a wide variety of motif classes. In the multi-domain podosome scaffold protein, Tks4 (*SH3PXD2B*; UniProt: A1X283), SLiMMine

correctly identified the experimentally validated tandem SH3-binding motif[36], actin capping protein (CP)-interacting motif[37] and CD2-associated protein (CD2AP)-binding motif[38] as *de novo* motifs, although none of them are present in ELM. While ELM stores 9 class variants for SH3-binding motifs, the tandem SH3-binding motif is not among them[39]. The CD2AP-binding motif of Tks4 does not completely fit the strict regex of the corresponding ELM class (class: LIG_SH3_CIN85_PxpxPR_1, regex: Px[PA]xPR) because it has a valine in the middle. However, results of peptide array screens support that the three SH3 domains of CD2AP have slightly different recognition specificities, and some of them also accept valine at this position[40]. The identified CP-interacting motif also does not fit the ELM class definition (LIG_ActinCP_CPI_1) due to a slight difference, however, subunits of the actin CP heterodimer were consistently identified as Tks4 binding partners by immunoprecipitation assays in five cancer cell lines[37], and therefore the proposed motif is most likely functional. On top of the interesting *de novo* hits, SLiMMine also pinpoints a Nck1/2 SH2-binding motif at phosphosite Y508, enabling the interaction between Tks4 and Nck1/2[41]. The SH2 domain of Src recognizes the same phosphorylated Tyr (Y508), while its SH3 domain binds to a Pro-rich peptide within Tks4[42] that SLiMMine also identified as a high-scoring hit. Although our institute has a long history in studying Tks4, we do not know of any other approved motifs in it[36–39,42]. Therefore, although not encountered during training, SLiMMine confidently identified all five known motifs of Tks4 (Figure 4/A). Of the ten *de novo* motif hits within the 911 residues long Tks4 (which cover <10% of the sequence) five match already validated motifs, while the remaining five could be yet uncovered motifs. The interaction partner of Tks4, CD2AP is also presented as a case study (Supplementary Material, SFigure 5).

The M-phase inducer phosphatase 1 Cdc25A (*CDC25A*; UniProt: P30304) is often used to illustrate how the versatility of SLiMs can regulate the life cycle and functioning of a protein[1,4]. It has 11 well-documented motifs, of which four are annotated ELM instances[8] (the cyclin-docking motif, a CK1 and a CDK phosphorylation site and the PP4 phosphatase docking motif). All four are successfully identified by SLiMMine (although in the case of the cyclin-docking motif and CK1 phosphorylation site, the motif boundaries are slightly shifted compared to the ELM instance (see SMaterial and SFigure 6 for explanation)). Three other motifs that fit the respective ELM class regex definitions are also all identified (the KEN-box degron[43], a canonical 14-3-3-binding motif[44] and the nuclear localisation signal[45]). The two copies of the SCF-beta-TrCP E3 ligase-specific degron[46] do not exactly fit the respective ELM regex, still, they are both found as *de novo* hits. Only two Cdc25A motifs were missed: the NES[45] that does not fit ELM NES class definitions and is too long to be identified as a *de novo* hit, and a second, non-canonical 14-3-3 binding site centered around T507[44] that is in a non-disordered region (AIUPred score < 0.4). In all, SLiMMine could identify most known motifs of Cdc25A as high-scoring hits (Figure 4/B).

Additional case studies, including E3 ubiquitin-protein ligase RNF13, Smaug1, CD2AP, LanC-like protein 2, and Early E1A from Human adenovirus 5 further illustrate the applicability of SLiMMine across diverse biological contexts (Supplementary Material, SFigure 3, SFigure 5, SFigure 7, SFigure 8*)*.

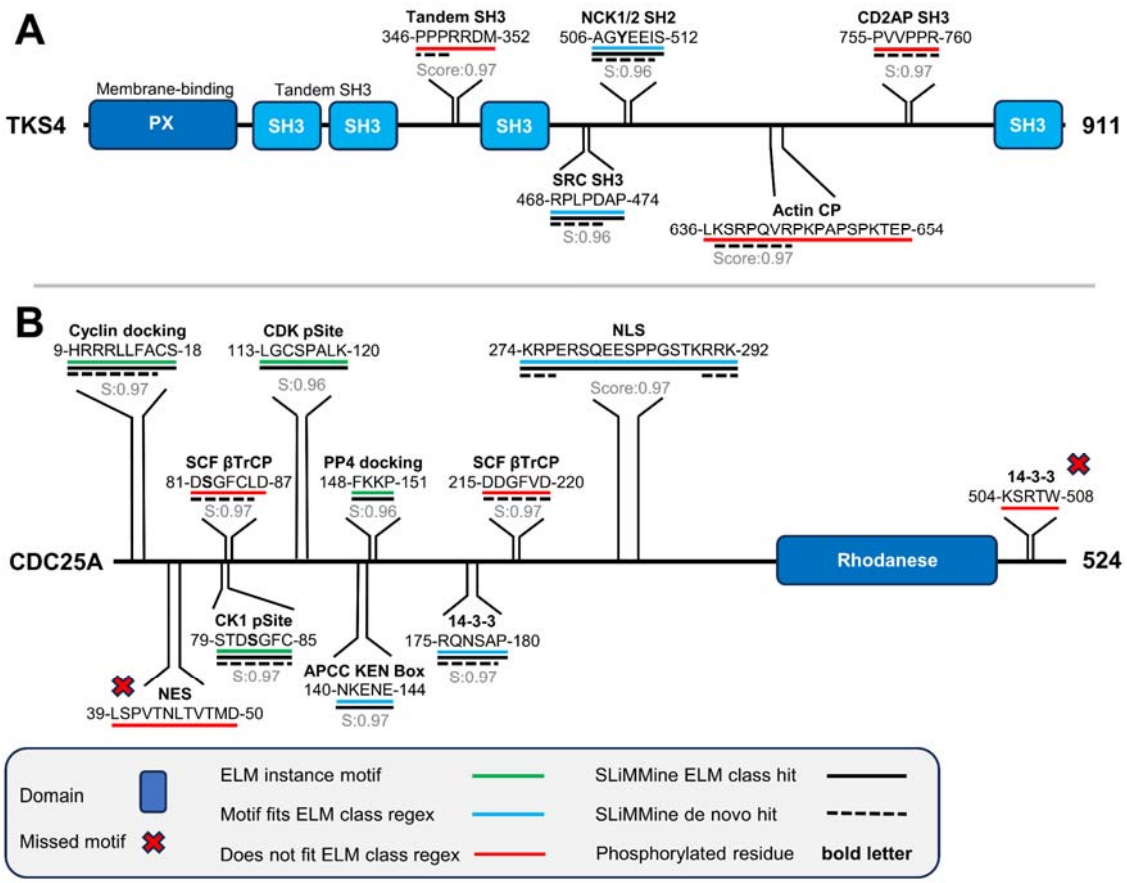

Figure 4: SLiMMine successfully identifies most known motifs of the human Tks4, CD2AP and Cdc25A proteins. *A) The domains (blue boxes) and SLiMs (sequence bits) are indicated on the domain maps of Tks4 and B) Cdc25A. For each known motif, the respective binding domains/motif names, sequences, residue boundaries, ELM status (underscoring in different colors) and SLiMMine predictions (underscoring with black simple (ELM class predictions) and/or dashed (de novo predictions) line) are indicated along with the respective SLiMMine Score (S).*

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[47]), IntAct[48] and BIOGRID[49]) 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, suggested underlying interaction mechanisms and interacting regions for more than 50 000 human PPIs (SMaterial, SFigure 9), for which no such data were previously available.

SLiMMine comes with a user-friendly web server facilitating motif discovery

SLiMMine is shared with the research community through a modern, easy-to-use web resource. To cover all possible use cases, the server provides three different options. Users can search for I) any protein in the human reference proteome, II) predefined ELM classes or III) custom regular expressions. The first option leads to the respective protein entry page with all predicted motifs displayed, while the latter two options provide a list of proteins that contain sequence matches to the selected ELM class/custom regex, along with the associated SLiMMine prediction scores.

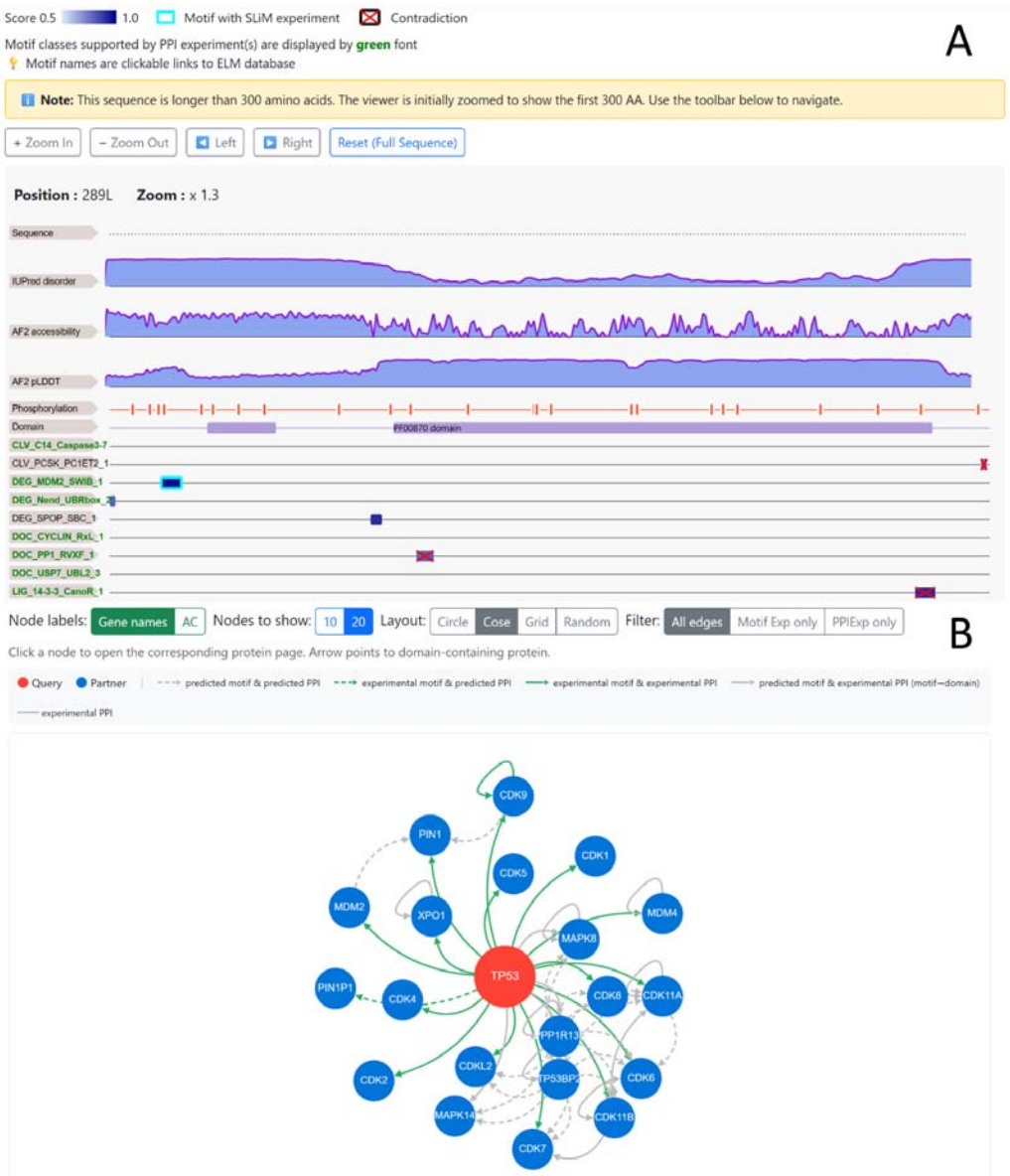

Figure 5: The SLiMMine web server represents an ideal tool for motif discovery. A) *Layout of the sequence viewer and B) the Protein interaction viewer.*

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 5/A) was designed to summarize all relevant information about the predicted motifs along the sequence and to display evidence-based filters that help the user decide whether the motifs could be functional. At the top of the panel, measures related to protein disorder are shown. Each motif class is shown below in a separate line. The different colors and frames of the boxes representing the motifs correspond to different evidence supporting the actual motif. For example, motifs with contradicting evidence (e.g., a phosphodependent motif without experimental phosphorylation evidence[50]) are crossed. Notably, when we defined these contextual filters, we found that they are unlikely to rule out functional motifs (SFigure 10) and they often help to pinpoint functional motifs among multiple hits (SFigure 11). Below the viewer, there is a table detailing supporting (or contradicting) evidence on the motifs.

The protein interaction network (Figure 5/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 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.

Discussion

SLiMMine is an alignment-free approach for predicting SLiMs. It is optimized for scenarios in which motif boundaries are known, for example, after scanning sequences with regular expressions derived from motif databases. The model evaluates whether the given region exhibits properties consistent with functional SLiMs. Although the method was trained and optimized on human protein data, we expect it to perform similarly well on other mammalian proteins.

A major obstacle during the development was the limited availability of experimentally validated motif data. Positive samples could be retrieved from ELM[8]. Defining true negatives was more challenging, as the absence of a motif annotation does not imply the absence of function (see SMaterial). However, emerging high-throughput experimental approaches are expected to eliminate this limitation in the long term[12,51,52]. 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. The current version of SLiMMine has not been tested for SLiMs in various pathogenic species, yet initial benchmarking shows promising results (see SMaterial, SFigure 8 and SFigure 12). Notably, SLiMMine was trained and tested on conserved motifs, so the prediction of fast-evolving viral mimicry motifs[53] may need further considerations[8].

SLiMMine classifies more than 80% of the proteome-scale sequence matches of ELM-derived regular expressions as false positives. 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 that some motif classes have permissive regular expressions with higher pattern probabilities and therefore occur tens of thousands of times within the human proteome. 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 are explicitly annotated as such. The SLiMMine web resource highlights these contextual annotations. 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 the provided supporting or contradicting evidence.

Beyond identifying candidate SLiMs, SLiMMine provides additional biological insight by proposing likely interaction partners for predicted motifs, and thereby enables functional hypothesis generation. The combined information can serve as a powerful input for both computational and experimental studies, and can substantially reduce the search space in large-scale interaction modeling efforts, such as AlphaFold-based domain–motif screening approaches[54], or guide the design of targeted validation assays, such as peptide pull-downs, mutational analyses, or cell-based interaction assays.

An important extension of SLiMMine is the *de novo* motif prediction module, which seeks to identify SLiM-like regions without predefined boundaries. This idea is motivated by the observation that many SLiMs exhibit “island-like” conservation patterns[55]: a short core region shows strong functional constraint, while flanking residues are more variable but still contribute to binding specificity or regulation. By evaluating the relationship between the predicted motif core and its surrounding residues, SLiMMine can highlight regions without defining the start and end position of a segment. Notably, if multiple motifs overlap, the interpretation of flanking regions becomes ambiguous. In all, the *de novo* approach can be valuable for identifying novel motif classes and atypical binding modes that are currently underrepresented in curated databases.

SLiMMine can serve as a starting point for addressing biologically relevant questions. Many regulatory proteins contain multiple SLiMs that engage several binding partners, leading to cooperative or competitive interaction networks[56]. Multivalency can increase effective binding affinity and enable dynamic assembly of signaling complexes[1]. In addition, clustering of multiple SLiMs may contribute to liquid–liquid phase separation, in which networks of weak, multivalent interactions drive the formation of dynamic biomolecular condensates that spatially organise signaling and regulatory processes[57,58]. 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, many disrupt SLiMs[18,59]. Loss (or gain) of a single motif can rewire interaction networks, leading to aberrant signaling, mislocalization, altered protein turnover or stability[60]. By highlighting predicted motifs and their potential interaction partners, SLiMMine can aid interpreting missense variants that fall outside classical functional domains.

In all, with its accurate predictions, SLiMMine represents a large step forward in identifying and understanding short linear motifs in human proteins. We plan to update SLiMMine when new data become available (i.e. when new motif class definitions can be extracted from novel, annotated SLiM databases or derived from large-scale experimental SLiM discovery datasets) and extend it to proteins of other organisms. In addition to predicting novel instances of already defined SLiM classes, SLiMMine could also identify ~32,500 *de novo* SLiMs in the human proteome. Among these, we identified several experimentally validated motifs that were not part of the training data and may not fit existing

motif definitions, providing strong support for the applicability of our method. Furthermore, reliable motif predictions from SLiMMine and the refined partner annotations 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 is a powerful tool with the 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 for the 319 motif classes in the ELM resource that are relevant to 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 their expected GO[18] biological processes and molecular functions are. Although ELM has its own GO annotations, those are mixed GO terms, meaning either relevant to the motif- or domain-containing counterparts of the interactions (or both), so they would have been problematic to use. We also identified the potential binding partners for the given ELM class - that is, the subset of proteins containing the relevant binding domain that has been shown to interact with that particular SLiM type. For this, either the same (usually PFAM) domain identifier was used as in ELM (if further specification of the partners was not required), or a more specific domain identifier from InterPro[21], or the 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[8]. 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[10] (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[23] > 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 motif boundaries from the 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[61]), 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[12].

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[12] 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 vary 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[62] (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[63]. We searched for the regular expression in the orthologous proteins. Since aligning disordered regions is challenging, we allowed 50 residues deviations 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. To visualize entropy for the Fibronectin example, we simply calculated the Shannon entropy at each position in the alignment.

Train and test sets

For each motif class, we defined a broad-level localization definition (extracellular/intracellular, 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 (i.e. start and end position in a sequence, that is usually defined by a regular expression) and predict whether that segment exhibits SLiM-like characteristics, that is reflected by a score between 0 and 1. 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 extracellular flanking regions the NN did not converge), 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[64] 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 for which 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, 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 to predict core and flanking regions. 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 outputs of the residue-level fully connected and convolutional NNs. The architectures of these predictors are visible on SFigure 13. 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/disorder-based motif prediction. We used AIUPred[23] results (calculating the mean over the residues within the motif) and conservation score from Chordata species (see “Defining motif labels”), then the mean of these two values defined the score of the traditional approach. We also compared SLiMMine with AIUPred-binding[14] (STable 18, Figure 1/A).

We also used additional motif data to test the robustness of our method. Cyclin-docking peptides were collected from Örd M. *et al.*[26], where ~100,000 peptides were assayed for relative binding strength to cyclins from five human cyclin families (D, E, A, B, and F) using quantitative intracellular binding and large-scale tiled peptide screening. We relied on Supplementary Data 2 of the cited paper, from which only high-confidence hits were retained; when the authors identified a motif match, we refined the peptide boundaries to the detected motif. We collected phospho-modulated motif interactions[25] identified via a phosphomimetic proteomic peptide–phage display (ProP-PD) assay using a library covering ~13,500 phospho-serine/threonine sites in intrinsically disordered regions. We relied on Table S5 of the cited publication and accepted only high-confidence non-redundant peptide hits corresponding to the wild-type form. We used a compilation of 14-3-3-binding motifs

from Egbert CM *et al.*[27]; however, we considered only a four-residue environment around the phosphorylated positions (listed in Table S2 of the cited paper). We also assessed the performance of SLiMMine on a list of essential human motifs required for normal cell proliferation based on large-scale mutation screening[28] (obtained from the Source data of Figure 2 of the cited paper). We further tested SLiMMine on human LMPID[9] motif entries supported by experimental evidence (STable 19).

We defined a decoy set by searching motif candidates with regular expressions in the ordered monomer test set also used for benchmarking Anchor[14] and in coiled-coil regions predicted by four different methods (from Supplementary Table 3 of Kalman *et al.*[65]). In addition, we analyzed disordered linker and terminal regions from DisProt[66] as a neutral set (covered by the functional annotation categories: “flexible C-terminal tail”, ”flexible N-terminal tail”, ”entropic chains”, and “flexible linker”) by pattern matching of ELM class regular expressions for motif candidates (STable 20). In the above described 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 STables 19-20.

De novo motif prediction

Using the constructed benchmark set, we evaluated multiple score cutoffs to define the most reliable *de novo* motif prediction. Using intracellular motifs of length 3-9 residues from the training and test sets, we benchmarked the correspondence between the final, the residue-level core and the residue-level flanking scores. 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).

Prediction of SLiMs in the human proteome

We scanned the human proteome with the regular expressions of the 319 ELM motif classes that are relevant to human proteins. Based on motif localization (intracellular or extracellular), we applied the appropriate predictors to each match, using the start and end positions defined by the regular expression. Motif scores were calculated as the mean probability across all final scores generated by the selected predictors.

For *de novo* prediction, we screened the human proteome for all segments of length 9, 7, 5, and 3. When a segment was identified as a motif, overlapping peptides were excluded from further analysis.

Protein-protein interaction data

For all predicted motif instances, we considered the list of annotated, possible partners of the respective SLiM class as potential interactors (STable1). If an interaction was validated in BIOGRID[49], IntAct[48] (with physical association or stronger annotation (MI:0915)), or it was included in the PDB (accessed via 3did[47]), ELM or LMPID, we accepted it as an experimentally verified protein-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[67], AIUPred disorder probabilities[23], and UniTmp[68] 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 fast 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: Further independent test sets (cyclin-binding motifs, phosphodependent motifs, 14-3-3 binding motifs, essential motifs required for normal cell proliferation and LMPID motif entries)
STable 20: Further independent test sets (monomeric structures, coiled-coils and DisProt linkers/termini)
STable 21: Sensitivity, specificity and motif boundary statistics for *de novo* predicted motifs at different thresholds
STable 22: SLiMMine *de novo* motif prediction results on the benchmark set with selected thresholds

SFigure 1: SLiMMine performance on Prop-PD dataset.
SFigure 2: The performance parameters of SLiMMine on the benchmark set.
SFigure 3: SLiMMine finds almost all experimentally validated motifs of Smaug1.
SFigure 4: SLiMMine *de novo* motif prediction finds multiple experimentally verified motifs.
SFigure 5: SLiMMine successfully identifies most known motifs of the human CD2AP.
SFigure 6: Overlapping and shifted motif detection may affect the visualization of ELM instances in SLiMMine.
SFigure 7: SLiMMine finds the nuclear localization signal in LanC-like protein 2 as a *de novo* motif.
SFigure 8: SLiMMine successfully identifies the motifs employed by the viral Early E1A protein.
SFigure 9: SLiMMine predictions explain a non-negligible fraction of validated human PPIs.
SFigure 10: The effect of contextual contradictions on experimentally verified motifs.
SFigure 11: Context filters work in the Piezo2 receptor.
SFigure 12: SLiMMine performance on motifs of pathogens.
SFigure 13: The Neural Network architectures of SLiMMine.

SMaterial: Extended discussion

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. Uyar, B., Weatheritt, R. J., Dinkel, H., Davey, N. E. & Gibson, T. J. Proteome-wide analysis of human disease mutations in short linear motifs: neglected players in cancer? *Mol Biosyst* **10**, 2626–2642 (2014).
6. Davey, N. E., Travé, G. & Gibson, T. J. How viruses hijack cell regulation. *Trends Biochem Sci* **36**, 159–169 (2011).
7. Davey, N. E. *et al.* Attributes of short linear motifs. *Mol. Biosyst.* **8**, 268–281 (2012).
8. Kumar, M. *et al.* ELM-the Eukaryotic Linear Motif resource-2024 update. *Nucleic Acids Res* **52**, D442–D455 (2024).
9. Sarkar, D., Jana, T. & Saha, S. LMPID: a manually curated database of linear motifs mediating protein-protein interactions. *Database (Oxford)* **2015**, (2015).
10. UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Res* **53**, D609–D617 (2025).
11. 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).

12. Benz, C. *et al.* Proteome-scale mapping of binding sites in the unstructured regions of the human proteome. *Mol Syst Biol* **18**, e10584 (2022).
13. Gogl, G. *et al.* Quantitative fragmentomics allow affinity mapping of interactomes. *Nat Commun* **13**, 5472 (2022).
14. Erdős, G., Deutsch, N. & Dosztányi, Z. AIUPred - Binding: Energy Embedding to Identify Disordered Binding Regions. *J Mol Biol* **437**, 169071 (2025).
15. Jahn, L. R., Marquet, C., Heinzinger, M. & Rost, B. Protein embeddings predict binding residues in disordered regions. *Sci Rep* **14**, 13566 (2024).
16. Chow, C. F. W. *et al.* SHARK-capture identifies functional motifs in intrinsically disordered protein regions. *Protein Sci* **34**, e70091 (2025).
17. 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).
18. Deutsch, N., Erdős, G. & Dosztányi, Z. Pathogenic variations illuminate functional constraints in intrinsically disordered proteins. *iScience* **29**, 115215 (2026).
19. Reys, V., Pons, J.-L. & Labesse, G. SLiMAn 2.0: meaningful navigation through peptide-protein interaction networks. *Nucleic Acids Res* **52**, W313–W317 (2024).
20. Gene Ontology Consortium *et al.* The Gene Ontology knowledgebase in 2023. *Genetics* **224**, (2023).
21. Blum, M. *et al.* InterPro: the protein sequence classification resource in 2025. *Nucleic Acids Res* **53**, D444–D456 (2025).
22. 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).
23. 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).
24. Dosztányi, Z., Mészáros, B. & Simon, I. ANCHOR: web server for predicting protein

binding regions in disordered proteins. *Bioinformatics* **25**, 2745–2746 (2009).

25. Kliche, J. *et al.* Large-scale phosphomimetic screening identifies phospho-modulated motif-based protein interactions. *Mol Syst Biol* **19**, e11164 (2023).
26. Örd, M. *et al.* High-throughput investigation of cyclin docking interactions reveals the complexity of motif binding determinants. *Nat Commun* **16**, 7622 (2025).
27. Egbert, C. M. *et al.* The Integration of Proteome-Wide PTM Data with Protein Structural and Sequence Features Identifies Phosphorylations that Mediate 14-3-3 Interactions. *J Mol Biol* **435**, 167890 (2023).
28. Ambjørn, S. M. *et al.* A proteome-wide dependency map of protein interaction motifs. *Nat Struct Mol Biol* **33**, 537–546 (2026).
29. Szappanos, B., Süveges, D., Nyitray, L., Perczel, A. & Gáspári, Z. Folded-unfolded cross-predictions and protein evolution: the case study of coiled-coils. *FEBS Lett* **584**, 1623–1627 (2010).
30. 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).
31. Kırlı, K. *et al.* A deep proteomics perspective on CRM1-mediated nuclear export and nucleocytoplasmic partitioning. *Elife* **4**, (2015).
32. 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).
33. 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).
34. 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).
35. Li, X. *et al.* A unified mechanism for intron and exon definition and back-splicing. *Nature* **573**, 375–380 (2019).
36. Merő, B. *et al.* Characterization of the Intramolecular Interactions and Regulatory Mechanisms of the Scaffold Protein Tks4. *Int J Mol Sci* **22**, (2021).

37. László, L. *et al.* Unveiling epithelial plasticity regulation in lung cancer: Exploring the cross-talk among Tks4 scaffold protein partners. *Mol Biol Cell* **35**, ar111 (2024).
38. 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).
39. Kalman, Z. E., Lazar, T., Dobson, L. & Pancsa, R. Definition and Discovery of Tandem SH3-Binding Motifs Interacting with Members of the p47-Related Protein Family. *Biomolecules* **15**, (2025).
40. 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).
41. Huttlin, E. L. *et al.* Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. *Cell* **184**, 3022–3040.e28 (2021).
42. Dülk, M. *et al.* EGF Regulates the Interaction of Tks4 with Src through Its SH2 and SH3 Domains. *Biochemistry* **57**, 4186–4196 (2018).
43. Kieffer, I., Lorenzo, C., Dozier, C., Schmitt, E. & Ducommun, B. Differential mitotic degradation of the CDC25B phosphatase variants. *Oncogene* **26**, 7847–7858 (2007).
44. Chen, M.-S., Ryan, C. E. & Piwnica-Worms, H. Chk1 kinase negatively regulates mitotic function of Cdc25A phosphatase through 14-3-3 binding. *Mol Cell Biol* **23**, 7488–7497 (2003).
45. Källström, H., Lindqvist, A., Pospisil, V., Lundgren, A. & Rosenthal, C. K. Cdc25A localisation and shuttling: characterisation of sequences mediating nuclear export and import. *Exp Cell Res* **303**, 89–100 (2005).
46. Jin, J. *et al.* SCFbeta-TRCP links Chk1 signaling to degradation of the Cdc25A protein phosphatase. *Genes Dev* **17**, 3062–3074 (2003).
47. 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).

48. Del Toro, N. *et al.* The IntAct database: efficient access to fine-grained molecular interaction data. *Nucleic Acids Res* **50**, D648–D653 (2022).
49. Oughtred, R. *et al.* The BioGRID interaction database: 2019 update. *Nucleic Acids Res* **47**, D529–D541 (2019).
50. Ochoa, D. *et al.* The functional landscape of the human phosphoproteome. *Nat Biotechnol* **38**, 365–373 (2020).
51. Mihalič, F. *et al.* Large-scale phage-based screening reveals extensive pan-viral mimicry of host short linear motifs. *Nat Commun* **14**, 2409 (2023).
52. Subbanna, M. S., Winters, M. J., Örd, M., Davey, N. E. & Pryciak, P. M. A quantitative intracellular peptide-binding assay reveals recognition determinants and context dependence of short linear motifs. *J Biol Chem* **301**, 108225 (2025).
53. Glavina, J., Palopoli, N. & Chemes, L. B. Evolution of SLiM-mediated hijack functions in intrinsically disordered viral proteins. *Essays Biochem* **66**, 945–958 (2022).
54. Lee, C. Y. *et al.* Systematic discovery of protein interaction interfaces using AlphaFold and experimental validation. *Mol Syst Biol* **20**, 75–97 (2024).
55. 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).
56. 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).
57. Case, L. B., Ditlev, J. A. & Rosen, M. K. Regulation of Transmembrane Signaling by Phase Separation. *Annu Rev Biophys* **48**, 465–494 (2019).
58. 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).
59. Dobson, L., Mészáros, B. & Tusnády, G. E. Structural Principles Governing Disease-Causing Germline Mutations. *J Mol Biol* **430**, 4955–4970 (2018).

60. Kliche, J. *et al.* Proteome-scale characterisation of motif-based interactome rewiring by disease mutations. *Mol Syst Biol* **20**, 1025–1048 (2024).
61. 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).
62. 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).
63. Sievers, F. & Higgins, D. G. Clustal Omega for making accurate alignments of many protein sequences. *Protein Sci* **27**, 135–145 (2018).
64. 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).
65. Kalman, Z. E., Mészáros, B., Gáspári, Z. & Dobson, L. Distribution of disease-causing germline mutations in coiled-coils implies an important role of their N-terminal region. *Sci Rep* **10**, 17333 (2020).
66. Nugnes, M. V. *et al.* DisProt in 2026: enhancing intrinsically disordered proteins accessibility, deposition, and annotation. *Nucleic Acids Res* **54**, D383–D392 (2026).
67. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
68. Dobson, L. *et al.* UniTmp: unified resources for transmembrane proteins. *Nucleic Acids Res* **52**, D572–D578 (2024).