Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Structural Biology

Finding functional motifs in protein sequences with deep learning and natural language models

Castrense Savojardo, Pier Luigi Martelli and Rita Casadio

Abstract

Recently, prediction of structural/functional motifs in protein sequences takes advantage of powerful machine learning based approaches. Protein encoding adopts protein language models overpassing standard procedures. Different combinations of machine learning and encoding schemas are available for predicting different structural/functional motifs. Particularly interesting is the adoption of protein language models to encode proteins in addition to evolution information and physicochemical parameters. A thorough analysis of recent predictors developed for annotating transmembrane regions, sorting signals, lipidation and phosphorylation sites allows to investigate the state-of-the-art focusing on the relevance of protein language models for the different tasks. This highlights that more experimental data are necessary to exploit available powerful machine learning methods.

Addresses

Biocomputing Group, Dept. of Pharmacy and Biotechnology, University of Bologna, Via San Giacomo 9/2, 40126 Bologna, Italy

Corresponding author: Casadio, Rita (rita.casadio@unibo.it)

Current Opinion in Structural Biology 2023, **81**:102641

This review comes from a themed issue on **Sequences and Topology (2023)**

Edited by **Madan Babu** and **Rita Casadio**

For complete overview of the section, please refer the article collection - Sequences and Topology (2023)

Available online 28 June 2023

https://doi.org/10.1016/j.sbi.2023.102641

Introduction

Protein sequence motifs are patterns of residues with different structural and/or functional features. They were recognized with protein multiple sequence/structural alignment methods as typical conserved signatures of protein families [1]. The general term "protein motif" includes different types of signatures ranging from short motifs (e.g., myristoylation, phosphorylation and glycosylation) to others which do not rely on conserved residues (e.g., transmembrane regions, signal sequences, cell sorting signals). Most of these motifs have been identified experimentally and have been adopted in the past thirty years or so for implementing simple statistical and machine learning based methods, suited for protein sequence analysis. They have provided practical tools for structurally and/or functionally annotating entire proteomes [2]. The increase of data alongside with the advent of deep learning techniques, and the application of natural language models to encode proteins (protein language models, pLMs) promote revisiting and refreshing most of the available predictors for protein motif recognition [3,4]. In the following, we will review recent methods based on deep learning and pLMs, or both, to highlight the state of the art when addressing the problem of finding motifs in proteins. The aim is that of stirring new approaches to find better solutions to old problems and to highlight those topics where improvement is still necessary.

Classical and deep machine learning

Historically, inference methods for motif prediction adopted statistics based on frequentist or Bayesian approaches or both. Starting from about thirty years ago, classical machine learning (ML) helped in extracting general rules of associations among the input (any given protein sequence) and the output (the residue or sequence fragment with the functional feature). In this scenario, data were modeled mainly through supervised ML procedures which allowed to set the model parameters with the goal of inferring the property at hand for never seen before sequences [5,6]. Importantly, this type of inference process allows to compute statistical scores suited to measure the performance of the method [7].

The relevance of the training set

With the advent of next generation sequencing machines in molecular biology, data bases of genes and proteins started increasing at an unprecedented rate. Concomitantly more ML tools have been developed in order to speed up the annotation process, including the ones based on Deep Learning (DL, see BOX 1) [8,9]. Learning becomes deep as soon as the number of hidden layers in a neural network exceeds two [8].

Routinely, during the annotation process, any protein sequence, although simply translated from the corresponding coding sequence, is endowed with structural

and functional features that are often obtained on the basis of prediction tools (e.g., the Biocuration Process at Uniprot, https://www.uniprot.org/help/biocuration; InterPro, https://www.ebi.ac.uk/interpro/).

Unfortunately, experiments were/are not synchronous with prediction outputs that were/are added to the protein file in the databases. These include UniProt, the largest archive of protein sequences (https://www.uniprot.org/). In UniProt files, GO terms describe functional annotation according to the Gene Ontology resource (GO; http://geneontology.org/). In order to distinguish among experimentally derived features (the ground truth data) and those acquired by sequence similarity with family templates and/or by computational means, each GO term is routinely endowed with a tag code, known as ECO code (these codes are from the Evidence and Conclusion Ontology, ECO codes; https://evidenceontology.org/). Given this scenario, most authors, considered in this review, when selecting sequences to be included in a training set declare that annotations are selected according to "manual assertion based on experiment," with ECO:0000269 and related ones.

When searching in a data base, such as UniProt, for sequences with specific protein motifs, we are basically facing two kinds of problems. One is that experiments probing motif functionality are costly and difficult to perform, and this often constrains the number of experimentally annotated sequences (with ECO:0000269 and related ones). The other is that inference of computed functionalities cannot be properly validated. This is so, rather independently of the performance of the computational method adopted, which is routinely scored on a small set of data.

As previously reported [5,6], the choice of the training set is therefore of critical relevance for the final output of the method. This is particularly true for predicting functional motifs in proteins, considering that they are signature of protein families and that, although pointing to the same functionality, they may have somehow different physicochemical characteristics in different organisms. In any case, either classical or deep machine learning requires proper training sets, possibly of well characterized experimental data (a larger volume of quality data, in the case of deep learning).

Standards for ML based methods

The more the volume of data bases increases, the more ML and DL methods have been/are adopted for their generalization capability, flexibility and the possibility of assessing reliability when addressing the problem of the correctness of the annotation.

The community of tool developers and users faces the problem of comparing different machine learning procedures for selecting the most reliable ones. For this, sets of rules have been put forward [10,11]. Recently, with the advent of more complicated and complex DL schemes (see BOX 1, where only the methods at the basis of the tools described in this review are briefly considered), it became even more necessary to set standards for machine learning, when developing methods [10–12].

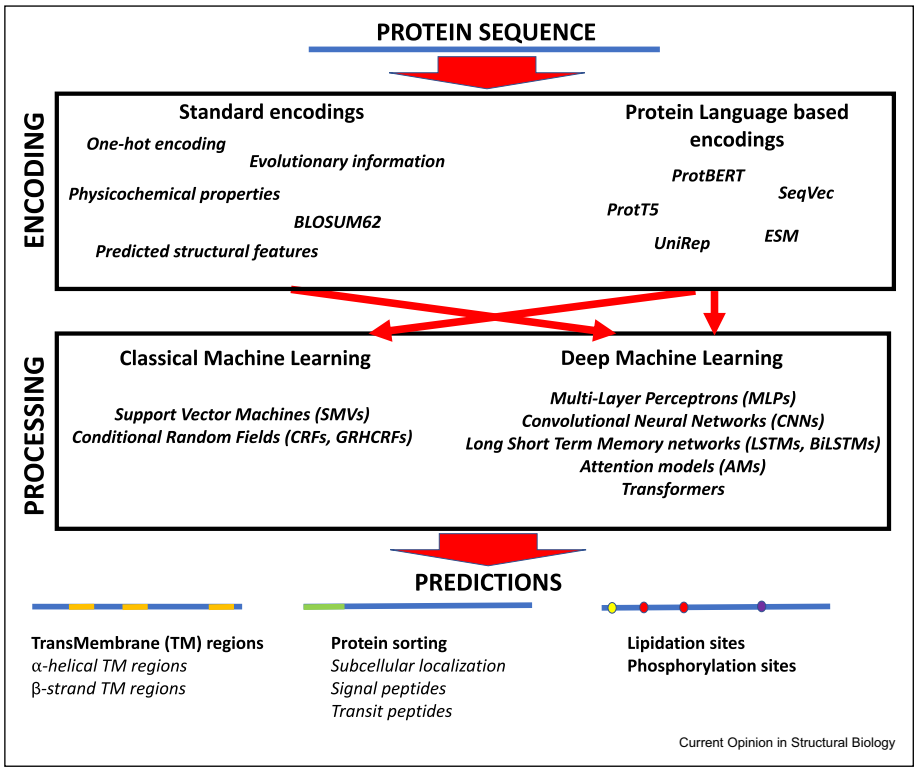
As discussed at length, method standardization requires that available data are divided into two sets: the *training* set used to train the model and the *testing* set used to evaluate the model, rigorously with non-similar protein sequences. A common variant to this procedure is the n-fold cross validation with n splits of the training set (routinely from 5 to 10). Training is performed over n-1 splits and testing is done on the 1. The final results are the average values of the n training runs. The n-fold cross validation is also run to mitigate information leakage (similarity between training and testing sets) and this is routinely performed by clustering all the protein sequences with a given threshold of identity (higher than 20%–30%) in the same split. Finally, a *blind* test set is adopted to validate the method performance and to compare it with previous tools addressing the same problem and performing at the State-Of-The-Art (SOTA). Validation on only the n sets of cross validation does not prevent information leakage.

In spite of this, when different methods are published a direct comparison on the basis of scoring indexes is impossible for different reasons: different training/testing sets and different encoding schemes are adopted even when the problem is the same, different architectures are implemented and sometimes even different scoring indexes are computed. For this reasons, international experiments have been/are carried out to benchmark methods on the same problem, such as CASP (https://predictioncenter.org/index.cgi, since 1994) for the critical assessment of protein structure prediction, CAFA (https://biofunctionprediction.org/cafa/, since 2010) for protein functional annotation and CAGI (https://genomeinterpretation.org/, since 2011) for genome interpretation.

Data encoding

If the methodological procedure is somehow standard, a particular and critical issue for ML/DL tools is how to encode input data for training. Different strategies have been/are adopted. By considering the tools quoted in this review, we may classify encoding procedures into classical and protein language-based ones (Figure 1). Classical ones include one-hot encoding, structural derived features and/or multiple sequence alignment to include evolutionary information [13]. Arrows between the ENCODING and the PROCESSING boxes emphasize the different possible combinations exploited by the methods reviewed.

Figure 1



Schematic view of encoding procedures associated to processing algorithms adopted in the prediction of motifs in protein sequence (described in Table 1).

Other encoding schemes derive from Natural Language Processing (pLMs, Figure 1). Some of them have been applied with success to the field of protein sequence and structure analysis [3,4,14]. These models take advantage of unifying all the possible information learned by filtering hundred million of protein sequences from all species with complex architectures of Neural Networks (see Box 1). Knowledge is casted into optimized weight values [3,4,14]. For each residue of the protein sequence, pLMs are also carrying residue context specific information derived again from all the sequences in the data bases. The different pLMs adopted when predicting sequence motifs, are listed in Figure 1, their difference depending on the volume of the data base they were trained on and/or the architecture/s by which they were computed (Box 1). pLM complexity routinely requires high computational costs and researchers adopt pre-computed pLMs in order to encode their specific training sets.

Recent results in predicting protein sequence motifs

Needless saying that recently protein sequence analysis has been carried out mainly by developing methods based on deep learning for training/testing. This improved SOTA over previous releases from the same or other groups. In the following, we will review recent methods based on machine learning addressing the problem of finding functional and structural motifs starting from the protein sequence. Out of literature, we selected methods with three constraints: i) post-print publication in the last five years, ii) based on DL and/or pLM embeddings, iii) availability to the research community.

Table 1 lists methods complying with our constraints. Researchers focused on three major groups of motifs: transmembrane regions, protein sorting and single residue Post Translational Modifications (PTMs). For each tool, the table lists the name together with the reference article and availability (first column, Name), major characteristics (second column, Methods), the encoding (Standard, StE and protein language model based, pLMs) and training procedure (classical (ML) and deep (DL) machine learning; third column, Type), the amplitude of the training set (fourth column, Training set), and (fifth column, Performance) the performance of the method as scored by the authors. When available, we detail the composition of the testing set and its identity to the training one. Some time, possibly due to

BOX 1. Encoding and Processing

Standard encoding

Standard encoding schemes provide numerical representations of the sequence and/or of the multiple sequence alignments (MSA) of similar proteins (13).

One-hot encoding: a representation of the residue sequence as a L (sequence length) ×20 matrix; in each sequence position, all elements are equal to zero, but the one corresponding to the residue type to be encoded.

Evolutionary information: a representation of the MSA as a Lx20 matrix, which contains, position by position, the frequency of each residue in the MSA (**sequence profiles**), or the log-odd of the frequency of each residue in the MSA with respect to the background frequency of the same residue in a large protein dataset (**Position Specific Scoring Matrices, PSSMs**).

Physicochemical properties: values characterizing each residue type (e.g., hydrophobicity) and derived from pre-compiled scales.

Predicted structural features: values derived from external tools that provide information on the putative protein structure (e.g., secondary structure and residue solvent accessibility) on the basis of inference processes routinely based on machine-learning.

Protein Language Models (pLMs)

pLMs provide an alternative way to encode protein sequences. They are based on deep learning models trained with self-supervised learning algorithms run on a large corpus of protein sequences. After training, pLMs are able to map a sequence into an internal representation which provides, for each position, high-dimensional and context-sensitive vector embeddings. Different pLMs are available, differing in the underlying deep-learning models and training datasets, which routinely include over 100 million (M) sequences [3,4].

UniRep model computes 64-, 256-, and 1900-dimensional embeddings of each residue in a sequence. It is based on a modified version of LSTM networks (see below). The training set consists of 24 M protein sequences from UniRef50 [35].

SeqVec model computes a 1024-dimensional embedding of each residue in a sequence. It is based on biLSTM networks (see below). The training set consists of 33 M protein sequences from UniRef50 [36].

ESM-1 model computes a 1280-dimensional embedding of each residue in a sequence. It is based on a biLSTM network (see below). It was trained on 250 M sequences from the UniParc dataset [37].

ProtT5 model computes a 1024-dimensional embedding of each residue in a sequence. It is based on the T5 Transformer architecture (see below). It was pre-trained on 2.1 B sequences of the metagenomic derived BFD database and fine-tuned on 45 M sequences from UniRef50 [38].

ProtBERT model computes a 1024-dimensional embedding of each residue in a sequence. It is based on a BERT Transformer architecture (see below). It was pre-trained on 2.1 B sequences of the metagenomic derived BFD database and 216 M sequences from UniRef100 [39].

Processing

Classical machine learning

Several algorithms have been implemented, before the advent of complex Deep Learning models, to solve classification and labelling problems in bioinformatics. Methods described in this review include the following.

Support vector machines (SVMs): binary classifiers separating two linearly separable sets, identifying an optimal separation hyperplane, and maximizing the distance between the classes. When coupled with kernel they can solve non-linearly separable problems [6].

Conditional random fields (CRFs): Markovian models devise to sequence labelling tasks. As for Hidden Markov Models, training and inference are based on dynamic programming algorithms [7]. A variant called **Grammatical-Restrained Hidden Random Fields (GRHCRF)** allows constraining the labelling according to user-defined regular grammar rules [39].

Deep machine learning

Deep learning (DL) techniques are based on complex architectures of neural networks, comprising several processing layers [8].

Multilayer perceptron (MLPs): the most basic DL model consisting of multiple layers of neurons, consecutively connected in a feed-forward architecture elaborating information from the input to the output layers [8,9].

Convolutional neural network (CNNs): DL architecture devised to pattern recognition (in images, time series, or sequences). Pattern detection is performed using a series of filters scanning the input and providing a feature mapping to be further elaborated by the cascading layers of the network [8,9].

Long-short term memory (LSTM) network: a neural network mapping input sequences to output sequences maintaining an internal state which aggregates information from the sequence following the processing direction. This allows keeping track of long- and short−range correlations along the input sequence. When the processing is bidirectional (begin-to-end and end-to-begin), the model is referred to as **bidirectional LSTM (BiLSTM)** [8,9].

Attention mechanism: a technique which allows enhancing important regions in the input and depressing less relevant parts, with respect to a specific prediction task. Different implementations of attention mechanisms have been devised. **Light Attention** is based on CNN architectures [24].

Transformer: a neural network model that transforms a sequence into another (e.g., language translation), based on self-attention mechanisms. Combinations of Transformer building blocks defines different architectures, such as T5 and BERT [9].

Scoring indexes

The performance of binary classifiers is routinely scored with different indexes based on the number of correct (true, T) or wrong (false, F) predictions in either the positive (P) or negative (N) classes [6].

Recall (REC) estimates the classifier ability to recognize the positive examples.

$$REC = \frac{TP}{TP + FN}$$

Precision (PRE) estimates the rate of success of the classifier when predicting the positive class.

$$PRE = \frac{TP}{TP + FP}$$

F1-score (F1) [40] combines precision and recall in a single number by computing their harmonic mean.

$$F1 = \frac{2 \times PRE \times REC}{PRE + REC}$$

Matthews correlation coefficient (MCC) provides a balanced measure accounting for true and false predictions in both classes.

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

It ranges between −1 and 1, being 1, 0, and −1 the scores obtained by perfect, random, and totally wrong predictions, respectively.

the limited size of training sets, only cross validation results are available. For each group of predictors, we report only one type of scoring index with the aim of giving a general overview of the performance in relation to the same problem. One should be aware that in each article the authors are benchmarking their method towards other ones, often including classical learning-approaches, from the same or other groups. From these data we can generally conclude that apparently the introduction of DL based methods helps in making a step forward in the prediction scores [11].

If this is so, then one open question is to which extent can we highlight the effect of the encoding scheme, and to which extent pLMs are relevant for making progress in the solution of the problem at hand.

We should keep in mind that the comparison among methods addressing the same problem cannot ground only on the direct comparison of scoring index values, and that these can only give an estimate of the different real performances unless an external benchmarking is done.

Table 1

Recent deep-learning and/or pLM embedding based prediction tools for discriminating various types of functional motifs.

Name	Methods	Type[a]	Training set	Performance
TransMembrane Regions (TMR)				
TMbed [15] (*α-helix and β-barrel TMRs*) (2022) https://github.com/BernhoferM/TMbed	ProtT5; CNN + Viterbi decoding	pLM DL	593 α-helical TMPs 65 β-barrel TMPs 3-D structures	F1 (segment-level, α-helix) = 0.83 F1 (segment-level, β-barrel) = 0.93 Blind set: 86 α-helical TMPs, 14 β-barrel TMPs; I ≤ 20%
DeepTMpred [16] (*α-helix TMRs*) (2022) https://github.com/ISYSLAB-HUST/DeepTMpred	ESM; CNN + CRF	pLM DL	582 α-helical TMPs 3-D structures	F1 (segment-level, α-helix) = 0.87 Blind set: 40 α-helical TMPs; I ≤ 30%
BetAware-Deep [17] (*β-barrel TMRs*) (2021) https://busca.biocomp.unibo.it/betaware2	Evolutionary information + profile-weighted hydrophobic moments; BiLSTM + GRHCRF	StE DL	58 β-barrel TMPs 3-D structures	F1 (segment-level, β-barrel) = 0.82 Blind set: 15 β-barrel TMPs; I ≤ 25%
MemBrain [18] (*α-helix TMRs*) (2020) http://www.csbio.sjtu.edu.cn/bioinf/MemBrain/	Evolutionary information + predicted structural features + physicochemical properties; CNN	StE DL	318 α-helical TMPs 3-D structures	F1 (segment-level, α-helix TMR) = 0.81 Blind set: 40 α-helical TMPs; I ≤ 20%
Protein sorting				
Subcellular localization: 2-class discrimination				
SCLpred-MEM [19] (*Secretory pathway membrane proteins localization*) (2021) http://distilldeep.ucd.ie/SCLpred-MEM	Evolutionary information; CNN	StE DL	SCLpred-MEM: 6988 proteins Exp. evidence	MCC = 0.62 Blind set: 240 proteins; I ≤ 30%
SCLpred-EMS [20] (*Secretory pathway localization*) (2020) http://distilldeep.ucd.ie/SCLpred2	Evolutionary information; CNN	StE DL	SCLpred-EMS: 19,579 proteins Exp. evidence	MCC = 0.82 Blind set: 593 proteins; I ≤ 30%
In-Pero [21] (*2 peroxisomal localizations*) (2021) https://organelx.hpc.rug.nl/fasta/	Concatenated UniRep and SeqVec; SVM	pLMs ML	160 proteins Literature evidence	MCC = 0.72 ± 0.06 10-fold cross-validation; I ≤ 40%
Subcellular localization: 4-class discrimination				
In-Mito [21] (*4 mitochondrial SLs*) (2021) https://organelx.hpc.rug.nl/fasta/	Concatenated UniRep and SeqVec; SVM	pLMs ML	424 proteins Exp. evidence	MCC (outer membrane) = 0.64 MCC (inner membrane) = 0.69 MCC (intermembrane) = 0.62 MCC (matrix) = 0.80 10-fold cross-validation; I ≤ 40%
DeepMito [22] (*4 mitochondrial SLs*) (2020) http://busca.biocomp.unibo.it/deepmito/	Position specific scoring matrices; CNN	StE DL	424 proteins Exp. evidence	MCC (outer membrane) = 0.46 MCC (inner membrane) = 0.47 MCC (intermembrane) = 0.53 MCC (matrix) = 0.65 10-fold cross-validation; I ≤ 40%

Subcellular localization: 10-class discrimination				
DeepLoc2.0 [23] (*10 SL compartments*) (2022) https://services.healthtech.dtu.dk/service.php?DeepLoc-2.0	ProtT5; attention mechanism + MLP	pLM DL	28,303 proteins (6684 multilocalized) Exp. evidence	See Table 2
LAProtT5 [24] (*10 SL compartments*) (2021) https://github.com/HannesStark/protein-localization https://embed.protein.properties/	ProtT5; light-attention mechanisms (based on CNN) + MLP	pLM DL	13,858 proteins (single localized) Exp. evidence	See Table 2
Signal peptides (SP)				
SignalP 6.0 [25] (2022) https://services.healthtech.dtu.dk/service.php?SignalP	Five types of SPs, SP detection and cleavage-site prediction (CSpred): ProtBERT embedding; CRF	pLM ML	4665 proteins with SP 15,625 proteins without SP Exp. evidence, manually reannotated	**Euk**: $MCC_{detection}$ = 0.87, REC_{CSpred} = 0.75 Blind set: 146 SP and 5581 non-SP proteins **Gram+**: $MCC_{detection}$ = 0.81, REC_{CSpred} = 0.80 Blind set: 156 SP and 81 non-SP proteins **Gram-**: $MCC_{detection}$ = 0.88, REC_{CSpred} = 0.64 Blind set: 374 SP and 133 non-SP proteins I $\leq$ 30%
Signal-3L 3.0 [26] (2020) http://www.csbio.sjtu.edu.cn/bioinf/Signal-3L/	SP detection: Evolutionary information; BiLSTM + Attention layer. SP cleavage site prediction (CSpred): Evolutionary information; Bidirectional CNN + CRF	StE DL	3309 proteins with SP 15,268 proteins without SP Exp. evidence	**Euk:** $MCC_{detection}$ = 0.93, REC_{CSpred} = 0.67 Blind set: 210 SP and 7247 non-SP proteins **Gram+**: $MCC_{detection}$ = 0.97, REC_{CSpred} = 0.76 Blind set: 90 SP and 153 non-SP proteins **Gram-**: $MCC_{detection}$ = 0.97, REC_{CSpred} = 0.67 Blind set: 25 SP and 89 non-SP proteins I $\leq$ 20%
DeepSig [27] (2018) https://deepsig.biocomp.unibo.it	SP detection: One-hot encoding; CNN SP cleavage-site prediction (CSpred): One-hot encoding; Attention + GRHCRF	StE DL	2271 proteins with SP 8032 proteins without SP Exp. evidence	**Euk**: $MCC_{detection}$ = 0.86, REC_{CSpred} = 0.76 Blind set: 46 SP and 1012 non-SP prot **Gram+**: $MCC_{detection}$ = 0.54, REC_{CSpred} = 0.44 Blind set: 9 SP and 429 non-SP proteins **Gram-**: $MCC_{detection}$ = 0.95, REC_{CSpred} = 0.78 Blind set: 23 SP and 188 non-SP proteins I $\leq$ 25%
Transit peptides (TP)				
TargetP 2.0 [28] (2019) https://services.healthtech.dtu.dk/service.php?TargetP-2.0	BLOSUM62 encoding; BiLSTM + Attention	StE DL	499 proteins (mitoTP) 227 proteins (chloroTP) 45 proteins (thylakTP) 19,234 proteins without TPs Exp. evidence	MCC (mitoTP) = 0.86 MCC (chloroTP) = 0.88 MCC (thylakTP) = 0.75 5-fold cross-validation; I<20%
Single residue post-translational modifications (PTMs)				
Lipidation				
NetGPI [29] (*GPI-anchor and ω-site prediction*) (2021) https://services.healthtech.dtu.dk/service.php?NetGPI	One hot-encoding; BiLSTM + Attention mechanism	StE DL	966 GPI-anchored proteins 2573 non GPI-anchored proteins Exp. evidence	$MCC_{detection}$ = 0.90, $REC_{\omega\text{-site}}$ = 0.47 Blind set: 160 GPI-anchored (50 with known ω-site) and 2573 non GPI-anchored proteins; I<30%

(*continued on next page*)

Table 1. (*continued*)

Name	Methods	Type[a]	Training set	Performance
Phosphorylation				
TransPhos [30] (2022) https://github.com/flyeagle0/TransPhos	One hot-encoding; Transformer + CNN	StE DL	20,964 phos-Ser 5685 phos-Thr 2163 phos-Tyr Exp. evidence	MCC (Ser) = 0.43 MCC (Thr) = 0.25 MCC (Tyr) = 0.15 Blind set: 5437 Ser, 1686 Thr, and 676 Tyr residues

For definition of F1, REC, and MCC scores, see Box 1.
I: pairwise sequence identity between training and testing sets.
In Methods column, semi-colons separate encoding methods from processing methods. For the explanation of the different input encodings and processing tools, see Box 1. When available, cross validation results report the standard error over the split test sets.
[a] In column Type: DL = Deep learning, pLM = Protein language models, StE = Standard encodings.

Prediction of transmembrane regions from the sequence

Historically the problem of predicting membrane protein topology is one of the first addressed with computational tools [31]. In Table 1, the present SOTA methods [15–17] adopt deep learning with post prediction refinements (CRF, GRHCRF, Viterbi decoding; Box 1) and pLM encodings to improve prediction scores [15,16]. TMBed [15] discriminates both alpha and beta motifs, whereas the other predictors focus on either alpha or beta regions [16–18].

Protein sorting

Protein sorting is the biological mechanism by which proteins are transported to their appropriate destinations within or outside the cell. After translation at the ribosomes, proteins can be targeted to the inner space of an organelle, different intracellular membranes, the plasma membrane, or to the exterior of the cell via secretion [32]. Apparently, information contained in the protein sequence triggers the delivery process. In Table 1, results on this important functional annotation are organised depending on how many classes of subcellular localisation are discriminated. Signal and transit peptide predictors are as well included.

Subcellular localization

2-class discrimination. SCLpred-MEM [19] and SCLpred-EMS [20] are from the same group. They adopt the same standard encoding method and the same DL (CNN; Box 1). The goal is different: discriminating the membrane proteins in the secretory pathways (SCLpred-MEM) and discriminating all the proteins in the secretory pathways (SCLpred-EMS), respectively. The training and testing sets are of different dimensions: larger in the second case. Results indicate that the size of the data is improving the method performance. In-Pero [21] adopts a concatenation of two pLMs (UniRep and SeqVec; Box 1) and a classical classifier (SVM; Box 1), and reports a 10-fold cross validation MCC value for the discrimination of proteins in the membrane and lumen localisation in peroxisomes.

Subcellular localization

4-class discrimination. In-Mito [21] and DeepMito [22] have been trained on the same small training/testing set (containing 424 proteins). In-Mito adopts a concatenated pLM model (as in In-Pero) and a classical classifier, whereas DeepMito adopts a standard encoding scheme with a deep learning architecture. It appears that the performance, adopting a 10-fold cross validation, is higher when pLMs are introduced.

Subcellular localization

10-class discrimination. The 10-class discrimination is the forefront, being ten the number of the main compartments routinely highlighted in eukaryotic cells [32]. DeepLoc2.0 [23] is a multilabel predictor that stands on

the ProtT5 pLM and on multilayer NNs endowed with an attention mechanism (Box 1); LAProtT5 is a single label predictor over the possible ten classes, encodes with ProtT5 and discriminates with a mechanism including "light attention" and multilayer NNs ([24]; Box 1). They are adopting different training/testing sets and their respective performances are listed in Table 2. LAProtT5 reports performances over the ten classes on a blind test set including 2768 proteins (second column in Table 2); DeepLoc2 reports scores in a 5-fold cross validation. What is interesting is that when both methods score themselves on different blind test sets (a so called setHard including 490 internally non redundant proteins in the case of LAProtT5, and a blind test set including 1717 human proteins from the Human Protein Atlas in the case of DeepLoc 2; third and fourth column, respectively), their performances significantly decrease.

Signal peptides

Three predictors are available. For SignalP, with different releases through the last decades, we select the last one which according to the authors is overperforming the previous ones (SignalP 6.0) [25]. The evolution of the methods through the last five years is quite clear in going from DeepSig [27] to Signal-3L 3.0 [26] and SignalP 6.0 [25]. Evidently, the richer input standard scheme of Signal-3L 3.0 together with an increased number of proteins in the training set, particularly for Gram positives (Gram+) is sufficient to overperform DeepSig in signal peptide detection, but not when predicting the cleavage sites (cs) (Recall cs sites, fifth column of Table 1; Box 1). SignalP 6.0 discriminates five types of signal peptides in relation to the different cleavage mechanisms [32], and in this it is superior to the previous two. When considering only the performance on eukaryotes, it appears that ProtBERT embedding associated to a classical ML classifier (CRF) is not outperforming Signal-3L 3.0. However, the cleavage site prediction score is comparable among the three predictors, with the exception of DeepSig on Gram+, which was scarcely populated.

Transit peptides

Transit peptides are responsible for the transport of a protein encoded by a nuclear gene to a particular organelle. For this problem only one predictor is available, TargetP 2.0 [28] which utilises DL with a standard encoding.

Single residue post translational modifications

For this particular type of motifs, requiring a one residue modification, only predictors implementing deep learning with standard encoding are presently available.

Lipidation

Several eukaryotic proteins associated to the extracellular leaflet of the plasma membrane carry a glycosylphosphatidylinositol (GPI) anchor, which is linked to the C-terminal residue after a proteolytic cleavage occurring at the so called ω-site [34]. NetGPI [29] predicts whether a protein undergoes a GPI

Table 2

LAProtT5 and DeepLoc2 predicting ten classes of eukaryotic subcellular localization.

Classes	LAProtT5[a] Blind:setDeepLoc I ≤ 30% (2768)	LAProtT5[b] Blind: setHARD I ≤ 30% (490)	DeepLoc2[c] Blind: HPA I ≤ 30% (1717)	DeepLoc2[d] Cross validation I ≤ 30%; 5 sets
Nucleus	0.84 (806)	0.70 (99)	0.44 (893)	0.69 ± 0.02 (9720)
Cytoplasm	0.69 (505)	0.49 (117)	0.36 (562)	0.62 ± 0.01 (9870)
Extracellular	0.94 (393)	0.78 (92)		0.85 ± 0.05 (3301)
Mitochondrion	0.87 (302)	0.75 (10)	0.56 (196)	0.76 ± 0.04 (2590)
Cell membrane	0.75 (273)	0.47 (98)	0.36 (287)	0.66 ± 0.02 (4187)
ER	0.62 (173)	0.47 (34)	0.17 (77)	0.56 ± 0.04 (2180)
Plastid	0.92 (152)	0.85 (11)		0.90 ± 0.04 (1047)
Golgi	0.55 (70)	0.10 (13)	0.31 (86)	0.31 ± 0.04 (1279)
Lysosome	0.17 (64)	0.19 (13)		0.28 ± 0.05 (1496)
Peroxysome	0.42 (30)	0.58 (3)		0.55 ± 0.04 (304)

Performance is scored with Matthews correlation coefficient (MCC). Cross validation results report the standard error over the 5 split test sets. Among brackets the number of proteins in each set and class.

ER: Endoplasmic reticulum.

I: pairwise sequence identity among proteins in training and testing sets.

[a] Data from the study by Thumuluri et el. [24]: dataset extracted from DeepLoc [33].

[b] Data computed on the internally non redundant setHARD testing set with the released LAProtT5 trained model [24].

[c] Data from the study by Stärk et al. [23], dataset from the Human Proteome Atlas (HPA) [23]. Out of 1717 proteins, 350 (20%), are multilocalized.

[d] Data from the study by Stärk et al. [23]. Out 28,303 proteins, 6684 (24%) are multilocalized.

modification at the C-terminus with a standard encoding embedding and DL, scoring with a high MCC value in detection. However, as to the ω-site prediction, scoring still deserves improvement.

Phosphorylation

Protein phosphorylation is a reversible post-translational modification of proteins in which mainly residues such as serine, threonine, and tyrosine in eukaryotes are phosphorylated by protein kinases, with the addition of a covalently bound phosphate group. This has important and well-characterized roles in signalling pathways and metabolism [32]. TransPhos [30] based on a standard encoding scheme and DL discriminates the phosphorylated residues, with scores that suggest improvement.

Conclusions and perspectives

Recent tools for the detection of protein sequence motifs stand on machine learning architectures of different complexities. The most recent ones adopt encoding based on protein language models (Box1).

Detecting transmembrane regions is more successful for predictors adopting pLMs and DL (TMbed, DeepTMpred).

Protein sorting includes different types of subcellular localisation discrimination: 1) for the 2-class discrimination and when the strategy is the same (standard encoding and deep learning) the score seems affected by the volume of the training/testing set (SCLpred-MEM vs SCLpred-EMS); when the training set is of small dimension, the adoption of pLMs and ML improves over previous versions of the same prediction task, although results of cross validation are not conclusive (In-Pero). 2) Both In-Mito and DeepMito were trained on the same small data set. Evidently pLMs and ML (In-Pero) outperform (however only in cross validation) on a standard encoding scheme and DL (DeepMito). 3) The prediction of subcellular localisation is upgraded to the point that apparently the ten different compartments of eukaryotes can be discriminated. DeepLoc2.0 and LAProtT5 take advantage of deep learning procedures including attention mechanisms and pLMs. Their performance is difficult to compare, given the multilabel vs the single label procedure (Table 2). According to the authors both methods poorly perform on unbiased blind test sets.

Signal peptide detection improved in the last five years from DeepSig (standard encoding and DL) to Signal-3L 3.0 (a richer standard encoding than DeepSig and DL) and up to the five-type signal peptide discrimination of SignalP 6.0 (pLM and ML). Interestingly the efficiency of the prediction of the cleavage site is not much affected by the different methods.

Transit peptides are satisfactorily predicted with standard encoding and DL (TargetP; results are in cross validation).

Single residue post translational modifications include NetGPI, based on standard encoding and DL, which discriminates GPI anchors with a high score and predicts with an ameliorable one the ω-site, and TransPhos that with a similar strategy predicts phosphorylation sites.

Overall pLMs, when applied with ML and/or DL appear to improve predictions over standard encoding mechanisms for the different tasks, suggesting that pLMs are powerful approaches to extract evolution information, overpassing multiple sequence and/or structural alignments. This is particularly true in the case of transmembrane region predictions, which can also rely on structural validation.

The volume of the training/testing sets is also relevant. Indeed, tasks such as the four-class mitochondrial discrimination, the transit peptides, the GPI-anchors, and single residue post translation modifications lack sufficient amount of experimental data to cope with the recent technological advancements. This suggests that experiments are necessary to populate the different classes of motifs. Furthermore, for a proper comparison of the different strategies, an external benchmarking, like CAFA or others to come, is necessary to understand which application among the ones available would get realistic scores.

Editorial Disclosure Statement

Given her role as Guest Editor, Rita Casadio had no involvement in the peer review of the article and has no access to information regarding its peer-review. Full responsibility for the editorial process of this article was delegated to Madan Babu.

Declaration of competing interest

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

Data availability

No data was used for the research described in the article.

Acknowledgments

The work was supported by PRIN2017 grant (project 2017483NH8_002), delivered to CS by the Italian Ministry of University and Research. We acknowledge ELIXIR-IIB, the Italian node of the ELIXIR infrastructure.

References

Papers of particular interest, published within the period of review, have been highlighted as:

* of special interest
** of outstanding interest

1. Bork P, Koonin EV: **Protein sequence motifs**. *Curr Opin Struct Biol* 1996, **6**:366–376.

2. Hou Q, Waury K, Gogishvili D, Feenstra KA: **Ten quick tips for sequence-based prediction of protein properties using machine learning**. *PLoS Comput Biol* 2022, **18**, e1010669.

3. ** Bepler T, Berger B: **Learning the protein language: evolution, structure, and function**. *Cell Syst* 2021, **12**:654. 669.e3.
A comprehensive review on protein language models and their applications in computational biology.

4. ** Ofer D, Brandes N, Linial M: **The language of proteins: NLP, machine learning and protein sequences**. *Comput Struct Biotechnol J* 2021, **19**:1750–1758.
A comprehensive review on the mechanisms of protein language models and their applications in computational biology.

5. Baldi P, Brunak S: *Bioinformatics: the machine learning approach*. MIT Press; 2001.

6. Bishop CM: *Pattern recognition and machine learning*. Springer; 2006.

7. Baldi P, Brunak S, Chauvin Y, Andersen CAF, Nielsen H: **Assessing the accuracy of prediction algorithms for classification: an overview**. *Bioinformatics* 2000, **16**:412–424.

8. Baldi P: *Deep learning in science*. Cambridge University Press; 2021.

9. Drori I: *The science of deep learning*. Cambridge University Press; 2023.

10. Jones DT: **Setting the standards for machine learning in biology**. *Nat Rev Mol Cell Biol* 2019, **20**:659–660.

11. Greener JG, Kandathil SM, Moffat L, Jones DT: **A guide to machine learning for biologists**. *Nat Rev Mol Cell Biol* 2021, https://doi.org/10.1038/s41580-021-00407-0.

12. Walsh I, Fishman D, Garcia-Gasulla D, Titma T, Pollastri G, , ELIXIR Machine Learning Focus Group, Capriotti E, Casadio R, Capella-Gutierrez S, Cirillo D, *et al.*: **DOME: recommendations for supervised machine learning validation in biology**. *Nat Methods* 2021, **18**:1122–1127.

13. Jing X, Dong Q, Hong D, Lu R: **Amino acid encoding methods for protein sequences: a comprehensive review and assessment**. *IEEE ACM Trans Comput Biol Bioinf* 2020, **17**:1918–1931.

14. Ibtehaz N, Kihara D: *Application of sequence embedding in protein sequence-based predictions*. 2021. arXiv:211007609 [q-bio].

15. Bernhofer M, Rost B: **TMbed: transmembrane proteins predicted through language model embeddings**. *BMC Bioinf* 2022, **23**:326.

16. Wang L, Zhong H, Xue Z, Wang Y: **Improving the topology prediction of α-helical transmembrane proteins with deep transfer learning**. *Comput Struct Biotechnol J* 2022, **20**:1993–2000.

17. Madeo G, Savojardo C, Martelli PL, Casadio R: **BetAware-deep: an accurate web server for discrimination and topology prediction of prokaryotic transmembrane β-barrel proteins**. *J Mol Biol* 2021, **433**:166729.

18. Feng S-H, Zhang W-X, Yang J, Yang Y, Shen H-B: **Topology prediction improvement of α-helical transmembrane proteins through helix-tail modeling and multiscale deep learning fusion**. *J Mol Biol* 2020, **432**:1279–1296.

19. Kaleel M, Ellinger L, Lalor C, Pollastri G, Mooney C, Sclpred-Mem: **Subcellular localization prediction of membrane proteins by deep N-to-1 convolutional neural networks**. *Proteins* 2021, **89**:1233–1239.

20. Kaleel M, Zheng Y, Chen J, Feng X, Simpson JC, Pollastri G, Mooney C, Sclpred-Ems: **Subcellular localization prediction of endomembrane system and secretory pathway proteins by Deep N-to-1 Convolutional Neural Networks**. *Bioinformatics* 2020, **36**:3343–3349.

21. Anteghini M, Martins dos Santos V, Saccenti E: **In-Pero: Exploiting deep learning embeddings of protein sequences to predict the localisation of peroxisomal proteins**. *Int J Mol Sci* 2021, **22**:6409.

22. Savojardo C, Bruciaferri N, Tartari G, Martelli PL, Casadio R: **DeepMito: accurate prediction of protein sub-mitochondrial localization using convolutional neural networks**. *Bioinformatics* 2019, https://doi.org/10.1093/bioinformatics/btz512.

23. * Thumuluri V, Almagro Armenteros JJ, Johansen AR, Nielsen H, Winther O: **DeepLoc 2.0: multi-label subcellular localization prediction using protein language models**. *Nucleic Acids Res* 2022, **50**:W228–W234.
Multilabel predictor discriminating ten different subcellular localizations in Eukaryotes with pLMs and DL.

24. * Stärk H, Dallago C, Heinzinger M, Rost B: **Light attention predicts protein location from the language of life**. *Bioinformatics Adv* 2021, **1**. vbab035.
Single-label predictor discriminating ten different subcellular localizations in Eukaryotes with pLMs and DL.

25. Teufel F, Almagro Armenteros JJ, Johansen AR, Gíslason MH, Pihl SI, Tsirigos KD, Winther O, Brunak S, von Heijne G, Nielsen H: **SignalP 6.0 predicts all five types of signal peptides using protein language models**. *Nat Biotechnol* 2022, **40**:1023–1025.

26. Zhang W-X, Pan X, Shen H-B: **Signal-3L 3.0: improving signal peptide prediction through combining attention deep learning with window-based scoring**. *J Chem Inf Model* 2020, **60**:3679–3686.

27. Savojardo C, Martelli PL, Fariselli P, Casadio R: **DeepSig: deep learning improves signal peptide detection in proteins**. *Bioinformatics* 2018, **34**:1690–1696.

28. Almagro Armenteros JJ, Salvatore M, Emanuelsson O, Winther O, von Heijne G, Elofsson A, Nielsen H: **Detecting sequence signals in targeting peptides using deep learning**. *Life Sci Alliance* 2019, **2**, e201900429.

29. Gíslason MH, Nielsen H, Almagro Armenteros JJ, Johansen AR: **Prediction of GPI-anchored proteins with pointer neural networks**. *Curr Res Biotechnol* 2021, **3**:6–13.

30. Wang X, Zhang Z, Zhang C, Meng X, Shi X, Qu P: **TransPhos: a deep-learning model for general phosphorylation site prediction based on transformer-encoder architecture**. *Int J Mol Sci* 2022, **23**:4263.

31. Chen CP, Rost B: **State-of-the-art in membrane protein prediction**. *Appl Bioinf* 2002, **1**:21–35.

32. Lodish H, Berk A, Kaiser CA, Krieger M, Bretscher A, Ploegh H, Martin KC, Yaffe MB, Amon A: *Molecular cell biology*. Macmillan Learning; 2021.

33. Almagro Armenteros JJ, Sonderby CK, Sonderby SK, Nielsen H, Winther O: **DeepLoc: prediction of protein subcellular localization using deep learning**. *Bioinformatics* 2017, **33**:3387–3395.

34. Mayor S, Riezman H: **Sorting GPI-anchored proteins**. *Nat Rev Mol Cell Biol* 2004, **5**:110–120.

35. Alley EC, Khimulya G, Biswas S, AlQuraishi M, Church GM: **Unified rational protein engineering with sequence-based deep representation learning**. *Nat Methods* 2019, **16**:1315–1322.

36. Heinzinger M, Elnaggar A, Wang Y, Dallago C, Nechaev D, Matthes F, Rost B: **Modeling aspects of the language of life**

through transfer-learning protein sequences. *BMC Bioinf* 2019, **20**:723.

37. Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, Guo D, Ott M, Zitnick CL, Ma J, *et al.*: **Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences**. *Proc Natl Acad Sci U S A* 2021, **118**, e2016239118.

38. Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, Gibbs T, Feher T, Angerer C, Steinegger M, *et al.*: **ProtTrans: towards cracking the language of lifes code through self-supervised deep learning and high performance computing**. *IEEE Trans Pattern Anal Mach Intell* 2021, https://doi.org/10.1109/TPAMI.2021.3095381.

39. Fariselli P, Savojardo C, Martelli PL, Casadio R: **Grammatical-restrained hidden conditional random fields for bioinformatics applications**. *Algorithm Mol Biol* 2009, **4**:13.

40. Blair DC: **Information retrieval 2nd ed. C.J. Van rijsbergen. London: butterworths; 1979: 208 pp**. *J Am Soc Inf Sci* 1979, **30**: 374–375.