Computational and Structural Biotechnology Journal 19 (2021) 5791–5799

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

DeepREx-WS: A web server for characterising protein–solvent interaction starting from sequence

Matteo Manfredi [a], Castrense Savojardo [a], Pier Luigi Martelli [a,*], Rita Casadio [a,b]

[a] Biocomputing Group, Department of Pharmacy and Biotechnology, University of Bologna, Bologna, Italy
[b] Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies (IBIOM), Italian National Research Council (CNR), Bari, Italy

ARTICLE INFO

Article history:
Received 5 August 2021
Received in revised form 7 October 2021
Accepted 7 October 2021
Available online 13 October 2021

Keywords:
Residue solvent accessibility
Deep Learning
Protein flexibility
Protein disorder
Surface engineering

ABSTRACT

Protein–solvent interaction provides important features for protein surface engineering when the structure is absent or partially solved. Presently, we can integrate the notion of solvent exposed/buried residues with that of their flexibility and intrinsic disorder to highlight regions where mutations may increase or decrease protein stability in order to modify proteins for biotechnological reasons, while preserving their functional integrity. Here we describe a web server, which provides the unique possibility of integrating knowledge of solvent and non-solvent exposure with that of residue conservation, flexibility and disorder of a protein sequence, for a better understanding of which regions are relevant for protein integrity. The core of the webserver is DeepREx, a novel deep learning-based tool that classifies each residue in the sequence as buried or exposed. DeepREx is trained on a high-quality, non-redundant dataset derived from the Protein Data Bank comprising 2332 monomeric protein chains and benchmarked on a blind test set including 200 protein sequences unrelated with the training set. Results show that DeepREx performs at the state-of-the-art in the field. In turn, the Web Server, DeepREx-WS, supplements the predictions of DeepREx with features that allow a better characterisation of exposed and buried regions: i) residue conservation derived from multiple sequence alignment; ii) local sequence hydrophobicity; iii) residue flexibility computed with MEDUSA; iv) a predictor of secondary structure; v) the presence of disordered regions as derived from MobiDB-Lite3.0. The web server allows browsing, selecting and intersecting the different features. We demonstrate a possible application of the DeepREx-WS for assisting the identification of residues to be variated in protein surface engineering processes.

1. Introduction

Knowledge of the exposure of a residue in the context of a folded protein allows defining the protein folding core and identifying residues that interact with the solvent and other molecules in physiological or artificial environments [1]. Solvent exposure is routinely measured by residue Solvent Accessible Surface Area (SASA) or its Relative Solvent Accessibility (RSA), in which the maximum surface area for each residue type is the normalizing factor [2–4]. Residues in any protein can be therefore classified as buried or exposed by defining a threshold on the RSA value, routinely set equal to 20%. Programs like DSSP [5] or PSAIA [6] estimate RSA starting from the Protein Data Bank (PDB) coordinates of a protein structure. When the three-dimensional structure of a protein is not or partially available, computational methods can predict solvent exposure from the protein sequence.

Different prediction tools, mainly based on machine-learning approaches, provide RSA estimation, classifying residues into buried or exposed [7–9]. Finer-grained predictions into three or four classes of solvent exposure are possible [10]. Recently, solvent exposure is computed with deep-learning approaches [10,11].

New developments in the protein structure prediction field led to the release of AlphaFold2 [12], a very powerful deep-learning based tool for the ab-initio prediction of protein three-dimensional (3D) structure from sequence. AlphaFold2 optimally scored in the most recent edition of the Critical Assessment of Structure Prediction (CASP, predictioncenter.org), although the accuracy is not uniform across all CASP target categories and still limited on difficult targets (e.g., the free-modelling ones). Despite the success of AlphaFold2, the availability of sequence-based predictors of protein features, like solvent exposure, are still important for many reasons. Accurate predictions of protein features

* Corresponding author.
E-mail address: pierluigi.martelli@unibo.it (P.L. Martelli).

https://doi.org/10.1016/j.csbj.2021.10.016

can be useful to validate models generated with AlphaFold2 (or with others ab-initio methods), particularly in those regions where the models are expected to be low quality. Moreover, predictions of solvent exposure can be helpful also in the perspective of being integrated into end-to-end deep-learning methods, even during the learning phase, to guide and refine the training process. Tools like AlphaFold2 are very demanding in terms of computational resources, whereas simple predictors of protein structural features can be easily adopted in the presence of time/resource constraints for the preliminary structural/functional characterization of large datasets of proteins. This allows the quick identification of interesting cases on which focusing the attention and, possibly, applying more sophisticated (and computationally demanding) approaches.

Computation of solvent exposure provides valuable information in different problems, which include defining constraints for *ab-initio* protein structure prediction tools, refining protein–protein interface predictors [13,14], and structurally and functionally characterizing sequence positions, which undergo pathogenic single-residue variations [15–17]. In biotechnological applications, knowledge of residue solvent exposure is of prominent importance. Rational surface engineering i.e., the chemical modification of key positions on the protein surface, is an effective tool for tailoring protein features to specific industrial and biotechnological demands [18,19] and references therein]. Applications of protein surface (re-)engineering include the improvement of protein solubility in different solvents [20,21], immobilization [22], and stabilization in aqueous or organic solvents [23,24]. In all these applications, computational prediction of protein solvent accessibility from sequence can provide constrains for screening the candidate sites to be considered for modifications when the experimental protein three-dimensional structure (or a validated structural model) is not available [19]. Other features, such as residue conservation in multiple sequence alignment, local protein flexibility, protein secondary structure and possibly the presence of intrinsically disordered regions can further reduce the search space, identifying residues not essential for protein function and/or located in external loops.

Here, we present DeepREx-WS, a web server providing a multi-dimensional characterization of exposed and buried positions of a protein starting from its residue sequence. A two-class prediction of protein solvent exposure is provided with a novel deep learning-based method, DeepREx. The new predictor described in this paper has been trained and tested on high-quality structures extracted from the PDB and performs at the state-of-the-art, when benchmarked against other methods available for the same task.

The server DeepREx-WS, for each position, supplements the exposure prediction of DeepREx with the Kyte-Doolittle hydrophobicity and residue conservation obtained from a multiple sequence alignment. Furthermore, three external resources, MEDUSA [25], PYTHIA [26] and MobiDB-Lite3.0 [27], are present to estimate, for each residue position, protein flexibility, protein secondary structure and the presence of intrinsically disordered regions, respectively.

We release DeepREx as both Python stand-alone program and Docker image.

2. Material and methods

2.1. DeepREx implementation

2.1.1. Datasets

DeepREx is trained and tested on a dataset extracted from the Protein Data Bank (PDB) [28] (accessed Oct 15, 2019), which includes 692,646 residues from 2532 non-redundant, monomeric proteins with an X-ray crystallographic structure at a resolution $\leq$ 2.5 Å and a coverage $\geq$ 70% of the corresponding UniProt sequence [29]. Mapping between PDB and UniProt sequences was retrieved with SIFTS [30]. Membrane proteins were excluded via a cross-check on the Orientations of Proteins in Membranes (OPM) database [31].

All proteins are declared by authors of the crystallography to be functional as monomers. The dataset was reduced by similarity, so that all protein sequences share $\leq$ 30% pairwise identity. The clustering and representative sequence selection have been performed using the MMseqs2 program [32]. Specifically, we used cluster mode 1 (single-linkage clustering) and 30% sequence identity threshold. No threshold has been set for coverage, allowing to cluster also sequences with very local sequence similarity. More details on the dataset collection are available in Supplementary Materials.

The absolute Solvent Accessible Surface Area (SASA) of each residue in the PDB file is computed using DSSP [5]. Relative Solvent Accessibility (RSA) values are then obtained dividing absolute SASA values by residue-specific maximal accessibility values, as extracted from the Sander and Rost scale [2]. Finally, each residue is classified as buried (B) if its RSA is $\leq$ 20%, and exposed (E) otherwise. This threshold divides the set of residues into two almost equally sized subsets, with 52% buried and 48% exposed residues and therefore provides a balanced dataset for training and testing.

The non-redundant dataset was then randomly split into a training set, comprising 2332 sequences, and a blind test set including 200 sequences. Proteins in the training set were further split into 10 equally sized sets for cross validation.

The blind test set includes 200 protein sequences (and their structures) from different organisms: 124 monomeric proteins from Bacteria, 56 from Eukaryotes, 15 from Archaea and 5 from Viruses. Moreover, these proteins cover a wide range of 3D SCOP/CATH [33,34] classes including 30 all-alpha proteins, 37 all-beta, 84 alpha/beta (a/b) and 16 alpha + beta (a + b) (32 proteins are unclassified). Overall, the 200 protein sequences contain 56,206 residues, 29,068 and 27,138 of which are buried and exposed, respectively, in the experimental 3D structure (for details, refer to Supplementary Table 1S).

Finally, we performed an additional comparative benchmark using 9 targets from the CASP14 experiment and previously used in literature for the evaluation of sequence-based prediction of protein features [26]. In particular, the chosen targets belong to the free modelling category i.e., no homologous sequences can be found for them and for this reason they are particularly challenging for structure prediction.

2.1.2. Input encoding

DeepREx is trained on 71 features, encoding for each position the protein sequence and information derived from Multiple Sequence Alignments (MSA).

MSA for each sequence in our dataset is generated with HHblits version 3 [35], setting two iterations and default parameters. Search is executed against the Uniclust30 database [36]. HHblits provides MSA and Hidden Markov Models (HMMs) adopted to guide the search of related sequences and from which we derived some of the features.

The 71-valued vector encoding each position *i* includes:

- The canonical residue one-hot encoding, representing primary-sequence information and accounting for 20 values.
- The protein sequence profile, computed from MSA and consisting of 21 values that account for the relative frequencies of each residue type (plus the gap) in the corresponding aligned position of the MSA.
- The HMM emission probabilities obtained from the match state in position *i* (20 values).

- The HMM transition probabilities (7 values), corresponding to all possible transitions between HMM states in position *i*.
- The 3 values of Neff_Match, Neff_Insertion and Neff_Deletion [35] computed by HHblits and encoding for the MSA local diversity around position *i*. These values provide the number of effective sequences (i.e., a sequence diversity estimation) for the subalignments comprising sequences having a match, an insertion and a deletion at position *i* of the full alignment, respectively.

2.1.3. The deep-learning architecture

Fig. 1 shows the deep architecture implemented in DeepREx.

Each sequence in the dataset is encoded as a *L*x71 matrix, where *L* is the protein length and 71 is the dimension of the encoding, as detailed in the previous section.

This input is firstly processed by three cascading Bidirectional Long-Short Term Memory (BLSTM) layers [37]. BLSTMs belong to the class of Long-Short Term Memory (LSTM) networks [38], a special recurrent neural network architecture well-suited for processing sequence data (e.g., protein sequences) and extracting relevant relations between elements of the sequence. Moreover, LSTMs have several advantages over traditional recurrent architectures in terms of stability of training and the proper handling of the vanishing gradient problem [39]. BLSTMs perform a double scanning of the input sequence, from left to right and vice versa, in order to better capture the sequential relations among sequence positions. Here, each BLSTM layer includes 32 activation units.

The output of the third recurrent layer is then provided as input to a time-distributed, fully-connected layer adopting a sigmoid activation function. This layer provides the final, binary classification of each residue in the sequence into buried or exposed classes. It computes a numerical output in the range [0,1] for each residue that can be interpreted as a probability for the residue to be exposed: all residues with $p \geq 0.5$ are predicted as exposed while those with $p < 0.5$ are classified as buried.

The method has been implemented with the Keras deep-learning Python library [40]. The total number of trainable parameters in the model is 76,353.

The output value *o* has been used to estimate the reliability index (RI) of the prediction:

$$RI = 2 \times |o - 0.5| \quad (1)$$

If *o* is close to 0.5 (uncertain classification), RI is close to 0. If *o* is close to 0 (strong classification in the buried class) or 1 (strong classification in the exposed class), RI is close to 1.

2.1.4. DeepREx training and evaluation

Training is performed by adopting a 10-fold cross-validation procedure, using 8 sets for training, one set for validation and early stopping (to avoid overfitting), and one for testing. Cross-validation results are reported as the average over performances computed on the testing sets. This training phase sets the optimal values of the architecture hyperparameters. Each model is trained for at most 1000 epochs. An early stopping procedure is adopted to reduce overfitting: the training procedure is stopped after 50 consecutive epochs when the error computed on the validation set does not decrease. The presence of sequences of variable length is handled using mini-batches of 64 sequences and zero-padding each sequence in the batch to the same length (i.e., the maximal length in the mini-batch). A masking layer, placed after the input layer, is used to ignore padded values. The ADAM optimizer [41] is adopted for gradient descent on the binary cross-entropy loss function. We run several complete cross-validations to select the optimal set of hyperparameters (number of activation units in LSTM layers, minibatch size, ADAM optimizer parameters). We chose the set of hyperparameters maximizing the performance of the method on the cross-validation validation sets.

Once the hyperparameters are fixed, the final DeepREx model for testing the blind set is obtained after training over the whole training set with the routinary procedure: 9/10 subsets are for

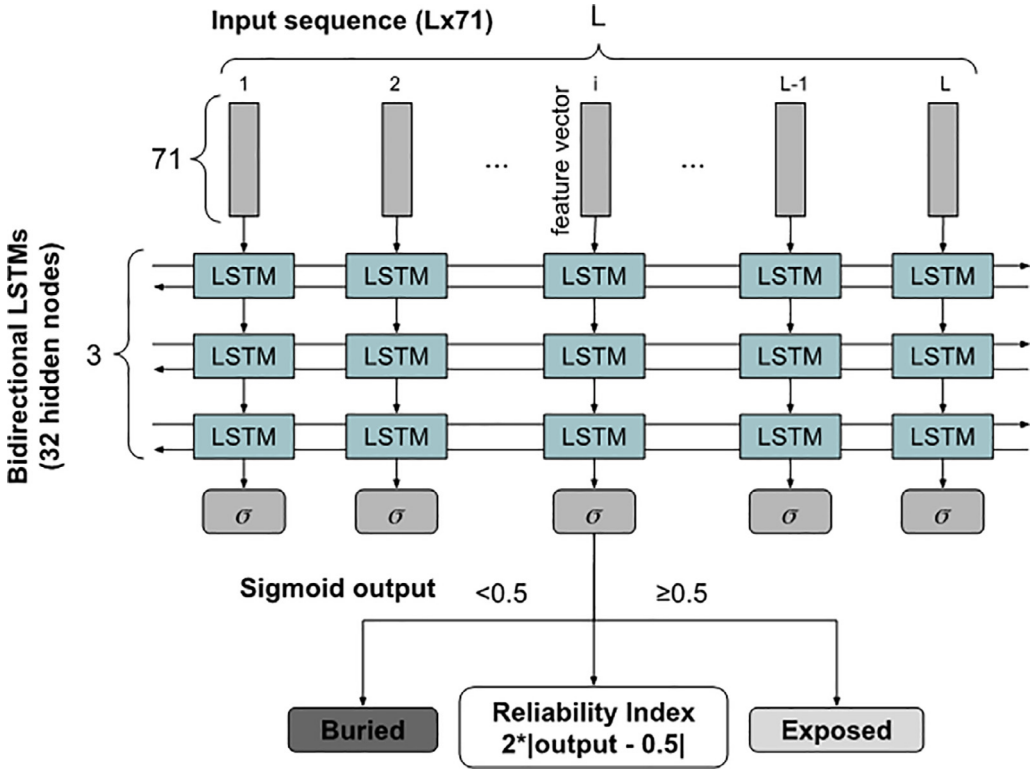


Fig. 1. Architecture of the deep neural network implemented in DeepREx to predict residue solvent exposure.

the actual training, while one random set among the 10 is adopted as validation set for early stopping. This final model is then tested on the 200 proteins of the blind test set and excluded from the training set to evaluate its performance.

2.1.5. Scoring indexes

The performance of the binary solvent accessibility classifiers is assessed with the following standard scores. Without loss of generality, we assume the exposed (E) and the buried (B) classes to be the positive and negative classes, respectively. In what follows, TP, TN, FP and FN are true positive, true negative, false positive and false negative predictions, respectively. The following scoring measures are computed:

- Accuracy (Q2), defined as:

$$Q_2 = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

- Precision:

$$Prec = \frac{TP}{TP + FP} \tag{4}$$

- Recall:

$$Rec = \frac{TP}{TP + FN} \tag{5}$$

- F1, the harmonic mean of the precision and recall, defined as:

$$F1 = \frac{2 * Prec * Rec}{Prec + Rec} \tag{6}$$

- Matthews Correlation Coefficient (MCC), defined as:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \tag{7}$$

2.2. The Web Server (DeepREx-WS) implementation

DeepREx-WS integrates DeepREx predictions with external resources. We include predictions obtained with MEDUSA [25], estimating residue flexibility of the proteins across five classes (0 = rigid, 4 = flexible). MEDUSA is based on a deep convolutional neural network architecture processing an input comprising evolutionary information, derived from MSAs and residue physicochemical properties [25].

We provide secondary structure prediction by means of PYTHIA, a protein local conformation prediction tool [26]. Specifically, PYTHIA (released in 2021, [26]) can be easily integrated in our web server, being released as a docker container. Furthermore, it runs fast, and it takes multiple sequence alignments as inputs. It is designed to predict local conformation in terms of Protein Blocks (PB). Overall, 16 PDB classes (labelled with lower-case letters, from *a* to *p*) are provided by PYTHIA: PB labels *a*, *b*, *c*, *d*, *e* and *f* represent different beta-strand regions (*c* is for the core of strand, *a*, *b* and *d*, *e* for N- anc C-terminal caps, respectively), PB labels *g*, *h*, *i* and *j* are all representing random coils while labels *k*, *l*, *m*, *n*, *o* and *p* map into alpha-helices (*m* for the helix core, *k, l* and *n, o* for N- and C-terminal caps, respectively). Here we mapped PB to secondary structure as follows: *c* to beta-strand (E), *m* to alpha-helix (H) and the remaining labels to random coil (C).

We integrate intrinsically disordered regions as predicted with MobiDB-Lite3.0 [27], providing a binary prediction for each residue (disordered/structured). MobiDB-Lite3.0 computes a consensus derived from the outputs of eight different predictors of disordered regions and applies a filtering procedure to get rid of spurious disorder predictions. All the three methods have been downloaded and are executed in-house.

Finally, DeepREx-WS also includes for each residue a hydrophobicity index, computed by averaging the Kyte-Doolittle hydropathy scale [42] over a window of 5 residues, and a conservation index computed from the MSA with the following equation:

$$CI(i) = 1.0 - \left(-\frac{1}{\log(20)} \sum_{a=1}^{20} f_a(i) \times \log\left[f_a(i) \right] \right) \tag{2}$$

where $f_a(i)$ is the frequency of the residue type a in the position i of the MSA. The CI ranges between 0 (not conserved) and 1 (fully conserved). The MSA used for computing the CI is the same provided in input to the DeepREx predictor and built for the input sequence using HHblits as detailed in section 2.1.2. The CI is only computed for MSA positions having at most 70% of gaps in the aligned column. For position with more than 70% gaps a default conservation of 0 is reported.

The web server is implemented using the Python Django application server (version 2.2.5), Apache (version 2) and Postgresql (version 11). The user interface is designed using Bootstrap (version 4), DataTable (version 1.10.22), the neXtProt feature viewer (version 1.0, https://github.com/calipho-sib/feature-viewer) and custom JavaScript-based validators for input data.

3. Results

3.1. Performance of the solvent accessibility DeepREx prediction

3.1.1. Cross-validation and blind test performance

DeepREx performance is scored using a 10-fold cross-validation procedure on our training dataset comprising 2332 proteins sequences and a blind set with 200 protein sequences, compiled to be non-redundant with respect to our training dataset. Results are reported in Table 1. DeepREx is quite robust, achieving similar performances in the two validation procedures. Overall, our method discriminates buried from exposed residues with 82% accuracy, 82% F1 and 0.63 MCC.

We further compared DeepREx with two recent state-of-the-art tools, both based on deep-learning approaches: PaleAle5 [10] and NetSurfP-2.0 [11]. PaleAle5, predicts exposure into 4 classes: E (exposed), e (partially exposed), b (partially buried) and B (buried). The threshold used by PaleAle5 authors to separate exposed (either E or e) from buried (either b or B) residues is 25% RSA, very close to the threshold adopted in this work. NetSurfP-2.0 directly predicts

Table 1
DeepREx performance in a 10-fold cross-validation and on the blind test set.

Scoring index	Cross-validation	Blind test
Precision	0.820 ± 0.002	0.82
Recall	0.800 ± 0.001	0.80
F1	0.810 ± 0.001	0.82
Q2	0.810 ± 0.001	0.82
MCC	0.620 ± 0.002	0.63

For index definition see section 2.1.5.

RSA real values: in this case we used our 20% RSA threshold to transform these values into a binary classification.

Comparative results on the blind test and on the CASP14 dataset are reported in Table 2. We should remark that the blind test set may not be blind for the other methods. Remarkably, all methods achieve a similar performance on both testing sets. DeepREx reports the most balanced results in the blind test set, as shown by the close values of precision and recall. When tested on the CASP14 dataset comprising 9 free-modelling targets, performances of all methods drop to lower values. The 9 targets are difficult to predict since they do belong to the free-modelling CASP category, without or with very few homologous in the data base. Nonetheless, the three approaches seem to have very close performances, as highlighted by the only small differences in the MCC values.

The three methods (DeepREx, PaleAle5 and NetSurfP-2) are all based on similar neural network architectures involving LSTMs and/or convolutional layers. Among the three, DeepREx adopts the simplest architecture, with only three cascading BiLSTM layers. This ensures the lowest number of parameters for the resulting model without affecting prediction performances that are comparable among the three approaches.

Differently from the other two methods, our DeepREx predictor has been trained on functional monomeric protein chains. This allows to properly define solvent exposure in physiological conditions and to avoid the introduction of biases in solvent exposure computation due to conformational changes at the interfaces upon protein complex formation. However, training only on monomers does not limit the adoption of our model for predicting solvent exposure of multimeric protein chains. To prove this, we performed an additional experiment testing DeepREx on a set of 984 multimeric protein chains extracted from the PaleAle5 independent dataset [10]. In this test, we registered only a slight decrease in the accuracy. The performances of both methods are listed in Table 2S (Supplementary Materials). This suggests that the exclusion of multimeric chains from our training dataset has a very limited impact on the overall performance of DeepREx.

Finally, a reliability index (RI) can be associated to each prediction by applying Eq. (1). RI close to 0 indicates a prediction output close to 0.5 while RI close to 1 indicates that the output is close to 0 (buried) or 1 (exposed). We performed tests to assess whether the RI value can be adopted to discriminate accurate from poor predictions. Results are reported in Supplementary Table 3S and indicate that the higher the RI value the most accurate is the prediction. Notably, most predictions have RI values higher than 0.6. Predictions with low RI values (<0.2) mostly pertain to proteins with very few sequences in the corresponding MSA and, therefore, with a poor input information.

3.2. The web server: DeepREx-WS

DeepREx-WS is available at https://deeprex.biocomp.unibo.it. The server input interface accepts a single sequence in FASTA format with length ranging between 50 and 5000 residues. Upon submission the user is redirected to the page where results will be available after job completion. This page automatically refreshes every 60 s and shows to the user the current status of the job (queued or running). The server also provides the user with a universal job identifier, which can be thereafter used to retrieve job results. The result page (Fig. 2) provides information about the job, including i) the identifier, ii) submission and completion time, iii) protein ID, iv) protein length and v) counts of buried and exposed predictions. After that, the output of the predictor is shown using an interactive viewer along the submitted protein sequence as well as in tabular format.

The following information is reported both in track and tabular form:

i) DeepREx output as two-class prediction of solvent exposure (E = exposed, B = buried).
ii) The RI associated to the DeepREx prediction.
iii) The Kyte-Doolittle hydropathy score [42], averaged over a window of five residues.
iv) The conservation index computed as in Equation (2).
v) The three-class prediction of secondary structure by PYTHIA [26].
vi) The five-class flexibility prediction provided by MEDUSA (0 = rigid, 4 = flexible) [25].
vii) The two-class prediction of intrinsically disordered regions provided by MobiDB-lite3.0 (S = structured, D = disordered) [27].

The feature viewer allows to navigate the sequence, visualizing the different predicted features along it. The user can zoom to specific regions and export a picture of the current visualization in PNG format.

Tabular data can be sorted according to any one of the reported outputs. Moreover, users can activate and combine filters for residue type, exposed or buried positions, reliability index, conservation index, flexibility level, secondary structure and disordered regions.

All results can be downloaded in Tab-Separated Values (TSV) format. If one or more filters are active, the downloaded TSV will report only results for selected residues.

3.3. DeepREx-WS output features

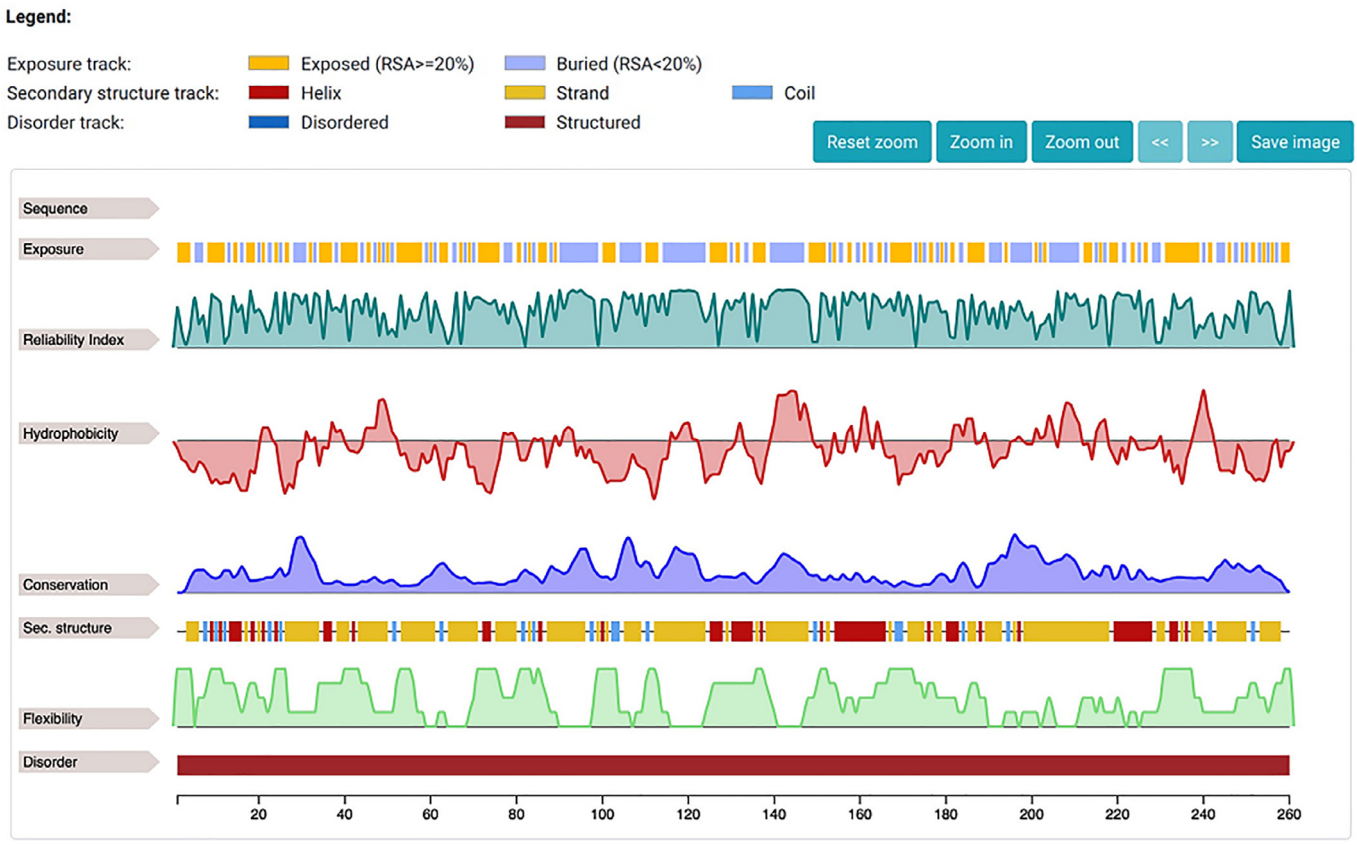
In this section we analyze the relation between solvent exposure and other features included in the DeepREx-WS output, comprising, as detailed above, hydrophobicity (Kyte-Doolittle), conservation index from MSA, flexibility (MEDUSA [25]), secondary structure (PYTHIA [26]) and disorder (MobiDB-Lite3.0 [27]).

All the correlation analyses (except for protein disorder) were performed on the 200 protein sequences included in our blind test (Table 3). Overall, the 200 proteins contain 56,206 residues, 29,068 and 27,138 of which are buried and exposed, respectively, in their experimental 3D structure. On this set DeepREx performs quite well, achieving a prediction accuracy of 82% and a MCC of 0.63

Table 2
Comparison of DeepREx and other protein solvent accessibility predictors on the blind test set and CASP14 targets.

Method	Dataset	Precision	Recall	F1	Q2	MCC
DeepREx	BlindTest	0.82	0.80	0.82	0.82	0.63
PaleAle5.0 [10]	BlindTest	0.78	0.85	0.82	0.82	0.65
NetSurfP2.0 [11]	BlindTest	0.92	0.77	0.82	0.83	0.66
DeepREx	CASP14	0.87	0.76	0.81	0.79	0.57
PaleAle5.0 [10]	CASP14	0.90	0.72	0.80	0.78	0.58
NetSurfP2.0 [11]	CASP14	0.81	0.89	0.85	0.81	0.59

For index definition see section 2.1.5.



Position	Residue	Exposure	Reliability Index	Hydrophobicity	Conservation	Flexibility	Disorder	Secondary structure
1	M	Exposed	0.7	-0.42	0.0	4	S	-
2	S	Exposed	0.28	-1.06	0.0	4	S	-
3	H	Exposed	0.05	-1.24	0.04369	4	S	E
4	H	Exposed	0.33	-1.7	0.16277	4	S	E
5	W	Buried	0.86	-1.8	0.24434	0	S	E
6	G	Buried	0.41	-1.24	0.25717	2	S	E
7	Y	Buried	0.55	-1.38	0.25732	2	S	C
8	G	Exposed	0.1	-1.84	0.21627	3	S	C
9	K	Exposed	0.92	-2.46	0.17615	4	S	H
10	H	Exposed	0.86	-2.28	0.16197	4	S	C
11	N	Exposed	0.78	-2.52	0.18599	4	S	H
12	G	Exposed	0.08	-2.44	0.2511	3	S	C
13	P	Buried	0.22	-2.44	0.25001	3	S	H
14	E	Exposed	0.89	-1.92	0.1766	3	S	H
15	H	Exposed	0.62	-2.48	0.21121	2	S	H
16	W	Buried	0.73	-2.94	0.29127	1	S	H
17	H	Exposed	0.26	-2.94	0.21777	2	S	E

Fig. 2. A screenshot of the DeepREx-WS result page.

(Table 2). The 200 proteins have a negligible disorder content according to MobiDB (less than 1%).

For the evaluation of the correlation between exposure and disorder we collected a dataset of 88 human proteins extracted from the DisProt database [43] and endowed with a disorder content ranging from 10% to 30%. We only compute correlation with respect to predicted exposure, since for disordered regions which, by definition, lack PDB structures, we cannot compute real solvent accessibility.

For what concerns secondary structure predictions, we report three different correlations between exposure and alpha-helix, beta-strand and coil predicted content, respectively.

Table 3
Pairwise Pearson's Correlation Coefficients (PCC) between predicted solvent exposure and the other features.

Feature	PCC with real solvent exposure [(a)]	PCC with predicted solvent exposure [(a)]
Flexibility (MEDUSA [25])	0.56±0.06	0.58±0.08
Alpha-helix (PYTHIA [26])	−0.10±0.10	−0.11±0.11
Beta-strand (PYTHIA [26])	−0.20±0.10	−0.21±0.10
Coil (PYTHIA [26])	0.24±0.08	0.25±0.08
Conservation from MSA	−0.37±0.11	−0.39±0.11
Hydrophobicity (Kyte-Doolittle [42])	−0.23±0.09	−0.24±0.10
Disorder (MobiDB-Lite3.0 [27]) [(b)]	–	0.27±0.11

[(a)] Average PCC computed per-protein and associated Standard Deviation values.
[(b)] Correlation computed on 88 proteins from DisProt [43] with disorder content ranging from 10% to 30%.

All correlation results are shown in Table 3 and are calculated per protein and then averaged.

Residue flexibility as predicted by MEDUSA well correlates with both real and predicted solvent exposure values (in Table 3, first line, average PCCs are 0.58 and 0.56, respectively). This can be partially explained by considering that MEDUSA adopts crystallographic B-factors as proxies for residue flexibility, and that these values tend to be higher at the protein surface. However, the correlation is not perfect, suggesting that the two features (i.e., residue solvent accessibility and flexibility) provide complementary information which can be profitably merged for a better understanding of residue structural properties from sequence.

Average correlation coefficients between exposure and helix and strand motifs are negative and close to 0, considering the significant deviations from the mean (in Table 3, second and third lines, respectively). This may indicate that exposed residues (both real and predicted) are not preferentially placed in helix or strand regions. Correlations with coils are slightly positive (in Table 3, fourth line), suggesting a weak propensity of exposed residues for coil regions.

Exposed residues (either real or predicted) tend to be localized in non-conserved positions, as highlighted by moderate anti-correlation reported in Table 3 between predicted and real solvent accessibility and conservation index (fifth line, average PCCs are −0.39 and −0.37, respectively). Moreover, as expected, solvent exposure anti-correlates with respect to hydrophobicity (in Table 3, sixth line, average PCCs are −0.24 and −0.23). Again, these results suggest that solvent accessibility cannot be completely explained by conservation or residue hydrophobicity alone, justifying the integration/combination of the different features for residue structural/functional characterization.

Finally, a modest correlation (PCC = 0.26) of exposure is also observed with protein disorder on a dataset of 88 proteins extracted from DisProt [43]. This may indicate a slight propensity of disordered regions for exposed positions.

Although the size of our protein sets is limited, the results presented in this section suggest that protein solvent exposure positively correlates with protein flexibility and negatively correlates with hydrophobicity and conservation. In general, all these features provide complementary information on residues and can be then combined to characterize proteins from a structural and functional point of view. This can be useful in many contexts such as protein surface engineering, where one looks for residues placed at the protein surface to be selected as candidate for site-specific mutagenesis. Routinely, selected positions are exposed residues characterized by low conservation indexes (in order to avoid functionally important sites) and placed in flexible loops. Starting from protein sequence, the combination of predicted exposure, flexibility and conservation can be helpful to reduce the search space in protein surface engineering. For instance, in our dataset of 200 proteins, selecting residues predicted as exposed, having a low conservation index (residue conservation lower than the median for each protein) and flexible (MEDUSA value ≥ 3) we obtain 12,068 residues, representing 21% of the total number of residues. This allows to significantly restrict the search space of candidate positions for surface engineering particularly when 3D structure is lacking.

3.4. Case study: DeepREx-WS to assist surface engineering

In this section, we benchmark DeepREx-WS in the context of protein surface charge engineering with an example. Surface charge engineering is particularly important for the industrial use of biocatalyst. Recently, much attention has been focused on halophilic enzymes that can be adopted in hypersaline environments (e.g., brines, ionic liquids or ionic detergents) [21]. Putative enzymes for the use in high-salt conditions have been traditionally identified among those available in natural systems. An alternative approach consists in the induction of halotolerance into an existing biocatalyst possessing the required features in terms of catalytic activity. Following this trend, in a recent study [21], authors considered the bovine carbonic anhydrase II (bCAII, UniProtKB: P00921) for the rational design of halotolerance by protein surface engineering. Specifically, in order to enhance bCAII halotolerance, authors adopted one of the possible mechanisms present in natural halophilic enzymes: the increase of the abundance of acidic residues in the protein surface. By this, 18 positions were identified and mutated into negative residues, after a rational choice procedure based on the available PDB bCAII structure (1V9E). The selection of positions to be mutated is not exhaustive and integrates considerations on solvent accessibility and/or side-chain steric bulks, and on the residue conservation in a multiple sequence alignment generated using 50 homologous sequences. The availability of the three-dimensional structure provides a large amount of information. However, what if the structure is not available as it is for many proteins? DeepREx-WS can assist the choice of residues to be mutated without the help of the structure. We submitted the 260-residue long sequence of the bCAII to the server and filtered the results to select possible positions for mutation into negative residues (Glutamic or Aspartic acid). Remarkably, the exposure prediction reaches a high MCC value (0.81). Mimicking the rational procedure described in [21] and considering the DeepREx-WS output for the whole protein sequence, we can select residues predicted as exposed, obtaining 139 positions, 112 of which are different from Glutamic or Aspartic acid, and then reducing the search space to 43% of the protein residues. All the 18 positions from [21] are included in this set. If we add a filter on protein conservation, selecting only lowly conserved residues (CI lower than the median on the protein equal to 0.2), we can further restrict to 78 possible target positions (30% of the sequence). Out of the 18 positions considered in [21], 13 are included in the set of 78 positions selected. Five out of 18 positions are not retained in our selection. Two of them (G8 and N24) have a conservation index (0.22) only slightly higher than the threshold used here (0.20). The remaining 3 positions (N62, N252 and Q254) are weakly variable in the MSA used in [21] and their selection in the study does not take into consideration conservation.

If exposed positions are intersected with most flexible ones (MEDUSA score equal to 3 or 4), 66 positions are selected, corresponding to 25% of the sequence. This set contains 12 out of 18 positions selected in [21]. Out of the 6 not included positions, 3 are predicted with a medium flexibility level (MEDUSA score equal to 2) and 3 are predicted with limited flexibility (MEDUSA score 1). Remarkably, none of them are predicted as rigid (MEDUSA score 0).

In Table 4 we report the complete output of DeepREx-WS for the 18 positions of interest reported in [20]. Interestingly, all the

Table 4
Analysis of relevant positions of the bovine carbonic anhydrase II protein (UniProtKB:P00921) reported in [21] with the DeepREx-WSs.

Pos	Res	SE[(a)]	RI[(b)]	HP[(c)]	CI[(d)]	Flexibility[(e)]	Disorder[(f)]
8	G	E	0.10	−1.84	0.22	3	S
18	K	E	0.98	−1.74	0.13	3	S
24	N	E	0.95	−0.22	0.22	4	S
36	K	E	0.98	−0.42	0.10	3	S
39	V	E	0.62	0.64	0.12	3	S
50	V	E	0.23	1.62	0.13	1	S
57	R	E	0.48	−1.72	0.09	1	S
62	N	E	0.54	−1.28	0.32	1	S
74	Q	E	0.76	−3.04	0.12	4	S
85	T	E	0.95	0.08	0.16	4	S
136	Q	E	0.11	−2.06	0.11	4	S
169	K	E	0.96	−2.56	0.1	4	S
177	N	E	0.78	−0.6	0.13	3	S
186	N	E	0.91	1.34	0.14	3	S
220	Q	E	0.95	−1.34	0.18	2	S
238	L	E	0.17	0.88	0.16	2	S
252	N	E	0.94	−2.32	0.28	3	S
254	Q	E	0.75	−2.36	0.25	2	S

(a) SE = Solvent Exposure, as predicted by DeepREx. E = Exposed, B = Buried.
(b) RI = DeepREx Reliability Index, as defined in Eq. (1).
(c) HP = Kyte-Doolittle Hydrophobicity [42].
(d) CI = Conservation Index, computed as in Eq. (2).
(e) Flexibility value, as predicted by Medusa [25]. It goes from 0 (rigid) to 4 (highly flexible).
(f) Disorder annotation as retrieved from MobiDB-Lite3.0 [27]. S = Structured, D = Disordered.

positions are correctly predicted as exposed, most of them with high reliability. Moreover, they are all characterized by a low conservation index (between 0.09 and 0.32), while most of them (12 out of 18) are predicted as localized in flexible regions (MEDUSA ≥ 3). Altogether, these features are in line with those required by the rational design performed in [21] and show that the DeepREx-WS prediction can reconstruct them starting from the protein sequence alone.

4. Conclusion

In this paper, we develop DeepREx, a novel deep-learning based tool for annotating residue solvent exposure into two classes (buried and exposed). DeepREx performance is evaluated on a blind dataset comprising 200 proteins and on a selected set of difficult targets from CASP14. Results show that DeepREx is competitive with other tools at the state-of-the-art. The method is made available as a web server (DeepREx-WS) and as a standalone tool, including a containerized version. This makes DeepREx well-suited for applications on large datasets and for easy integration into higher-level workflows. The web server which integrates the predictor of solvent accessibility (DeepREx-WS) is implemented to allow the intersection of DeepREx outputs with other protein features such as residue flexibility, conservation, hydrophobicity and inclusion in intrinsically disordered regions. Our results on 200 proteins indicate that solvent accessibility well correlates with flexibility and negatively correlates with conservation and hydrophobicity. Disorder is apparently negligible for this analysis. Furthermore, with the example of the bovine carbonic anhydrase II [21] and comparing with residue selection done directly on the protein structure, we confirm that the integration of the server outputs can profitably allow a primary selection of candidate positions for surface residue modification starting from the protein sequence alone. We propose our web server to highlight likely positions in protein sequence for surface engineering and as a valuable alternative when protein structure is not or partially available.

5. Data and method availability

The DeepREx web server and datasets are available at https://deeprex.biocomp.unibo.it.

The DeepREx standalone tool Python source code is available at https://github.com/BolognaBiocomp/deeprex. The program has been tested with Python version 3.8. External dependencies include the Biopython package (tested version 1.78), the Keras (tested version 2.4.3) deep-learning library as well as a working installation of the HHsuite (tested version 3.3.0) for multiple sequence alignment building.

DeepREx has been also released as a Docker container available at https://hub.docker.com/r/bolognabiocomp/deeprex. In both cases, the program takes in input: i) a FASTA file containing one or more sequences; ii) a valid sequence database for HHblits alignments; iii) a file name where an output TSV file will be written after termination.

Funding

The work was supported by the PRIN2017 grant (project 2017483NH8_002), delivered to CS from the Italian Ministry of University and Research.

CRediT authorship contribution statement

Matteo Manfredi: Methodology, Software, Validation. **Castrense Savojardo:** Conceptualization, Data curation. **Pier Luigi Martelli:** Conceptualization, Supervision. **Rita Casadio:** Conceptualization, Supervision.

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.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2021.10.016.

References

[1] Miller S, Lesk AM, Janin J, Chothia C. The accessible surface area and stability of oligomeric proteins. Nature 1987;328(6133):834–6. https://doi.org/10.1038/328834a0.

[2] Rost B, Sander C. Conservation and prediction of solvent accessibility in protein families. Proteins 1994;20(3):216–26. https://doi.org/10.1002/prot.340200303.

[3] Tien MZ, Meyer AG, Sydykova DK, Spielman SJ, Wilke CO, Porollo A. Maximum allowed solvent accessibilites of residues in proteins. PLoS ONE 2013;8(11):e80635. https://doi.org/10.1371/journal.pone.0080635.

[4] Rose G, Geselowitz A, Lesser G, Lee R, Zehfus M. Hydrophobicity of amino acid residues in globular proteins. Science 1985;229(4716):834–8. https://doi.org/10.1126/science.4023714.

[5] Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22(12):2577–637. https://doi.org/10.1002/bip.360221211.

[6] Mihel J, Šikić M, Tomić S, Jeren B, Vlahoviček K. PSAIA – Protein Structure and Interaction Analyzer. BMC Struct Biol 2008;8(1):21. https://doi.org/10.1186/1472-6807-8-21.

[7] Adamczak R, Porollo A, Meller J. Combining prediction of secondary structure and solvent accessibility in proteins. Proteins 2005;59(3):467–75. https://doi.org/10.1002/prot.20441.

[8] Magnan CN, Baldi P. SSpro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. Bioinformatics 2014;30(18):2592–7. https://doi.org/10.1093/bioinformatics/btu352.

[9] Pollastri G, Martin AJ, Mooney C, Vullo A. Accurate prediction of protein secondary structure and solvent accessibility by consensus combiners of sequence and structure information. BMC Bioinf 2007;8(1):201. https://doi.org/10.1186/1471-2105-8-201.

[10] Kaleel M, Torrisi M, Mooney C, Pollastri G. PaleAle 5.0: prediction of protein relative solvent accessibility by deep learning. Amino Acids 2019;51(9):1289–96. https://doi.org/10.1007/s00726-019-02767-6.

[11] Klausen MS, Jespersen MC, Nielsen H, Jensen KK, Jurtz VI, Sønderby CK, et al. NetSurfP-2.0: improved prediction of protein structural features by integrated deep learning. Proteins Struct Funct Bioinf 2019;87(6):520–7. https://doi.org/10.1002/prot.v87.610.1002/prot.25674.

[12] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, et al. (2021) Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583-589. doi: 10.1038/s41586-021-03819-2. Epub 2021 Jul 15. PMID: 34265844; PMCID: PMC8371605.

[13] Savojardo C, Fariselli P, Martelli PL, Casadio R. ISPRED4: interaction sites PREDiction in protein structures with a refining grammar model. Bioinformatics 2017;33(11):1656–63. https://doi.org/10.1093/bioinformatics/btx044.

[14] Porollo A, Meller J. Prediction-based fingerprints of protein-protein interactions. Proteins 2007;66(3):630–45. https://doi.org/10.1002/prot.21248.

[15] Savojardo C, Manfredi M, Martelli PL, Casadio R. Solvent accessibility of residues undergoing pathogenic variations in humans: from protein structures to protein sequences. Front Mol Biosci 2021;7:. https://doi.org/10.3389/fmolb.2020.626363626363.

[16] Martelli PL, Fariselli P, Savojardo C, Babbi G, Aggazio F, et al. Large scale analysis of protein stability in OMIM disease related human protein variants. BMC Genomics 2016;17(S2):397. https://doi.org/10.1186/s12864-016-2726-v.

[17] Savojardo C, Babbi G, Martelli PL, Casadio R. Functional and structural features of disease-related protein variants. IJMS 2019;20(7):1530. https://doi.org/10.3390/ijms20071530.

[18] Pedersen JN, Zhou Ye, Guo Z, Pérez B. Genetic and chemical approaches for surface charge engineering of enzymes and their applicability in biocatalysis: A review. Biotechnol Bioeng 2019;116(7):1795–812. https://doi.org/10.1002/bit.v116.710.1002/bit.26979.

[19] Shivange AV, Hoeffken HW, Haefner S, Schwaneberg U. Protein consensus-based surface engineering (ProCoS): a computer-assisted method for directed protein evolution. Biotechniques 2016;61(6):305–14. https://doi.org/10.2144/000114483.

[20] Simeonov P, Berger-Hoffmann R, Hoffmann R, Strater N, Zuchner T. Surface supercharged human enteropeptidase light chain shows improved solubility and refolding yield. Protein Eng Des Sel 2011;24(3):261–8. https://doi.org/10.1093/protein/gzq104.

[21] Warden AC, Williams M, Peat TS, Seabrook SA, Newman J, Dojchinov G, et al. Rational engineering of a mesohalophilic carbonic anhydrase to an extreme halotolerant biocatalyst. Nat Commun 2015;6(1). https://doi.org/10.1038/ncomms10278.

[22] Qi Y, Chilkoti A. Protein-polymer conjugation-moving beyond PEGylation. Curr Opin Chem Biol 2015;28:181–93. https://doi.org/10.1016/j.cbpa.2015.08.009.

[23] Turunen O, Vuorio M, Fenel F, Leisola M. Engineering of multiple arginines into the Ser/Thr surface of Trichoderma reesei endo-1,4-beta-xylanase II increases the thermotolerance and shifts the pH optimum towards alkaline pH. Protein Eng 2002;15:141–5. https://doi.org/10.1093/protein/15.2.141.

[24] Takagi H, Hirai K, Maeda Y, Matsuzawa H, Nakamori S. Engineering subtilisin E for enhanced stability and activity in polar organic solvents. J Biochem 2000;127:617–25. https://doi.org/10.1093/oxfordjournals.jbchem.a022649. PMID: 10739954.

[25] Meersche YV, Cretin G, de Brevern AG, Gelly J-C, Galochkina T. MEDUSA: prediction of protein flexibility from sequence ISSN 0022-2836. J Mol Biol 2021;166882. https://doi.org/10.1016/j.jmb.2021.166882.

[26] Cretin G, Galochkina T, de Brevern AG, Gelly JC. (2021) PYTHIA: Deep learning approach for local protein conformation prediction. Int J Mol Sci., 22(16), 8831. Published 2021 Aug 17. doi:10.3390/ijms22168831

[27] Necci M, Piovesan D, Clementel D, Dosztányi Z, Tosatto SCE (2020) MobiDB-lite 3.0: fast consensus annotation of intrinsic disorder flavors in proteins, Bioinformatics, https://doi.org/10.1093/bioinformatics/btaa1045

[28] Berman HM. The Protein Data Bank. Nucleic Acids Res 2000;28(1):235–42. https://doi.org/10.1093/nar/28.1.235.

[29] UniProt Consortium. UniProt: A worldwide hub of protein knowledge. Nucleic Acids Res 2019;47(D1):D506–15. https://doi.org/10.1093/nar/gky1049.

[30] Dana JM, Gutmanas A, Tyagi N, Qi G, O'Donovan C, et al. (2019) 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 (D1), D482–D489. https://doi.org/10.1093/nar/gky1114.

[31] Lomize MA, Pogozheva ID, Joo H, Mosberg HI, Lomize AL. OPM database and PPM web server: resources for positioning of proteins in membranes. Nucleic Acids Res 2012;40(D1):D370–6. https://doi.org/10.1093/nar/gkr703.

[32] Fox NK, Brenner SE, Chandonia JM. SCOPe: Structural Classification of Proteins–extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Res 2014;42(Database issue):D304–9. https://doi.org/10.1093/nar/gkt1240.

[33] Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol 2017;35(11):1026–8. https://doi.org/10.1038/nbt.3988. Epub 2017 Oct 16 PMID: 29035372.

[34] Sillitoe I, Bordin N, Dawson N, Waman VP, Ashford P, Scholes HM, et al. CATH: increased structural coverage of functional space. Nucleic Acids Res, 2021;49(D1):D266-D273. doi: 10.1093/nar/gkaa1079.

[35] Steinegger M, Meier M, Mirdita M, Vöhringer H, Haunsberger SJ, et al. HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinf 2019;20(1):473. https://doi.org/10.1186/s12859-019-3019-7.

[36] Mirdita M, von den Driesch L, Galiez C, Martin MJ, Söding J, et al. Uniclust databases of clustered and deeply annotated protein sequences and alignments. Nucleic Acids Res 2017;45(D1):D170–6. https://doi.org/10.1093/nar/gkw1081.

[37] Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks 2005;18(5–6):602–10. https://doi.org/10.1016/j.neunet.2005.06.042.

[38] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;9(8):1735–80. https://doi.org/10.1162/neco.1997.9.8.1735.

[39] Pascanu R, Mikolov T, Bengio Y (2013) On the difficulty of training recurrent neural networks. arXiv:1211.5063 [cs].

[40] Chollet F (2015) Keras; GitHub.

[41] Kingma DP, Ba J (2017) Adam: A method for stochastic optimization. arXiv:1412.6980 [cs].

[42] Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol 1982;157(1):105–32.

[43] Hatos A, Hajdu-Soltész B, Monzon AM, Palopoli N, Álvarez L, et al. DisProt: intrinsic protein disorder annotation in 2020. Nucleic Acids Res 2020;48(D1):D269–76. https://doi.org/10.1093/nar/gkz975. PMID: 31713636; PMCID: PMC7145575.