Computational and Structural Biotechnology Journal xxx (xxxx) xxx

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

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

Review

Deep learning methods in protein structure prediction

Mirko Torrisi [b], Gianluca Pollastri [b], Quan Le [a,*]

[a] Centre for Applied Data Analytics Research, University College Dublin, Ireland
[b] School of Computer Science, University College Dublin, Ireland

ARTICLE INFO

Article history:
Received 15 October 2019
Received in revised form 19 December 2019
Accepted 20 December 2019
Available online xxxx

Keywords:
Deep learning
Protein structure prediction
Machine learning

ABSTRACT

Protein Structure Prediction is a central topic in Structural Bioinformatics. Since the '60s statistical methods, followed by increasingly complex Machine Learning and recently Deep Learning methods, have been employed to predict protein structural information at various levels of detail. In this review, we briefly introduce the problem of protein structure prediction and essential elements of Deep Learning (such as Convolutional Neural Networks, Recurrent Neural Networks and basic feed-forward Neural Networks they are founded on), after which we discuss the evolution of predictive methods for one-dimensional and two-dimensional Protein Structure Annotations, from the simple statistical methods of the early days, to the computationally intensive highly-sophisticated Deep Learning algorithms of the last decade. In the process, we review the growth of the databases these algorithms are based on, and how this has impacted our ability to leverage knowledge about evolution and co-evolution to achieve improved predictions. We conclude this review outlining the current role of Deep Learning techniques within the wider pipelines to predict protein structures and trying to anticipate what challenges and opportunities may arise next.

© 2020 The Authors. Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Contents

1. Introduction

Proteins hold a unique position in Structural Bioinformatics. In fact, the origins of the field itself can be traced to Max Perutz and John Kendrew's pioneering work to determine the structure of globular proteins (which also led to the 1962 Nobel Prize in Chemistry) [1,2]. The ultimate goal of Structural Bioinformatics, when it comes to proteins, is to unearth the relationship between the residues forming a protein and its function, i.e., in essence, the relationship between genotype and phenotype. The ability to disentangle this relationship can potentially be used to identify, or even design, proteins able to bind specific targets [3], catalyse novel reactions [4] or guide advances in biology, biotechnology

* Corresponding author.
E-mail address: quan.le@ucd.ie (Q. Le).

https://doi.org/10.1016/j.csbj.2019.12.011
2001-0370/© 2020 The Authors. Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology.
This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Please cite this article as: M. Torrisi, G. Pollastri and Q. Le, Deep learning methods in protein structure prediction, Computational and Structural Biotechnology Journal, https://doi.org/10.1016/j.csbj.2019.12.011

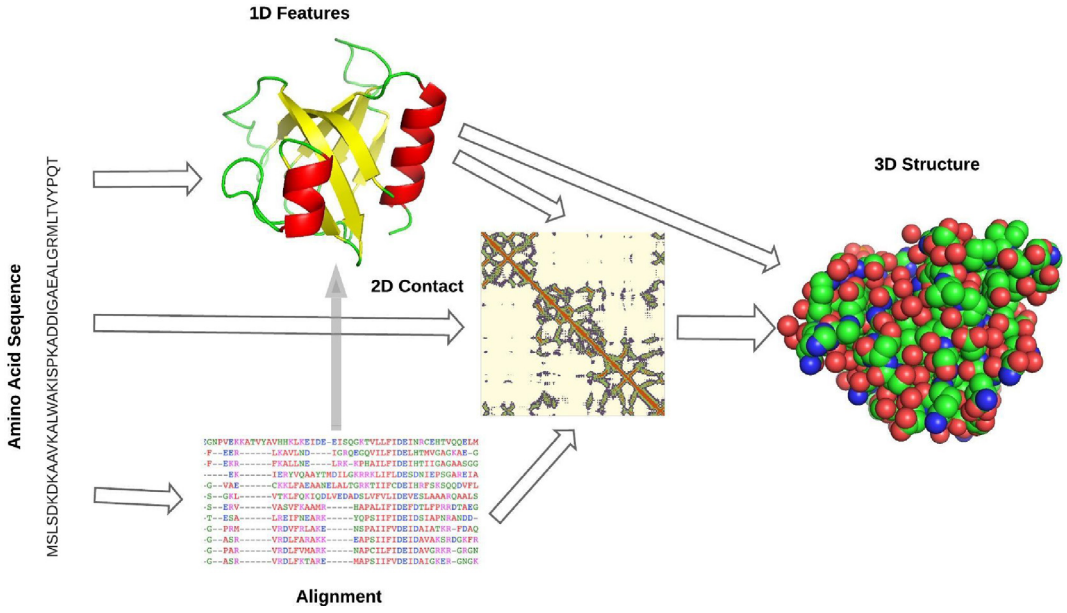

Fig. 1. A generic pipeline for ab initio Protein Structure Prediction, in which evolutionary information in the form of alignments, 1D and 2D PSA are intermediate steps.

and medicine [5], e.g. editing specific locations of the genome with CRISPR-Cas9 [6].

According to Anfinsen's thermodynamic hypothesis, all the information that governs how proteins fold is contained in their respective primary sequences, i.e. the chains of amino acids (AA, also called residues) forming the proteins [7,8]. Anfinsen's hypothesis led to the development of computer simulations to score protein conformations, and, thus, search through potential states looking for that with the lowest free energy, i.e. the native state [9,8]. The key issue with this energy-driven approach is the explosion of the conformational search space size as a function of a protein's chain length. A solution to this problem consists in the exploitation of simpler, typically coarser, abstractions to gradually guide the search, as proteins appear to fold locally and non-locally at the same time but incrementally forming more complex shapes [10].

A standard pipeline for Protein Structure Prediction envisages intermediate prediction steps where abstractions are inferred which are simpler than the full, detailed 3D structure, yet structurally informative - what we call Protein Structure Annotations (PSA) [11]. The most commonly adopted PSA are secondary structure, solvent accessibility and contact maps. The former two are one-dimensional (1D) abstractions which describe the arrangement of the protein backbone, while the latter is a two-dimensional (2D) projection of the protein tertiary structure in which any 2 AA in a protein are labelled by their spatial distance, quantised in some way (e.g. greater or smaller than a given distance threshold). Several other PSA, e.g. torsion angles or contact density, and variations of the aforementioned ones, e.g. half-sphere exposure and distance maps, have been developed to describe protein structures [11]. Fig. 1 depicts a pipeline for the prediction of protein structure from the sequence in which the intermediate role of 1D and 2D PSA is highlighted.

It should be noted that protein intrinsic disorder [12–14] can be regarded as a further 1D PSA with an important structural and functional role [15], which has been predicted by Machine Learning and increasingly Deep Learning methods similar to those adopted for the prediction of other 1D PSA properties [16–22], sometimes alongside them [23]. However, given its role in protein structure prediction pipelines is less clear than for other PSA, we will not explicitly focus on disorder in this article and refer the reader to specialised reviews on disorder prediction, e.g. [24–26].

The slow but steady growth in the number of protein structures available at atomic resolution has led to the development of PSA predictors relying also on homology detection (“template-based predictors”), i.e. predictors directly exploiting proteins of known structure (“templates”) that are considered to be structurally similar based on sequence identity [27–30]. However, a majority PSA predictors are “ab initio”, that is, they do not rely on templates. Ab-initio predictors leverage extensive evolutionary information searches at the sequence level, relying on ever-growing data banks of known sequences and constantly improving algorithms to detect

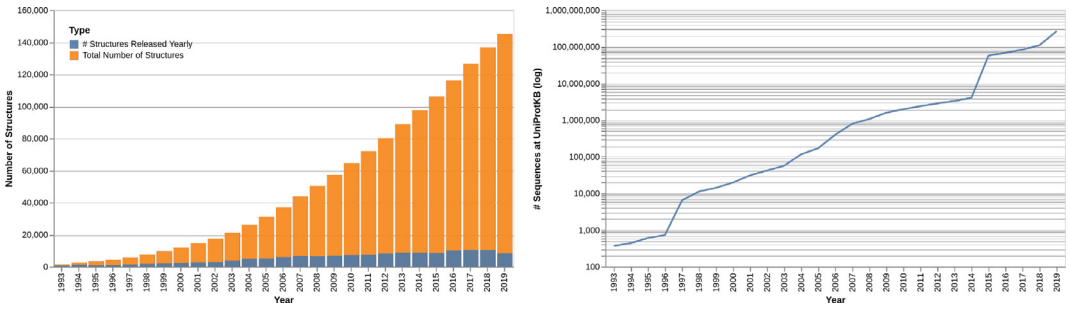

Fig. 2. Growth of known structures in the Protein Data Bank (left) and known sequences in Uniprot (right). The y-axis is shown in logarithmic scale for the Uniprot.

Please cite this article as: M. Torrisi, G. Pollastri and Q. Le, Deep learning methods in protein structure prediction, Computational and Structural Biotechnology Journal, https://doi.org/10.1016/j.csbj.2019.12.011

similarity among them [31–33]. Fig. 2 shows the growth in the number of known structures in the Protein Data Bank (PDB) [34] and sequences in the Uniprot [35] - the difference in pace is evident, with an almost constant number of new structures having been added to the PDB each year for the last few years while the number of known sequences is growing close to exponentially.

1.1. Feed forward neural networks

A Feed Forward Neural Network (FFNN) is an artificial neural network [36] containing no cycles. In particular layered FFNN are FFNN whose nodes can be partitioned into groups (layers) that are ordered and in which the outputs of layer i are inputs to and only to layer $i+1$. The first layer is known as Input layer, the last as Output layer and any layer in between is a Hidden layer whose units form an intermediate representation of an instance. Layered FFNN, which may be trained from examples using the back-propagation algorithm [36] and have been proven to have universal approximation properties [37], have been used to predict 1D PSA since the '80s [38–40]. These networks have typically been used in their so-called "windowed" version, in which each segment of a fixed number of amino acids in a sequence is treated as the input for a separate example, the target for the segment being the PSA of interest for one of the amino acids in the segment (usually the central one).

1.2. Deep Learning

Deep Learning [41] is a sub-field of Machine Learning based on artificial neural networks, which emphasises the use of multiple connected layers to transform inputs into features amenable to predict corresponding outputs. Given a sufficiently large dataset of input–output pairs, a training algorithm can be used to automatically learn the mapping from inputs to outputs by tuning a set of parameters at each layer in the network.

While in many cases the elementary building blocks of a Deep Learning system are FFNN or similar elementary cells, these are combined into deep stacks using various patterns of connectivity. This architectural flexibility allows Deep Learning models to be customised for any particular type of data. Deep Learning models can generally be trained on examples by back-propagation [36], which leads to efficient internal representations of the data being learned for a task. This automatic feature learning largely removes the need to do manual feature engineering, a laborious and potentially error-prone process which involves expert domain knowledge and is required in other Machine Learning approaches. However, Deep Learning models easily contain large numbers of internal parameters and are thus data-greedy - the most successful applications of Deep Learning to date have been in fields in which very large numbers of examples are available [41]. In the remainder of this section we summarise the main Deep Learning modules which are used in previous research in Protein Structure Prediction.

Convolutional Neural Networks (CNN) [42] are an architecture designed to process data which is organised with regular spatial dependency (like the tokens in a sequence or the pixels in an image). A CNN layer takes advantage of this regularity by applying the same set of local convolutional filters across positions in the data, thus brings two advantages: it avoids the overfitting problem by having a very small number of weights to tune with respect to the input layer and the next layer dimensionality, and it is translation invariant. A CNN module is usually composed of multiple consecutive CNN layers so that the nodes at later layers have larger receptive fields and can encode more complex features. It should be noted that "windowed" FFNN discussed above can be regarded as a particular, shallow, version of CNN, although we will keep referring to them as FFNN in this review to follow the historical naming practice in the literature.

Recurrent Neural Networks (RNN) [43] are designed to learn global features from sequential data. When processing an input sequence, a RNN module uses an internal state vector to summarise the information from the processed elements of the sequence: it has a parameterised sub-module which takes as inputs the previous internal state vector and the current input element of the sequence to produce the current internal state vector; the final state vector will summarise the whole input sequence. Since the same function is applied repeatedly across the elements of a sequence, RNN modules easily suffer from the gradient vanishing or gradient explosion problem [44] when applying the back propagation algorithm to train them. Gated recurrent neural network modules like Long Short Term Memory (LSTM) [45] or Gated Recurrent Unit (GRU) [46] are designed to alleviate these problems. Bidirectional versions of RNNs (BRNN) are also possible [47] and particularly appropriate in PSA predictions, where data instances are not sequences in time but in space and propagation of contextual information in both directions is desirable.

Even though the depth of a Deep Learning model increases its expressiveness, increasing depth also makes it more difficult to optimise the network weights due to gradients vanishing or exploding. In [48] Residual Networks have been proposed to solve these problems. By adding a skip connection from one layer to the next one, a Residual Network is initialised to be near the identity function thus avoids large multiplicative interactions in the gradient flow. Moreover, skip connections act as "shortcuts", providing shorter input–output paths for the gradient to flow in otherwise deep networks.

2. Methods for 1D Protein Structural Annotations

First generation PSA predictors relied on statistical calculations of propensities of single AA towards structural conformations, usually secondary structures [49–52], which were then combined into actual predictions via hand-crafted rules. While these methods predicted at better than chance accuracy, they were quite limited - especially on novel protein structures [53], with per-AA accuracies usually not exceeding 60%.

In a second generation of predictors [54], information from more than one AA at a time was fed to various methods, including FFNN to predict secondary structure [38,39], and least squares, i.e. a standard regression analysis, to predict hydrophobicity values [55]. This step change was made possible by the increasing number of resolved structures available. These methods were somewhat more accurate than first generation ones, with secondary structure accuracies of 63–64% reported [38].

The third generation of PSA predictors has been characterised by the adoption of evolutionary information [56] in the form of alignments of multiple homologous sequences as input to the predictive systems, which are almost universally Machine Learning, or Deep Learning algorithms. One of the early systems from this generation, PHD [56], arguably the first to predict secondary structure at over 70% accuracy, was implemented as two cascaded FFNN taking segments of 13 AA and 17 secondary structure predictions as inputs, containing 5,000–15,000 free tunable parameters, and trained by back-propagation.

Subsequent sources of improvement were more sensitive tools for mining evolutionary information such as PSI-BLAST [32] or HMMER [57], and the ever increasing nature of both the databases of available structures and sequences, with PSIPRED [58], based on a similar stack of FFNN to that used in PHD, albeit somewhat larger,

Please cite this article as: M. Torrisi, G. Pollastri and Q. Le, Deep learning methods in protein structure prediction, Computational and Structural Biotechnology Journal, https://doi.org/10.1016/j.csbj.2019.12.011

achieving state of the art performances at the time of development, with sustained 76% secondary structure prediction accuracy.

2.1. Deep Learning methods for 1D PSA prediction

Various Deep Learning algorithms have been routinely adopted for PSA prediction since the advent of the third generation of predictors [11], alongside more classic Machine Learning methods such as k-Nearest Neighbors [63,64], Linear Regression [65], Hidden Markov Models [66], Support Vector Machines (SVM) [67] and Support Vector Regression [68].

PHD, PSIPRED, and JPred [69] are among the first notable examples in which cascaded FFNN are used to predict 1D PSA, in particular secondary structure. DESTRUCT [70] expands on this approach by simultaneously predicting secondary structure and torsion angles by an initial FFNN, then having a filtering FFNN map first stage predictions into new predictions, and then iterating, with all copies of the filtering network sharing their internal parameters.

SPIDER2 [59] builds on this approach adding solvent accessibility to the set of features predicted and training an independent set of weights for each iteration. The entire set of PSA predicted is used, along with the input features of the first stage, to feed the second and third stage. Each stage is composed of a window-based ($w = 17$) 3-layered FFNN with 150 hidden units each [59].

SSpro is a secondary structure predictor based on a Bidirectional RNN architecture followed by a 1D CNN stage. The architecture was shown to be able to identify the terminus of the protein sequence and was quite compact with only between 1400 and 2900 free parameters [47]. Subsequent versions of SSpro increased the size of the training datasets and networks [71]. Similar architectures have been implemented to predict solvent accessibility and contact density [72]. The latest version of SSpro adds a final refinement step based on a PSI-BLAST search of structurally similar proteins [30], i.e. is a template-based predictor.

A variant to plain BRNN-CNN architectures are stacks of Recurrent and Convolutional Neural Networks [73,27,74,31,75]. In these a first BRNN-CNN stage is followed by a second structurally similar stage fed with averages over segments of predictions from the first stage. Porter, PaleAle, BrownAle and Porter+ (Brewery) are Deep Learning methods employing these architectures to predict secondary structure, solvent accessibility, contact density and torsion angles, respectively [60,11]. The latest version of Porter (v5) is composed by an ensemble of 7 models with 40,000–60,000 free parameters each, using multiple methods to mine evolutionary information [31,76]. The same architecture has also been trained on a combination of sequence and structural data [27,28], and in a cascaded approach similar to that of DESTRUCT and SPIDER2 in which multiple PSA are predicted at once and the prediction is iterated [77].

SPIDER3 [61] substitutes the FFNN architecture of SPIDER2 with a Bidirectional RNN with LSTM cells [45] followed by a FFNN, predicts 4 PSA at once, and iterates the prediction 4 times. Each of the 4 iterations of SPIDER3 is made of 256 LSTM cells per direction per layer, followed by 1024 and 512 hidden units per layer in the FFNN. Adam optimiser and Dropout (with a ratio of 50%) [78] are used to train the over 1 million free parameters of the model. SPIDER2 and SPIDER3 are the only described methods which employ seven representative physio-chemical properties in input along with both HHblits and PSI-BLAST outputs.

2.2. Convolutional neural networks

RaptorX-Property is a collection of 1D PSA predictors released since 2010 and based on Conditional Neural Fields (CNF), i.e. Neural Networks possessing an output layer made of Conditional Random Fields (CRF) [79]. The most recent version of RaptorX-Property is based on Deep Convolutional Neural Fields (DeepCNF), i.e. CNN with CRF output [80,23]. This version has 5 convolutional layers containing 100 hidden units with a window size of 11 each, i.e. roughly 500,000 free parameters (10 times and 100 times as many as Porter5 and PHD, respectively). The latest version of RaptorX-Property depends on HHblits instead of PSI-BLAST for the evolutionary information fed to DeepCNF models [23].

NetSurfP-2.0 is a recently developed predictor which employs either HHblits or MMsEqs. 2 [76,81], depending on the number of sequences in input [62]. NetSurfP-2.0 is made of two CNN layers, consisting of 32 filters with 129 and 257 units, respectively, and two BRNN layers, consisting of 1024 LSTM cells per direction per layer. The CNN input is fed to the BRNN stage as well. NetSurfP-2.0 predicts secondary structure, solvent accessibility, torsion angles and structural disorder with a different fully connected layer per PSA.

In Fig. 3 we report a scatterplot of performances of secondary structure predictors vs. the year of their release. Gradual, continuing improvements are evident from the plot, as well as the transition from statistical methods to classical Machine Learning and later Deep Learning methods. A set of surveys of recent methods for the prediction of protein secondary structure can be found in [82–85] and a thorough comparative assessment of high-throughput predictors in [86].

3. Methods for 2D Protein Structural Annotations

A typical pipeline to predict protein structure envisages a step in which 2D PSA of some nature are predicted [11]. In fact, most of the recent progress in Protein Structure Prediction has been driven by Deep Learning methods applied to the prediction of contact or distance maps [87,88].

Contact maps have been adopted to reconstruct the full three-dimensional (3D) protein structure since the '90s [89–91]. Although the 2D-3D reconstruction is known to be a NP-hard problem [92], heuristic methods have been devised to solve it approximately [89,93,94] and optimised for computational efficiency [90]. The robustness of these heuristic methods has been tested against noise in the contact map [95].

Distance maps and multi-class contact maps (i.e. maps in which distances are quantised into more than 2 states) typically lead to more accurate 3D structures than binary maps and tend to be more robust when random noise is introduced in the map [29,96]. Nonetheless, one contact every twelve residues may be sufficient to allow robust and accurate topology-level protein structure modeling [97].

Predicted contact maps can also be helpful to score and, thus, guide the search for 3D models [98].

One of the earliest examples of 2D PSA annotations are $\beta - sheet$ pairings, i.e. AA partners in parallel and anti-parallel $\beta - sheet$ conformations. Machine/Deep Learning methods such as FFNN [99], BRNN [100] and multi-stage approaches [101] have been used since the late '90s to predict whether any 2 residues are partners in a $\beta - sheet$. Similarly, disulphide bridges (formed by $cysteine - cysteine$ residues) have been predicted by the Edmonds-Gabow algorithm and Monte Carlo simulation annealing [102], or hybrid solutions such as Hidden Markov Models and FFNN [103], and multi-stage FFNN, SVM and BRNN [104], alongside classic Machine Learning models such as SVM [105], pure Deep Learning models such as BRNN [106], and FFNN [107].

The prediction of a contact map's principal eigenvector (using BRNN) is instead an example of 1D PSA used to infer 2D characteristics [108]. The predictions of $\beta - sheet$ pairings, disulphide bridges and principal eigenvectors have been prompted by the need for

Please cite this article as: M. Torrisi, G. Pollastri and Q. Le, Deep learning methods in protein structure prediction, Computational and Structural Biotechnology Journal, https://doi.org/10.1016/j.csbj.2019.12.011

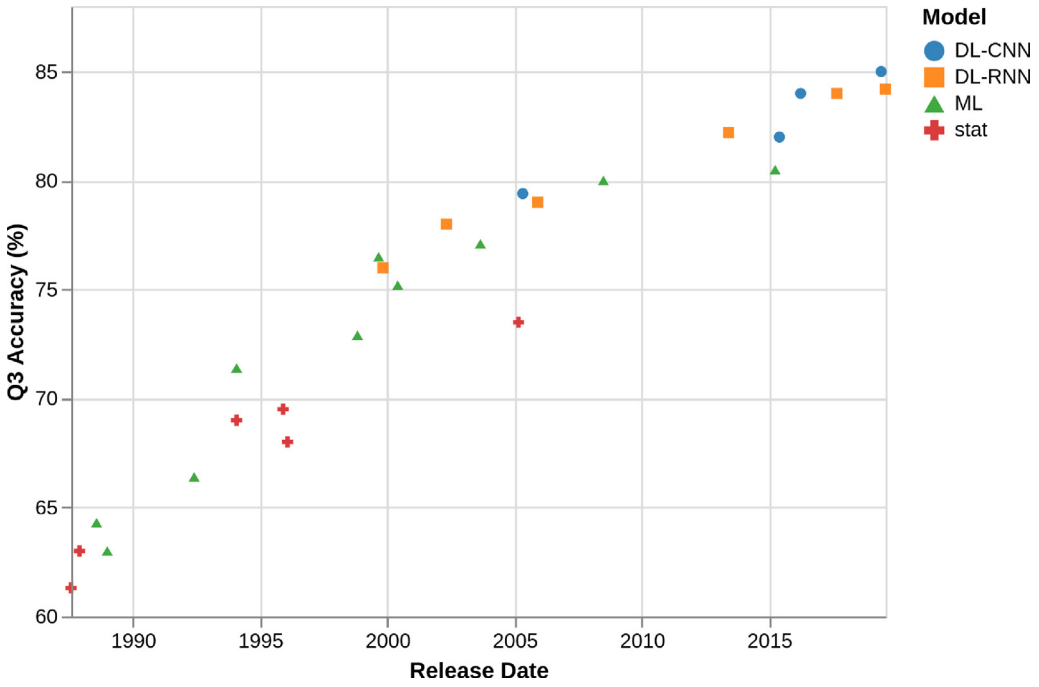

Fig. 3. Performances of secondary structure predictors over the years. "stat" are predictors based on statistical methods other than Neural Networks. "ML" are predictors based on shallow Neural Networks or Support Vector Machines. "DL-CNN" are Deep Learning methods based on Convolutional Neural Networks. "DL-RNN" are Deep Learning methods based on Recurrent Neural Networks. Data extracted from accompanying publications of predictors referenced in this article.

"easy-to-predict", informative abstractions which can be used to guide the prediction of more complex 2D PSA such as contact or distance maps. Ultimately, however, most interest in 2D PSA has been in the direct prediction of contact and distance maps as these contain most, if not all, the information necessary for the reconstruction of a protein's tertiary structure [89,29,96], while being translation and rotation invariant [91] which is a desirable property for the target of Machine Learning and Deep Learning algorithms.

Early methods for contact map prediction typically focused on simple, binary maps, and relied on statistical features extracted from evolutionary information in the form of alignments of multiple sequences. Features such as correlated mutations, sequence conservation, alignment stability and family size were inferred from multiple alignments and were shown to be informative for contact map prediction since the '90s [109,110]. Early methods often relied on simple linear combinations of features, though FFNN [111] and other Machine Learning algorithms such as Self-Organizing Maps [112] and SVM [113] quickly followed.

3.1. Modern and deep learning methods for 2D PSA prediction

2D-BRNN [72,124] are an extension to the BRNN architecture used to predict 1D PSA. These models, which are designed to process 2D maps of variable sizes, have 4 state vectors summarising information about the 4 cardinal corners of a map. 2D-BRNN have been applied to predict contact maps [72,124,108,125], multi-class contact maps [29], and distance maps [96]. Contact map predictions by 2D-BRNN have also been refined using cascaded FFNN [126]. Both ab initio and template-based predictors have been developed to predict maps (as well as 1D PSA) [29,96]. In particular, template-based contact and distance map predictors rely both on the sequence and structural information and, thus, are often better than ab initio predictors even when only dubious templates are available [29,96].

More recently, growing abundance of evolutionary information data and computational resources has led to substantial breakthroughs in contact map prediction [127]. More sophisticated statistical methods have been developed to calculate mutual information without the influence of entropy and phylogeny [128], co-evolution coupling [129], direct-coupling analysis (DCA) [130] and sparse inverse covariance estimation [131]. The ever-growing number of known sequences has led to the development of more optimised and, thus, faster tools [132] able to also run on GPU [133]. PSICOV [131], FreeContact [132] and CCMpred [133], which are notable results of this development, have allowed the exploitation of ever growing data banks and prompted a new wave of Deep Learning methods.

MetaPSICOV is a notable example of a Deep Learning method applied to PSICOV, FreeContact and CCMpred, as well as 1D features (such as predicted 1D PSA) [134]. MetaPSICOV is a two-stage FFNN with one hidden layer with 55 units. MetaPSICOV2, the following version, is a two-stage FFNN with two hidden layers with 160 units each and also a template-based predictor [114].

DeepCDpred is a multi-class contact map ab initio predictor which attempts to extend MetaPSICOV [115]. In particular, PSICOV is substituted with QUIC - a similarly accurate but significantly faster implementation of the sparse inverse covariance estimation - and the two-stage FFNN with an ensemble of 7 deeper FFNN (with 8 hidden layers) which are trained on different targets and, thus, result in a multi-class map predictor.

RaptorX-Contact is one of the first examples of contact map predictor based on a Residual CNN architecture [116]. RaptorX-Contact has been trained on CCMpred, mutual information, pairwise potential extraction and RaptorX-Property's output, i.e. secondary structure and solvent accessibility predictions [23]. RaptorX-Contact uses filters of size 3×3 and 5×5, 60 hidden units per layer and a total of 60 convolutional layers.

DNCON2 is a two-stage CNN trained on a set of input features similar to MetaPSICOV [117]. The first stage is composed of an ensemble of 5 CNN trained on 5 different thresholds, which feeds a following refining stage of CNN. The first stage of DNCON2 can be seen as a multi-class contact map predictor.

Please cite this article as: M. Torrisi, G. Pollastri and Q. Le, Deep learning methods in protein structure prediction, Computational and Structural Biotechnology Journal, https://doi.org/10.1016/j.csbj.2019.12.011

DeepContact (also known as i_Fold1) aims to demonstrate the superiority of CNN over FFNN to predict contact maps [118]. DeepContact is a 9-layer Residual CNN with 32 filters of size 5×5 trained on the same set of features used by MetaPSICOV. The outputs of the third, sixth and ninth layers are concatenated with the original input and fed to a last hidden layer to perform the final prediction.

DeepCov uses CNN to predict contact maps when limited evolutionary information is available [119]. In particular, DeepCov has been trained on a very limited set of input features: pair frequencies and covariance. This is one of the first notable examples of 2D PSA predictors which entirely skips the prediction of 1D PSA in its pipeline.

PconsC4 is a CNN with limited input features to significantly speed-up prediction time [120]. In particular, PconsC4 uses predicted 1D PSA, the GaussDCA score, APC-corrected mutual information, normalised APC-corrected mutual information and cross-entropy. PconsC4 requires only a recent version of Python and a GCC compiler with no need for any further external programs and appears to be significantly faster (and more accurate) than MetaPSICOV [120,114].

SPOT-Contact has been inspired by RaptorX-Contact and extends it by adding a 2D-RNN stage downstream of a CNN stage [121]. SPOT-Contact is an ensemble of models based on 120 convolutional filters – half 3×3 and half 5×5 – followed by a 2D-BRNN with 800 units – 200 LSTM cells for each of the 4 directions – and a final hidden layer composed of 400 units. Adam, a 50% dropout rate and layer normalization are among the Deep Learning techniques implemented to train this predictor. CCMpred, mutual and direct-coupling information are used as inputs as well as the output of SPIDER3, i.e. predictions of solvent accessibility, half-Sphere exposures, torsion angles and secondary structure [61].

TripletRes [122] is a contact map predictor that ranked first in the Contact Predictions category of the latest edition of CASP, a bi-annual blind competition for Protein Structure Prediction [135]. TripletRes is composed of 4 CNN trained end-to-end. More in detail, 3 coevolutionary inputs, i.e. the covariance matrix, precision matrix and coupling parameters of the Potts model, are fed to 3 different CNN which are then fused in a unique CNN downstream. Each CNN is composed of 24 residual convolutional layers with a kernel of size $3 \times 3 \times 64$. The training of TripletRes required 4 GPUs running concurrently - using Adam and a 80% dropout rate. TripletRes successfully identified and predicted both globally and locally multi-domain proteins following a divide et impera strategy.

AlphaFold [123] is a Protein Structure Prediction method that achieved the best performance in the Ab initio category of CASP13 [135]. Central to AlphaFold is a distance map predictor implemented as a very deep residual neural networks with 220 residual blocks processing a representation of dimensionality $64 \times 64 \times 128$ – corresponding to input features calculated from two 64 amino acid fragments. Each residual block has three layers including a 3×3 dilated convolutional layer – the blocks cycle through dilation of values 1, 2, 4, and 8. In total the model has 21 millions parameters. The network uses a combination of 1D and 2D inputs, including evolutionary profiles from different sources and co-evolution features. Alongside a distance map in the form of a very finely-grained histogram of distances, AlphaFold predicts Φ and Ψ angles for each residue which are used to create the initial predicted 3D structure. The AlphaFold authors concluded that the depth of the model, its large crop size, the large training set of roughly 29,000 proteins, modern Deep Learning techniques, and the richness of information from the predicted histogram of distances helped AlphaFold achieve a high contact map prediction precision.

Constant improvements in contact and distance map predictions over the last few years have directly resulted in improved 3D predictions. Fig. 4 reports the average quality of predictions submitted to the CASP competition for free modelling targets, i.e. proteins for which no suitable templates are available and predictions are therefore fully ab initio, between CASP9 (2010) and CASP13 (2018). Improvements especially over the last two editions are largely to be attributed to improved map predictions [127,136].

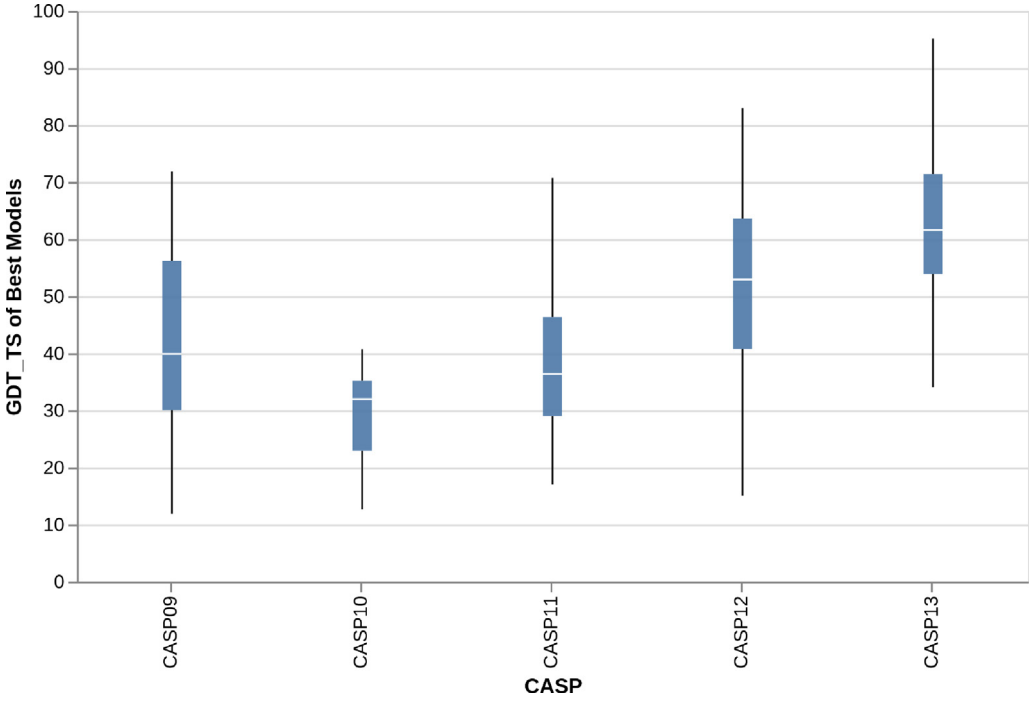

Fig. 4. Improvements in quality of 3D predictions for free modelling (ab initio) targets between CASP9 and CASP13.

Please cite this article as: M. Torrisi, G. Pollastri and Q. Le, Deep learning methods in protein structure prediction, Computational and Structural Biotechnology Journal, https://doi.org/10.1016/j.csbj.2019.12.011

Table 1
Deep Learning methods for 1D PSA prediction, along with models adopted and tools to gather evolutionary information, respectively. Secondary structure (SS), solvent accessibility (SA), torsion angles (TA), contact density (CD) and disordered regions (DR) are the PSA predicted.

Predictor	PSA	Model	Evolutionary Information
SPIDER2 [59]	SS, SA	Multi-stage FFNN	PSI-BLAST
SSpro/ACCpro5 [30]	SS, SA	BRNN-CNN	PSI-BLAST
Brewery [60]	SS, SA, TA, CD	Multi-stage BRNN-CNN	PSI-BLAST, HHblits
SPIDER3 [61]	SS, SA, TA, CD	BLSTM	PSI-BLAST, HHblits
RaptorX-Property [23]	SS, SA, DR	CNF	PSI-BLAST, HHblits
NetSurfP-2.0 [62]	SS, SA, TA, DR	BLSTM	HHblits, (or) MMseqs2

Table 2
Modern and Deep Learning methods for 2D PSA prediction, along with models adopted and tools to gather evolutionary information, respectively. Contact maps (CM), multi-class CM and distance maps (DM) are the PSA predicted.

Predictor	PSA	Model	Evolutionary Information
MetaPSICOV2 [114]	CM	Multi-stage FFNN	HHblits, JackHMMer
DeepCDpred [115]	multi-class CM	Multi-stage FFNN	HHblits
RaptorX-Contact [116]	multi-class CM	Residual CNN	HHblits
DNCON2 [117]	CM	Multi-stage CNN	HHblits, jackHMMer
DeepContact [118]	CM	Residual CNN	HHblits, jackHMMer
DeepCov [119]	CM	CNN	HHblits
Pconsc4 [120]	CM	CNN	HHblits
SPOT-Contact [121]	CM	Residual CNN 2D-BLSTM	HHblits, PSI-BLAST
TripletRes [122]	CM	Multi-stage residual CNN	HHblits, jackHMMer, HMMER
AlphaFold [123]	DM	Residual CNN	HHblits, PSI-BLAST

4. Summary and outlook

Proteins fold spontaneously in 3D conformations based only on the information present in their residues [7]. Protein Structure predictors are systems able to extract from the protein sequence information constraining the set of possible local and global conformations and use this to guide the folding of the protein itself. Deep Learning methods are successful at producing higher abstractions/representations while ignoring irrelevant variations of the input when sufficient amounts of data are provided to them [137]. Both characteristics together with the availability of rapidly growing protein databases increasingly make Deep Learning methods the preferred techniques to aid Protein Structure Prediction (see Tables 1 and 2). The highly complex landscape of protein conformations make Protein Structural Annotations one of the main research topics of interest within Protein Structure Prediction [11]. In particular, 1D annotations have been a central topic since the '60s [1,2] while the focus is progressively shifting towards more informative and complex 2D annotations such as contact maps and distance maps. This change of paradigm is mainly motivated by technological breakthroughs which result in continuous growth in computational power and protein sequences available thanks to next-generation sequencing and metagenomics [76,81].

Recent work on the prediction of 1D structural annotations [11,31,75,61], contact map prediction [117,122], and on overall structure prediction systems [123,138], emphasises the importance of more sophisticated pipelines to find and exploit evolutionary information from ever growing databases. This is often achieved by running several tools to find multiple homologous sequences in parallel [32,76,81] and, increasingly, by deploying Machine/Deep Learning techniques to independently process the sequence before fusing their outputs into the final prediction. The correlation between sequence alignment quality and accuracy of PSA predictors has been empirically demonstrated [139–141]. How to best gather and process homologous sequences is an active research topic, e.g. RawMSA is a suite of predictors which proposes to substitute the pre-processing of sequence alignments with an embedding step in order to learn a representation of protein sequences instead of pre-compressing homologous sequences into input features [142].

The same trend towards end-to-end systems has been attempted in the pipeline from processed homologous sequences to 3D structure, e.g. in NEMO [143], a differentiable simulator, and RGN (Recurrent Geometrical Network) [144], an end-to-end differentiable learning of protein structure. However, state-of-the-art structure predictors are still typically composed of multiple intelligent systems. The last mile of Protein Structure Prediction, i.e. the building, ranking and scoring of structural models, is also fertile ground for Machine Learning and Deep Learning methods [145,146]. E.g. MULTICOM exploits DNCON2 - a multi-class contact map predictor - to build structural models and to feed DeepRank - an ensemble of FFNN to rank such models [138]. DeepFragLib is, instead, a Deep Learning method to sample fragments (for ab initio structure prediction) [147]. The current need for multiple intelligent systems is supported by empirical results, especially in the case of hard predictions. Splitting proteins into composing domains, predicting 1D PSA, and optimising each component of the pipeline is particularly useful especially when alignment quality is poor [148].

Today, state-of-the-art systems for Protein Structure Prediction are composed by multiple specialised components [123,138,11] in which Deep Learning systems have an increasing, often crucial role, while end-to-end prediction systems entirely based on Deep Learning techniques, e.g. Deep Reinforcement Learning, may be on the horizon but are at present still immature. Progress in this field over the last few years has been substantial, even dramatic especially in the prediction of contact and distance maps [127,136], but the essential role of structural, evolutionary, and co-evolutionary information in this progress cannot be understated, with ab initio prediction quality still lagging that of template-based predictions, proteins with poor alignments being still a weak spot and prediction of protein structure from a single sequence being a challenge that is far from solved [149], although some progress has recently been observed for proteins with shallow alignments [150]. More generally, given that our current structure prediction pipelines rely almost exclusively on increasingly sophisticated and sensitive techniques to detect similarity to known structures and sequences, it is unclear whether predictions truly represent low energy structures unless we know they are correct. The prediction of protein misfolding [151,152] presents a

Please cite this article as: M. Torrisi, G. Pollastri and Q. Le, Deep learning methods in protein structure prediction, Computational and Structural Biotechnology Journal, https://doi.org/10.1016/j.csbj.2019.12.011

further challenge for the current prediction paradigm, with Machine Learning methods only making slow inroads [153]. Nevertheless, as more computational resources, novel techniques and ultimately, critically, increasing amounts of experimental data will become available [137], further improvements are to be expected.

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.

References

[1] Kendrew JC, Dickerson RE, Strandberg BE, Hart RG, Davies DR, Phillips DC, Shore VC. Structure of myoglobin: a three-dimensional Fourier synthesis at 2 A. resolution. Nature 1960;185:422–7.
[2] Perutz MF, Rossmann MG, Cullis AF, Muirhead H, Will G, North AC. Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-A. resolution, obtained by X-ray analysis. Nature 1960;185:416–22.
[3] Fleishman SJ, Whitehead TA, Ekiert DC, Dreyfus C, Corn JE, Strauch E-M, Wilson IA, Baker D. Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. Science 2011;332:816–21.
[4] Siegel JB, Zanghellini A, Lovick HM, Kiss G, Lambert AR, St.Clair JL, Gallaher JL, Hilvert D, Gelb MH, Stoddard BL, Houk KN, Michael FE, Baker D. Computational design of an enzyme catalyst for a stereoselective bimolecular diels-alder reaction. Science 2010;329:309–13.
[5] Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D. Design of a Novel globular protein fold with atomic-level accuracy. Science 2003;302:1364–8.
[6] Hsu PD, Lander ES, Zhang F. Development and applications of CRISPR-Cas9 for genome engineering. Cell 2014;157:1262–78.
[7] Anfinsen CB, Haber E, Sela M, White FH. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide Chain. Proc Nat Acad Sci 1961;47:1309–14.
[8] Anfinsen CB. Principles that govern the folding of protein chains. Science 1973;181:223–30.
[9] Levitt M, Warshel A. Computer simulation of protein folding. Nature 1975;253:694–8.
[10] Dill KA, MacCallum JL. The protein-folding problem, 50 years on. Science 2012;338:1042–6.
[11] Torrisi M, Pollastri G. Protein Structure Annotations. In: Shaik NA, Hakeem KR, Banaganapalli B, Elango R, editors. Essentials of Bioinformatics, Volume I: Understanding Bioinformatics: Genes to Proteins. Cham: Springer International Publishing; 2019. p. 201–34.
[12] Wright PE, Dyson HJ. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J Mol Biol 1999;293:321–31.
[13] Dunker AK, Obradovic Z. The protein trinity-linking function and disorder. Nat Biotechnol 2001;19:805.
[14] Tompa P. Intrinsically unstructured proteins. Trends Biochem Sci 2002;27:527–33.
[15] Dunker AK, Silman I, Uversky VN, Sussman JL. Function and structure of inherently disordered proteins. Curr Opin Struct Biol 2008;18:756–64.
[16] Ward JJ, McGuffin LJ, Bryson K, Buxton BF, Jones DT. The disopred server for the prediction of protein disorder. Bioinformatics 2004;20:2138–9.
[17] Schlessinger A, Punta M, Yachdav G, Kajan L, Rost B. Improved disorder prediction by combination of orthogonal approaches. PloS One 2009;4:e4433.
[18] Deng X, Eickholt J, Cheng J. Predisorder: ab initio sequence-based prediction of protein disordered regions. BMC Bioinf 2009;10:436.
[19] Walsh I, Martin AJ, Di Domenico T, Tosatto SC. Espritz: accurate and fast prediction of protein disorder. Bioinformatics 2011;28:503–9.
[20] Walsh I, Martin AJ, Di Domenico T, Vullo A, Pollastri G, Tosatto SC. Cspritz: accurate prediction of protein disorder segments with annotation for homology, secondary structure and linear motifs. Nucl Acids Res 2011;39:W190–6.
[21] Wang S, Ma J, Xu J. Aucpred: proteome-level protein disorder prediction by auc-maximized deep convolutional neural fields. Bioinformatics 2016;32:i672–9.
[22] Hanson J, Yang Y, Paliwal K, Zhou Y. Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks. Bioinformatics 2017;33:685–92.
[23] Wang S, Li W, Liu S, Xu J. RaptorX-Property: a web server for protein structure property prediction. Nucl Acids Res 2016;44:W430–5.
[24] Ferron F, Longhi S, Canard B, Karlin D. A practical overview of protein disorder prediction methods. Proteins: Struct Function Bioinf 2006;65:1–14.
[25] Deng X, Eickholt J, Cheng J. A comprehensive overview of computational protein disorder prediction methods. Mol BioSyst 2012;8:114–21.
[26] Meng F, Uversky VN, Kurgan L. Comprehensive review of methods for prediction of intrinsic disorder and its molecular functions. Cell Mol Life Sci 2017;74:3069–90.
[27] 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;8:201.
[28] Mooney C, Pollastri G. Beyond the Twilight Zone: automated prediction of structural properties of proteins by recursive neural networks and remote homology information. Proteins: Struct, Function, Bioinf 2009;77:181–90.
[29] Walsh I, Ba D, Martin AJ, Mooney C, Vullo A, Pollastri G. Ab initio and template-based prediction of multi-class distance maps by two-dimensional recursive neural networks. BMC Struct Biol 2009;9:5.
[30] 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;30:2592–7.
[31] Torrisi M, Kaleel M, Pollastri G. Deeper profiles and cascaded recurrent and convolutional neural networks for state-of-the-art protein secondary structure prediction. Sci Rep 2019;9:1–12.
[32] Altschul SF, Madden TL, Schffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl Acids Res 1997;25:3389–402.
[33] Remmert M, Biegert A, Hauser A, Sding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods 2012;9:173–5.
[34] Berman HM, Bourne PE, Westbrook J, Zardecki C. The protein data bank. In: Protein Structure. CRC Press; 2003. p. 394–410.
[35] Consortium U. Uniprot: a hub for protein information. Nucl Acids Res 2014;43:D204–12.
[36] Rumelhart D, Hinton G, Williams R. Learning representations by back-propagating errors. Nature 1986:533–6.
[37] Cybenko G. Approximation by superpositions of a sigmoidal function. Math Control Signals Syst 1989;2:303–14.
[38] Qian N, Sejnowski TJ. Predicting the secondary structure of globular proteins using neural network models. J Mol Biol 1988;202:865–84.
[39] Holley LH, Karplus M. Protein secondary structure prediction with a neural network. Proc Natl Acad Sci USA 1989;86:152–6.
[40] Holbrook SR, Muskal SM, Kim SH. Predicting surface exposure of amino acids from protein sequence. Protein Eng 1990;3:659–65.
[41] Goodfellow I, Bengio Y, Courville A. Deep learning. MIT Press; 2016.
[42] LeCun Y, Bottou L, Bengio Y, Haffner P, et al. Gradient-based learning applied to document recognition. Proc IEEE 1998;86:2278–324.
[43] Elman JL. Finding structure in time. Cognitive Sci 1990;14:179–211.
[44] Bengio Y, Simard P, Frasconi P, et al. Learning long-term dependencies with gradient descent is difficult. IEEE Trans Neural Networks 1994;5:157–66.
[45] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;9:1735–80.
[46] Cho K, van Merrienboer B, Bahdanau D, Bengio Y. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches, arXiv:1409.1259 [cs, stat]; 2014. .
[47] Baldi P, Brunak S, Frasconi P, Soda G, Pollastri G. Exploiting the past and the future in protein secondary structure prediction. Bioinformatics (Oxford, England) 1999;15:937–46.
[48] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. p. 770–78. .
[49] Davies DR. A correlation between amino acid composition and protein structure. J Mol Biol 1964;9:605–9.
[50] Chou PY, Fasman GD. Prediction of protein conformation. Biochemistry 1974;13:222–45.
[51] Lim VI. Structural principles of the globular organization of protein chains. A stereochemical theory of globular protein secondary structure. J Mol Biol 1974;88:857–72.
[52] Garnier J, Osguthorpe DJ, Robson B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J Mol Biol 1978;120:97–120.
[53] Kabsch W, Sander C. How good are predictions of protein secondary structure? FEBS Lett 1983;155:179–82.
[54] Rost B. Review: protein secondary structure prediction continues to rise. J Struct Biol 2001;134:204–18.
[55] Cornette JL, Cease KB, Margalit H, Spouge JL, Berzofsky JA, DeLisi C. Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. J Mol Biol 1987;195:659–85.
[56] Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol 1993;232:584–99.
[57] Eddy SR. Hidden Markov models. Curr Opin Struct Biol 1996;6:361–5.
[58] Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
[59] Heffernan R, Paliwal K, Lyons J, Dehzangi A, Sharma A, Wang J, Sattar A, Yang Y, Zhou Y. Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. Sci Rep 2015;5.
[60] Torrisi M, Kaleel M, Pollastri G. Brewery: state-of-the-art ab initio prediction of 1d protein structure annotations. Poster presented at BITS18 and CASP13; 2018..
[61] Heffernan R, Yang Y, Paliwal K, Zhou Y. Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility. Bioinformatics 2017;33:2842–9.
[62] Klausen MS, Jespersen MC, Nielsen H, Jensen KK, Jurtz VI, Snderby CK, Sommer MOA, Winther O, Nielsen M, Petersen B, Marcatili P. NetSurfP- 2.0:

improved prediction of protein structural features by integrated deep learning. Proteins: Struct, Funct, Bioinf 2019;87:520–7.
[63] Yi TM, Lander ES. Protein secondary structure prediction using nearest-neighbor methods. J Mol Biol 1993;232:1117–29.
[64] Levin JM. Exploring the limits of nearest neighbour secondary structure prediction. Protein Eng, Des Selection 1997;10:771–6.
[65] Xia Li, Xian-Ming Pan. New method for accurate prediction of solvent accessibility from protein sequence. Proteins: Struct, Function, Bioinf 2000;42:1–5.
[66] Bystroff C, Thorsson V, Baker D. HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. J Mol Biol 2000;301:173–90.
[67] Kuang R, Leslie CS, Yang A-S. Protein backbone angle prediction with machine learning approaches. Bioinformatics 2004;20:1612–21.
[68] Yuan Z. Better prediction of protein contact number using a support vector regression analysis of amino acid sequence. BMC Bioinf 2005;6:248.
[69] Cuff JA, Barton GJ. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins: Struct, Funct, Bioinf 2000;40:502–11.
[70] Wood MJ, Hirst JD. Protein secondary structure prediction with dihedral angles. Proteins: Struct, Funct, Bioinf 2005;59:476–81.
[71] Pollastri G, Przybylski D, Rost B, Baldi P. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. Proteins 2002;47:228–35.
[72] Pollastri G, Baldi P. Prediction of contact maps by GIOHMMs and recurrent neural networks using lateral propagation from all four cardinal corners. Bioinformatics 2002;18:S62–70.
[73] Pollastri G, McLysaght A. Porter: a new, accurate server for protein secondary structure prediction. Bioinformatics 2005;21:1719–20.
[74] Mirabello C, Pollastri G. Porter, PaleAle 4.0: high-accuracy prediction of protein secondary structure and relative solvent accessibility. Bioinformatics 2013;29:2056–8.
[75] Kaleel M, Torrisi M, Mooney C, Pollastri G. PaleAle 5.0: prediction of protein relative solvent accessibility by deep learning. Amino Acids 2019;51:1289–96.
[76] Steinegger M, Meier M, Mirdita M, Vhringer H, Haunsberger SJ, Sding J. HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinf 2019;20:473.
[77] Mooney C, Vullo A, Pollastri G. Protein structural motif prediction in multidimensional ø-ψ space leads to improved secondary structure prediction. J Comput Biol 2006;13:1489–502.
[78] Srivastava N, Hinton GE, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 2014;15:1929–58.
[79] Wang Z, Zhao F, Peng J, Xu J. Protein 8-class secondary structure prediction using Conditional Neural Fields. In: 2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). p. 109–14. .
[80] Wang S, Peng J, Ma J, Xu J. Protein secondary structure prediction using deep convolutional neural fields. Sci Rep 2016;6:18962.
[81] Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol 2017;35:1026–8.
[82] Chen K, Kurgan L. Computational prediction of secondary and supersecondary structures. In: Protein supersecondary structures. Springer; 2012. p. 63–86.
[83] Meng F, Kurgan L. Computational prediction of protein secondary structure from sequence. Curr Protocols Protein Sci 2016;86:2–3.
[84] Jiang Q, Jin X, Lee S-J, Yao S. Protein secondary structure prediction: a survey of the state of the art. J Mol Graph Model 2017;76:379–402.
[85] Oldfield CJ, Chen K, Kurgan L. Computational prediction of secondary and supersecondary structures from protein sequences. In: Protein Supersecondary Structures. Springer; 2019. p. 73–100.
[86] Zhang H, Zhang T, Chen K, Kedarisetti KD, Mizianty MJ, Bao Q, Stach W, Kurgan L. Critical assessment of high-throughput standalone methods for secondary structure prediction. Briefings Bioinf 2011;12:672–88.
[87] Cheng J, Choe M-H, Elofsson A, Han K-S, Hou J, Maghrabi AHA, McGuffin LJ, Menndez-Hurtado D, Olechnovic K, Schwede T, Studer G, Uziela K, Venclovas E, Wallner B. Estimation of model accuracy in CASP13. Proteins: Struct, Funct, Bioinf 2019;87:1361–77.
[88] Kuhlman B, Bradley P. Advances in protein structure prediction and design. Nat Rev Mol Cell Biol 2019.
[89] Vendruscolo M, Kussell E, Domany E. Recovery of protein structure from contact maps. Fold Des 1997;2:295–306.
[90] Vassura M, Margara L, Di Lena P, Medri F, Fariselli P, Casadio R. Reconstruction of 3d Structures From Protein Contact Maps. IEEE/ACM Trans Comput Biol Bioinf 2008;5:357–67.
[91] Bartoli L, Capriotti E, Fariselli P, Martelli PL, Casadio R. The pros and cons of predicting protein contact maps, Methods in Molecular. Biology (Clifton, N.J.) 2008;413:199–217.
[92] Breu H, Kirkpatrick DG. Unit disk graph recognition is NP-hard. Comput Geometry 1998;9:3–24.
[93] Ba D, Martin AJ, Mooney C, Vullo A, Walsh I, Pollastri G. Distill: a suite of web servers for the prediction of one-, two- and three-dimensional structural features of proteins. BMC Bioinf 2006;7:402.
[94] Zhang C, Mortuza SM, He B, Wang Y, Zhang Y. Template-based and free modeling of I-TASSER and QUARK pipelines using predicted contact maps in CASP12. Proteins: Struct, Funct, Bioinf 2018;86:136–51.
[95] Vassura M, Di Lena P, Margara L, Mirto M, Aloisio G, Fariselli P, Casadio R. Blurring contact maps of thousands of proteins: what we can learn by reconstructing 3d structure. BioData Mining 2011;4:1.
[96] Kukic P, Mirabello C, Tradigo G, Walsh I, Veltri P, Pollastri G. Toward an accurate prediction of inter-residue distances in proteins using 2d recursive neural networks. BMC Bioinf 2014;15:6.
[97] Kim DE, DiMaio F, Wang RY-R, Song Y, Baker D. One contact for every twelve residues allows robust and accurate topology-level protein structure modeling. Proteins 2014;82:208–18.
[98] Tress ML, Valencia A. Predicted residue-residue contacts can help the scoring of 3d models. Proteins: Struct, Funct, Bioinf 2010;78:1980–91.
[99] Asogawa M. Beta-sheet prediction using inter-strand residue pairs and refinement with hopfield neural network. Genome Inf 1996;7:198–9.
[100] Baldi P, Pollastri G, Andersen CA, Brunak S. Matching protein beta-sheet partners by feedforward and recurrent neural networks. In: Proceedings. International Conference on Intelligent Systems for Molecular Biology 8 (2000) 25–36. .
[101] Cheng J, Baldi P. Three-stage prediction of protein -sheets by neural networks, alignments and graph algorithms. Bioinformatics 2005;21:i75–84.
[102] Fariselli P, Casadio R. Prediction of disulfide connectivity in proteins. Bioinformatics 2001;17:957–64.
[103] Martelli PL, Fariselli P, Malaguti L, Casadio R. Prediction of the disulfide bonding state of cysteines in proteins with hidden neural networks. Protein Eng, Des Selection 2002;15:951–3.
[104] Ceroni A, Passerini A, Vullo A, Frasconi P. DISULFIND: a disulfide bonding state and cysteine connectivity prediction server. Nucl Acids Res 2006;34: W177–81.
[105] Tsai C-H, Chen B-J, Chan C-H, Liu H-L, Kao C-Y. Improving disulfide connectivity prediction with sequential distance between oxidized cysteines. Bioinformatics 2005;21:4416–9.
[106] Vullo A, Frasconi P. Disulfide connectivity prediction using recursive neural networks and evolutionary information. Bioinformatics 2004;20:653–9.
[107] Ferr F, Clote P. DiANNA: a web server for disulfide connectivity prediction. Nucl Acids Res 2005;33:W230–2.
[108] Vullo A, Walsh I, Pollastri G. A two-stage approach for improved prediction of residue contact maps. BMC Bioinf 2006;7:180.
[109] Göbel U, Sander C, Schneider R, Valencia A. Correlated mutations and residue contacts in proteins. Proteins: Struct, Funct, Bioinf 1994;18:309–17.
[110] Pazos F, Helmer-Citterich M, Ausiello G, Valencia A. Correlated mutations contain information about protein-protein interaction 11edited by A.R. Fersht. J Mol Biol 1997;271:511–23.
[111] Fariselli P, Olmea O, Valencia A, Casadio R. Prediction of contact maps with neural networks and correlated mutations. Protein Eng, Des Selection 2001;14:835–43.
[112] MacCallum RM. Striped sheets and protein contact prediction. Bioinformatics 2004;20:i224–31.
[113] Cheng J, Baldi P. Improved residue contact prediction using support vector machines and a large feature set. BMC Bioinf 2007;8:113.
[114] Buchan DWA, Jones DT. Improved protein contact predictions with the MetaPSICOV2 server in CASP12. Proteins 2018;86(Suppl 1):78–83.
[115] Ji S, Oru T, Mead L, Rehman MF, Thomas CM, Butterworth S, Winn PJ. DeepCDpred: inter-residue distance and contact prediction for improved prediction of protein structure. PLOS ONE 2019;14.
[116] Wang S, Sun S, Li Z, Zhang R, Xu J. Accurate de novo prediction of protein contact map by ultra-deep learning model. PLOS Comput Biol 2017;13: e1005324.
[117] Adhikari B, Hou J, Cheng J. DNCON2: improved protein contact prediction using two-level deep convolutional neural networks. Bioinformatics 2018;34:1466–72.
[118] Liu Y, Palmedo P, Ye Q, Berger B, Peng J. Enhancing evolutionary couplings with deep convolutional neural networks. Cell Syst 2018;6:65–74.e3.
[119] Jones DT, Kandathil SM. High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features. Bioinformatics 2018;34:3308–15.
[120] Michel M, Menndez Hurtado D, Elofsson A. PconsC4: fast, accurate and hassle-free contact predictions. Bioinformatics 2019;35:2677–9.
[121] Hanson J, Paliwal K, Litfin T, Yang Y, Zhou Y. Accurate prediction of protein contact maps by coupling residual two-dimensional bidirectional long short-term memory with convolutional neural networks. Bioinformatics 2018;34:4039–45.
[122] Li Y, Zhang C, Bell EW, Yu D-J, Zhang Y. Ensembling multiple raw coevolutionary features with deep residual neural networks for contact-map prediction in CASP13. Proteins: Struct, Funct, Bioinf 2019.
[123] Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, Qin C, Žídek A, Nelson AWR, Bridgland A, Penedones H, Petersen S, Simonyan K, Crossan S, Kohli P, Jones DT, Silver D, Kavukcuoglu K, Hassabis D. Protein structure prediction using multiple deep neural networks in CASP13. Proteins: Struct, Funct, Bioinf 2019.
[124] Baldi P, Pollastri G. The principled design of large-scale recursive neural network architectures-DAG-RNNs and the protein structure prediction problem. J Mach Learn Res 2003;4:575–602.
[125] Tegge AN, Wang Z, Eickholt J, Cheng J. NNcon: improved protein contact map prediction using 2d-recursive neural networks. Nucl Acids Res 2009;37: W515–8.
[126] Di Lena P, Nagata K, Baldi P. Deep architectures for protein contact map prediction. Bioinformatics 2012;28:2449–57.

Please cite this article as: M. Torrisi, G. Pollastri and Q. Le, Deep learning methods in protein structure prediction, Computational and Structural Biotechnology Journal, https://doi.org/10.1016/j.csbj.2019.12.011

[127] Schaarschmidt J, Monastyrskyy B, Kryshtafovych A, Bonvin AM. Assessment of contact predictions in casp12: co-evolution and deep learning coming of age. Proteins: Struct, Funct, Bioinf 2018;86:51–66.
[128] Dunn SD, Wahl LM, Gloor GB. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics 2008;24:333–40.
[129] Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, Sander C. Protein 3d structure computed from evolutionary sequence variation. PLoS ONE 2011;6.
[130] Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc Nat Acad Sci 2011;108:E1293–301.
[131] Jones DT, Buchan DWA, Cozzetto D, Pontil M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics 2012;28:184–90.
[132] Kaján L, Hopf TA, Kalaš M, Marks DS, Rost B. FreeContact: fast and free software for protein contact prediction from residue co-evolution. BMC Bioinf 2014;15:85.
[133] Seemayer S, Gruber M, Söding J. CCMpred-fast and precise prediction of protein residue-residue contacts from correlated mutations. Bioinformatics 2014;30:3128–30.
[134] Jones DT, Singh T, Kosciolek T, Tetchner S. MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. Bioinformatics 2015;31:999–1006.
[135] Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J. Critical assessment of methods of protein structure prediction (casp) – round xiii. Proteins: Struct, Funct, Bioinf 2019;87:1011–20.
[136] Shrestha R, Fajardo E, Gil N, Fidelis K, Kryshtafovych A, Monastyrskyy B, Fiser A. Assessing the accuracy of contact predictions in casp13. Proteins: Struct, Funct, Bioinf 2019;87:1058–68.
[137] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015:436–44.
[138] Hou J, Wu T, Cao R, Cheng J. Protein tertiary structure modeling driven by deep learning and contact distance prediction in CASP13. Proteins: Struct, Funct, Bioinf 2019;87:1165–78.
[139] Fox G, Sievers F, Higgins DG. Using de novo protein structure predictions to measure the quality of very large multiple sequence alignments. Bioinformatics 2015;32:814–20.
[140] Le Q, Sievers F, Higgins DG. Protein multiple sequence alignment benchmarking through secondary structure prediction. Bioinformatics 2017;33:1331–7.
[141] Sievers F, Higgins DG. Quantest2: benchmarking multiple sequence alignments using secondary structure prediction. Bioinformatics 2020.
[142] Mirabello C, Wallner B. rawMSA: end-to-end deep learning using raw multiple sequence alignments. PLOS ONE 2019;14: e0220182.
[143] Ingraham J, Riesselman A, Sander C, Marks D. Learning Protein Structure with a Differentiable Simulator. In: International Conference on Learning Representations. .
[144] AlQuraishi M. End-to-end differentiable learning of protein structure. Cell Syst 2019;8. 292–301.e3.
[145] Martin AJM, Mirabello C, Pollastri G. Neural network pairwise interaction fields for protein model quality assessment and ab initio protein folding. Curr Protein Peptide Sci 2011;12:549–62.
[146] Cao R, Adhikari B, Bhattacharya D, Sun M, Hou J, Cheng J. QAcon: single model quality assessment using protein structural and contact information with machine learning techniques. Bioinformatics 2017;33:586–8.
[147] Wang T, Qiao Y, Ding W, Mao W, Zhou Y, Gong H. Improved fragment sampling for ab initio protein structure prediction using deep neural networks. Nat Mach Intell 2019;1:347–55.
[148] Wu T, Hou J, Adhikari B, Cheng J. Analysis of several key factors influencing deep learning-based inter-residue contact prediction. Bioinformatics 2019.
[149] Kandathil SM, Greener JG, Jones DT. Recent developments in deep learning applied to protein structure prediction. Proteins: Struct, Funct, Bioinf 2019;87:1179–89.
[150] Abriata LA, Tamò GE, Dal Peraro M. A further leap of improvement in tertiary structure prediction in casp13 prompts new routes for future assessments. Proteins: Struct, Funct, Bioinf 2019;87:1100–12.
[151] Knowles TP, Vendruscolo M, Dobson CM. The amyloid state and its association with protein misfolding diseases. Nat Rev Mol Cell Biol 2014;15:384–96.
[152] Luheshi LM, Dobson CM. Bridging the gap: from protein misfolding to protein misfolding diseases. FEBS Lett 2009;583:2581–6.
[153] Walsh I, Seno F, Tosatto SC, Trovato A. Pasta 2.0: an improved server for protein aggregation prediction. Nucl Acids Res 2014;42:W301–7.

Please cite this article as: M. Torrisi, G. Pollastri and Q. Le, Deep learning methods in protein structure prediction, Computational and Structural Biotechnology Journal, https://doi.org/10.1016/j.csbj.2019.12.011