Received: 25 November 2021 | Revised: 3 March 2022 | Accepted: 7 March 2022

DOI: 10.1002/wcms.1618

ADVANCED REVIEW

WIREs COMPUTATIONAL MOLECULAR SCIENCE

WILEY

Machine learning solutions for predicting protein–protein interactions

Rita Casadio | **Pier Luigi Martelli** | **Castrense Savojardo**

Biocomputing Group, University of Bologna, Bologna, Italy

Correspondence
Pier Luigi Martelli, Biocomputing Group, University of Bologna, Bologna, Italy.
Email: pierluigi.martelli@unibo.it

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

Edited by: Modesto Orozco, Associate Editor and Peter R. Schreiner, Editor-in-Chief

Abstract

Proteins are "social molecules." Recent experimental evidence supports the notion that large protein aggregates, known as biomolecular condensates, affect structurally and functionally many biological processes. Condensate formation may be permanent and/or time dependent, suggesting that biological processes can occur locally, depending on the cell needs. The question then arises as to which extent we can monitor protein-aggregate formation, both experimentally and theoretically and then predict/simulate functional aggregate formation. Available data are relative to mesoscopic interacting networks at a proteome level, to protein-binding affinity data, and to interacting protein complexes, solved with atomic resolution. Powerful algorithms based on machine learning (ML) can extract information from data sets and infer properties of never-seen-before examples. ML tools address the problem of protein–protein interactions (PPIs) adopting different data sets, input features, and architectures. According to recent publications, deep learning is the most successful method. However, in ML-computational biology, convincing evidence of a success story comes out by performing general benchmarks on blind data sets. Results indicate that the state-of-the-art ML approaches, based on traditional and/or deep learning, can still be ameliorated, irrespectively of the power of the method and richness in input features. This being the case, it is quite evident that powerful methods still are not trained on the whole possible spectrum of PPIs and that more investigations are necessary to complete our knowledge of PPI-functional interactions.

This article is categorized under:

Software > Molecular Modeling
Structure and Mechanism > Computational Biochemistry and Biophysics
Data Science > Artificial Intelligence/Machine Learning
Molecular and Statistical Mechanics > Molecular Interactions

KEYWORDS

deep learning, machine learning, protein-protein interactions

1 | INTRODUCTION

Proteins are large, complex molecules that play many critical roles, participating in a variety of biological functional processes. They are required for the organization, function, and regulation of the life span of any cell type. They can perform chemical catalysis, supporting billions of biochemical reactions, and can be part of larger macromolecular machines, whose structure and functional role has been partly highlighted and is the subject of ongoing research. Proteins can interact with other molecules. Interaction partners include ions, small organic molecules, membrane lipids, nucleic acids, small peptides, and proteins, to generate homo- and hetero-complexes. In the crowded cellular environment, protein through evolution have been able to develop and maintain efficiency and binding specificity for function.[1]

Since the past decade, interest is growing in understanding the organization of the cell interior and its dynamics in relation to its physiology.[2–4] A new vision leads descriptions: apparently, proteins and other biomolecules aggregate either transiently or permanently, depending on the cell needs, and generate molecular condensates, broadly defined as concentrated foci, lacking a surrounding membrane, or membrane-less organelles.[5] Biomolecular condensates have been documented in different compartments of eukaryotic cells, including the nucleus, the nucleolus, and the cytoplasm. The presence of different types of membrane-less organelles is now well established, after the first report, over a century ago, of the Cajal inclusion bodies in neuron nuclei. Condensates of different dimensions seem to have a widespread role in cell biology, allowing the formation of stable and/or transient aggregates whose role is under investigation to unravel cell function and complexity in normal and pathological conditions.[6–9]

In this dynamic scenario, the problem of protein–protein molecular interactions (PPI) is evidently an issue. Proteins can interact with genomic DNA and RNA to trigger transcription and protein biosynthesis, and monomeric protein chains can give rise to stable functional complexes. Less documented is PPI that drive the formation of transient complexes, apparently necessary to assemble condensates for functioning.[10] A question, above all, is becoming urgent. If PPIs are necessary to trigger biochemical reaction mechanisms and to enhance activity, how can we distinguish functional PPI from spontaneous forms of aggregates that eventually may occur due to nonspecific short-range interactions?

Our knowledge of PPI is mainly based on two different types of evidence. The first one is grounds on the presence of complexes known with atomic details in the Protein Data Bank (PDB, https://www.rcsb.org/). The other, at a higher and much broader scale, stands from the results of different techniques which investigate the formation of protein complexes at large in a cell proteome. Data analysis produces networks of interacting proteins which routinely cover fractions of the different proteomes.[11,12]

Possibly we should be able to establish links among these layers of information to model condensate formation and understand the role of PPI with a bottom-up approach.[5]

While biological descriptions seem to have reached unprecedented levels of information, powerful computational approaches became available for data analysis. They are based on machine learning (ML), a procedure that can discover relations, if existing, among dependent and independent data sets. By now, ML has a long-standing tradition in computational biology, and many tools are publicly available to address different problems, from bioimage analysis to protein structure prediction and their interactions.[13]

In the following we will shortly describe ML and focus on the problem of PPI, listing major resources of data available to explore data associations with ML approaches. The strategy here is to briefly highlight the problems, also including our biophysical knowledge, and how they have been tackled with ML. The main focus is however on the prediction of PPI as network of interactions, three-dimensional (3D) protein aggregates, and PPI prediction sites (PPIs) on structures and sequences.

2 | MACHINE LEARNING

ML refers to many algorithms able to automatically build models for inference and clustering, starting from a set of data called *training set*. ML can be *unsupervised* or *supervised*.

Unsupervised ML procedures aim to discover patterns and similarities in training data and to identify meaningful clusters and data representations.

In turn, *supervised* ML algorithms infer mapping between two spaces (input and output), based on known training examples.[14] Supervised ML aims to implement tools that generalize the learned associations to new examples.

Supervised ML methods can be adopted for classification or regression, depending on the discrete or continuous property of the output space.

In the context of computing PPIs, supervised ML is most relevant, and we will focus on it.

Different ML approaches have been developed during the last decades and the most adopted ones are shortlisted in Box 1, being the exhaustive description of all relevant algorithms beyond the scope of this review. Methods referred to as traditional in the recent literature,[15] include shallow neural networks (NN), support vector machines (SVMs), random forests (RFs), and probabilistic methods such as Hidden Markov Models (HMMs) and conditional random fields (CRFs). Starting from about 2010, the field of ML-based applications has been more and more influenced by the so-called *deep* ML approaches, basically derived from the evolution of traditional NNs.[13,16]

Notwithstanding the huge variety of available approaches, ML methods share a few key issues related to *data quality, data representation, training algorithms*, and *validation procedures*.

Training data are at the core of the learning process, since the inference rules are automatically extracted from them with the learning procedure, ideally using the minimal amount of a priori assumptions. Therefore, the dimension and the quality of the training set are of outmost importance. Training data should derive from experimental measures affected by small errors. They should be reproducible and high quality and should uniformly represent all the input space, avoiding redundancies that can bias learning toward given classes of examples. The accurate selection of the training set is crucial for achieving good generalization performance.

A second relevant issue is *data representation*. Data must be wisely represented with all the features potentially relevant for inference. The definition of relevant features routinely requires previous knowledge, preliminary analysis, and data preprocessing, especially when traditional ML approaches are adopted. Deep learning approaches are more effective in extracting important features and their relations, when the training data set is large and enough representative. Therefore, they allow the implementation of the so-called end-to-end models that integrate preprocessing pipelines.

All supervised ML methods rely on training algorithms that set the value of a (routinely large) number of trainable parameters with the goal of finding the optimal fitting between input and output, starting from known examples presented during the training phase. The optimization criteria differ depending on the algorithm. Traditional and deep NNs minimize an error (or cost) function with a gradient descent algorithm called back-propagation; probabilistic methods such as HMMs and CRFs maximize probability functions with expectation–maximization or gradient-ascent protocols.

BOX 1 Machine learning methods

Traditional machine learning (ML) includes different computational methods briefly listed below.

Support vector machines (SVM) are methods estimating the optimal linear separation between two classes of data. Nonlinear classifications can be achieved using kernels.

Probabilistic models (such as *Hidden Markov Models*, HMMs) are adopted to learn the most probable labeling of input samples (sequences, structures, and graphs) taking into consideration complex contexts.

Random forests (RF) are ensemble learning methods than decrease the prediction error rates by averaging a multitude of simple decision trees.

Shallow feed-forward neural networks (NN) consist of simple computational units, called neurons, that communicate between each other through connections whose weights can be trained with the back-propagation algorithm. Neurons are routinely organized in layers: one encodes the input; one provides the output. Input and output can be separated by few hidden layers.

Deep learning methods are NNs with many hidden layers able to extract complex relations among input features. Based on the type of hidden layers and connection topology different classes of deep NNs are defined.[13]

Recurrent networks extract relationships in sequential data through memory layers, feedback, and time-delay loops. Long short-term memory networks belong to this class.

In *convolutional networks*, hidden layers consist of several filters that extract and pool local relations from input layers organized as matrixes or tensors.

Graph convolutional networks extend learning to structures where the relations among neurons are described by graphs. Methods tackling more complex structures are known as *geometric deep learning*.

In *attention networks*, an additional layer flexibly identifies the most relevant parts of the input.

Training procedures are often iterative, as in the case of the gradient-based ones, and require important computational resources, particularly for deep ML methods. Besides the trainable parameters, ML models are characterized by several hyper-parameters that define their overall architecture. Differently from trainable parameters, hyper-parameters are not optimized during the learning procedure and their values must be selected by performing a search in the hyper-parameter space.[13]

The *validation procedure* is a critical step for assessing the generalization performance of trained methods, which is its effectiveness in inferring the correct output from input data not used to learn the mapping. To this aim, a subset of known examples must be taken apart to generate a testing set, used to evaluate different statistical scores of performances, including *accuracy*, *recall*, *precision*, and *correlation indexes*.[17]

Non-redundancy among training and testing data is of outmost relevance for avoiding polarization of the method toward a particular class of examples. Best practices require the adoption of four different and independent sets of known examples: the training set for learning the trainable parameters, the validation set to optimize the hyper-parameters, the testing set to evaluate the performance, and a blind set (not including data of the training and testing set) to ultimately score the method and for benchmarking different methods. Different schemas, such as *cross-validation*, can be also applied to increase the statistical confidence of the evaluation. The adoption of a rigorous validation procedure is fundamental to minimize the overfitting risk. Guidelines and recommendation for the use of ML methods in computational biology are extensively reported in recent literature.[18,19]

The need for comparative evaluation of different methods requires the compilation of blind test sets independent of the training sets of all the evaluated tools. To this aim, different communities organize international critical assessment experiments in which computational methods are tested on examples whose solutions are unknown and are released only after the deadline for submitting predictions. In the context of protein–protein interaction (PPI), the critical assessment of prediction of interactions (CAPRI, https://www.ebi.ac.uk/msd-srv/capri/) is regularly organized.[20]

The training of deep learning is mostly associated with end-to-end learning, where a complex learning system is trained by applying a gradient-based learning to the system as a whole. The most striking success story of a combination of deep ML procedures and biophysical and bioinformatic knowledge, derived from the protein universe (protein structure, its representation in terms of contact maps, and evolutionary information as derived from multiple sequence alignments) is Alpha Fold2[21] and its score in predicting protein structures at the last benchmark of CASP14 (Critical Assessment of Protein Structure Prediction, 14th edition, https://predictioncenter.org/casp14/). Although the algorithm has still a modest performance in correctly recognizing intra-protein domain interactions for chains poorly represented in the database, perspectives are promising for its extension to the computation of inter-protein interaction sites[22] (see also Section 4.3).

3 | PPIs AT DIFFERENT SCALES AND THEIR PREDICTIONS

As mentioned in the Introduction (Section 1), several studies highlight all the functions that are supported by condensate formations.[5] Data are still sparse, not yet collected in relational databases. They have been recently reviewed at length, focusing on the different mechanisms of biological processes.[5] However, ML applications routinely require the development of training data sets that should be shared for comparison and benchmarking among different implementations. We will list here which repositories contain data useful for ML developments in the field of PPIs.

3.1 | Detection of PPI at a proteome scale and their prediction

Data on PPI can describe putative direct and non-direct interactions occurring at a mesoscopic level in the cell.

In the last decades, possible insights in PPI at a mesoscopic level became available in many organisms and human tissues, resulting from the applications of different techniques[12] (Box 2 for details). Routinely, PPIs are represented as networks, where nodes are the proteins and links are the detected interactions.

Despite many efforts, no single technique can capture all the possible interactions in a cell considering the different sensitivity of the methods, and the intrinsic changing in protein expression, which makes it difficult to capture the real protein content and its continuous changing over time.[23] Possibly, with the advent of single-cell proteomics, more homogeneous and non-ambiguous data will become available.[24,25]

Different databases have been implemented over the years, collecting data from different types of experiments.[26]

BOX 2 Experimental methods for characterizing PPIs

In the last 20 years, technology-enabled experimental procedures for the large-scale determination of all putative PPIs in a system, allowing to chart the interactome of whole organisms.[12]

Two-hybrid screening detects binary interactions in eukaryotic cells. Two domains of a transcription complex are fused to two different proteins (bait and prey). The interaction, also weak, between bait and prey activates the expression of a reporter gene.[31]

Affinity purification is a chromatographic assay for isolating all the interactors of a bait protein from a mixture[32] while in *co-fractionation* experiments native complexes are separated with physicochemical techniques.[33] In both cases, proteomics techniques, mainly based on mass spectrometry, are used to recognize proteins.[34]

Experimental methods largely differ in sensitivity and precision, also in relation, to the binding affinity of the interacting proteins, their subcellular localization, and their ability in discriminating direct from indirect interactions.[35]

Large-scale techniques cannot however characterize the region of the protein surfaces in which the interaction takes place. To collect this information, routinely experimental data on the structure of the complex are required, mainly based on *x-ray diffractometry* and/or *nuclear magnetic resonance spectrometry*, notoriously two low-throughput techniques. Recently, *cryogenic electron microscopy* expanded the possibility to resolve the structure of large complexes, but it is still unapplicable at the whole interactome scale.[36]

Comprehensive and popular databases of PPIs are IntAct (https://www.ebi.ac.uk/intact/home) and BioGRID4.4 (https://thebiogrid.org/). Both provide free, open-source databases, and analysis tools for molecular interaction data. All interactions derive from literature curation or direct user submissions.

Search Tool for the Retrieval of Interacting Genes/proteins (STRING, version 11.5, https://string-db.org/) is by far the largest collection of PPIs, presently including over 20 billion interactions in about 14,000 organisms, relative to about 68 million proteins. STRING is a database of known and predicted PPIs. The interactions include direct (physical) and indirect (functional) associations. Indirect interactions derive from computational predictions, from knowledge transfer between organisms, and from interactions aggregated from other (primary) databases. Other specialized databases collecting experimental, as well as computed interactions at a proteome scale, are available.[26]

Recently, deep learning applications for predicting at large networks of PPIs were proposed.[27,28] These approaches are however validated towards data sets whose completeness and reproducibility may be an issue. While condensates increase the complexity of the scenario, it is very hard to assess to which extent the above-mentioned data are complete. The ratio of the number of proteins and the number of coding genes for a given organism can give an estimate of the completeness of the proteome space; however, the number of interactions is still unknown and difficult to evaluate for any organism, even on theoretical grounds. Recently, some improvements in networks validation were described, including an analysis of network paths[29] and its variant integrating complementary interface and gene duplication.[30] More to it, when organism-specific interactomes are compared, networks of large dimensions have routinely small overlap. This is often due to the different experimental approaches, error rates of the experimental procedures, different levels of protein expression and post-translational processing in the expression systems.[23]

3.2 | Detection and computational prediction of protein–protein binding affinities

Macromolecular assemblies in vivo are explained by a full range of molecular mechanisms, classified as active processes, which consume energy for generating the condensate and passive thermodynamic processes, including liquid–liquid phase separation (LLPS).[5]

In phase behavior, like in protein phase separation,[3] besides pairwise interactions, higher multibody interactions can occur to mediate membrane-less foci formation. The basic idea is that proteins involved in aggregate formation should be in principle endowed with different and flexible interaction patches for their multiple interactions within the foci and the environment.[3,5] Most of our knowledge on affinity derives from in vitro experiments. For decades, measures of the pairwise binding affinities among proteins focused on characterizing thermodynamically and kinetically

conformational equilibria, in which the environmental solvent effect (routinely polar) is included. This allowed the categorization of pairwise interactions as short-lived with low binding affinity, and "obligatory," long-lived with high binding affinity. However, the spectrum includes any possible value among the two extremes.[37]

A major problem is therefore the extension of binding affinities from in vitro to in vivo experiments,[4,10] where multiple interactions may affect binding, including assembly cooperativity, molecular concentration, and properties.[3,5]

Rate constants ($K_D = k_{on}/k_{off}$) for pairwise protein association span six orders of magnitude, from $<10^3$ M^{-1} s^{-1} to $>10^9$ M^{-1} s^{-1}, while rate constants for protein dissociations span some eight orders of magnitude, from $<10^{-6}$ M^{-1} s^{-1} to $>10^2$ M^{-1} s^{-1}. Apparently, fast associations are electrostatically driven, while slower ones result from major structural rearrangements upon complexations.[37]

PDBbind (http://pdbbind.org.cn/) is a comprehensive collection of experimentally measured binding affinity data for all biomolecular complexes deposited in the Protein Data Bank (PDB). The current release (version 2020) provides binding affinity data for a total of 23,496 biomolecular complexes in the PDB, comprising protein–ligand (19,443), protein–protein (2852), protein–nucleic acid (1052), and nucleic acid–ligand complexes (149).[38] PDBbind includes a core set, providing a relatively small set of high-quality protein–ligand complexes for validating docking/scoring methods. The data set contributes to the popular Comparative Assessment of Scoring Functions (CASF) benchmark (http://www.pdbbind.org.cn/casf.php).

Structural Kinetic and Energetic Database of Mutant Protein Interactions (SKEMPI) contains data on the changes in thermodynamic parameters and kinetic rate constants upon mutation, for PPIs for which a structure of the complex is solved and is available in PDB (https://life.bsc.es/pid/skempi2/).[39]

3.3 | Detection of the binding affinity of protein–protein complexes with atomic resolution

Several computational tools are available for binding affinity prediction.[40] They include methods based on force fields and docking, knowledge-based scoring of single protein–protein complexes, ensemble-based approaches, and binding-free energy simulations.

A broad spectrum of ML machine-learning techniques, including supervised machine-learning, convolutional NNs, and RFs have been adopted for the implementation of integrated computational tools to predict ligand-binding affinity, relying on the atomic coordinates of protein-ligand complexes. Supervised machine-learning is applied for developing protein-targeted scoring functions for the prediction of binding affinity[41,42] (for an extensive description of recent docking methods, see Reference 40 and references therein).

The Protein Data Bank (PDB, https://www.rcsb.org/) is the main source for data with atomic resolution to ground our knowledge of PPIs. The current version (November 2021) contains 160,543 protein files, out of which about 60% contain complexes. Refinement resolution varies from <0.5 Å up to >4.5 Å, with most of the structures with average values of 2–2.5 Å. The gap with the number of sequences contained in UniProt (219,174,961) is still of three orders of magnitude. Specific and derived databases organize structures according to given properties,[39] like ProtCID, a data resource for structural information on protein interactions.[43] Furthermore, curated and processed small datasets are shared to enable benchmarking of novel methods.[17]

3.3.1 | Properties and representation of protein–protein interfaces

In ML, input encoding is an issue, given that a proper representation of the data is an important step for optimizing training sets and output results. This problem can be addressed by considering the biophysical properties of the protein interfaces as derived by a thorough analysis of the complexes known with atomic resolution in the PDB, when tackling the problem of Protein–Protein Interaction sites prediction.

Being proteins extremely heterogeneous molecules, with a large variety of binding affinity values, properties of protein–protein interfaces in PDB complexes (transient or obligatory) are different and often specific for a given set of complexes.[3,37]

In this respect, a main problem is the recognition of PDB functional protein–protein interfaces from nonspecific interactions due to the crystallization process and molecular packing into the unit crystal cell. In other words, not all the complexes in the PDB are functional, and this is should be taken into consideration when selecting protein sets for

ML training. A very general trend is that on average the size of the interface, measured as solvent accessible area buried upon complex formation, is larger in biological interfaces compared to crystallographic ones.[3] Apparently, also the residue composition of the biological interfaces differs from the crystallographic ones, enriching aliphatic and aromatic moieties. However, properties of nonfunctional interfaces seem to overlap with those of transient complexes and therefore a clear distinction is impossible based only on physicochemical and geometrical properties.[37]

Recently,[23] we analyzed a large data set of PDB complexes (19,360) from different organisms and downloaded with the constraints of being functional and solved with high resolution (in the range of 1–2.5 Å). We focused on the problem of distinguishing between homo- and heterointerfaces, finding that cysteine and to a lesser extent tryptophan are more prone to form interfaces in heterocomplexes. In turn, phenylalanine and leucine are more abundant in homointerfaces. Average areas of homo and heterointerfaces are about 3946 and 3551 Å,[2] respectively, confirming previous observations.[3,44]

Evolutionary conservation is another important feature which can help in the detection of functionally important residues, which are conserved in proteins forming complexes in related species.[3,23] Conservation can be estimated by means of multiple sequence/structural alignments (MSA). Shannon entropy and its variants are routinely adopted to score positional conservation for each column of the MSA.[23] Analyzing 9301 protein chains, we found that interface residues tend to be slightly more conserved than the other surface residues,[23] confirming previous observations in smaller data sets.

In general, conservation and composition alone are not sufficient to accurately discriminate interface residues from the remaining surface ones.

Interactions patches have been measured mainly on geometrical assumptions. A residue is defined accessible when endowed with a relative solvent accessibility higher than 20%. Once the monomer surface is computed, two main definitions of interface residues are widespread. The first, as mentioned above, is based on the different solvent accessibility between the bound (complex) and unbound (monomer). The second definition is based on the computation of inter-residue distances: interface residues are those having at least one residue of another subunit at a distance below a defined threshold (routinely between 5 and 8 Å).[23]

Recently an alternative general framework to learn protein surface fingerprints was introduced to perform a geometric deep learning. The method describes the geometric structure of the surface through its geometric features (shape index, distance-dependent curvature) and geodesic polar coordinates.[45]

All the varieties of interface properties and their representations are important when ML is applied and training is performed, as discussed in the following sections, were input features of the ML methods are listed.

4 | ML LEARNING APPROACHES FOR PPI PREDICTION

Routinely ML methods for PPI prediction take as input protein sequences or structures. Depending on the specific task at hand, we may group methods as sequence based and structure based.

4.1 | Expanding PPI networks

Considering the sequence-based methods we can start distinguishing methods that focus on expanding PPI networks (Section 3.1). Protein-level prediction of PPI refers to the problem of inferring an interaction score given a pair of putatively interacting proteins. Approaches in this field allow to extend the current knowledge on interaction networks by adding new edges to the graph.

Many computational tools have been devised for this task in the last decade (Table 1). These methods routinely start from a pair of protein sequences and produce as output a probability/score for their interaction.

A major issue is the representation of variable-length protein sequences. A task is the definition of a proper and effective procedure to transform input sequences of variable lengths into fixed-size vector encodings, to be then provided in input to the computational machinery. Several computational frameworks are now available for extracting complex information from protein sequences and profiles of interacting and noninteracting proteins.

Feature encodings adopted in this field include basic residue composition,[46] sequence profiles or Position-Specific Scoring Matrixes (PSSMs),[27,47,48] and residue physicochemical features.[28,46,49–52] In all cases, residue-level encodings are aggregated to obtain a fixed-size vector for the entire protein sequence. Methods to perform this aggregation are

TABLE 1 Expanding PPI networks with ML methods

Name	Year	Method details	Input features	Dataset/s	URL
EnAmDNN[53]	2020	Ensemble of Deep Neural Networks and Attention mechanisms	Autocovariance and Multiscale Local Descriptors,[50] pseudo residue composition.	Different data sets taken from IntAct (Section 3.1)	Web server not available
Yang et al.[28]	2020	Graph embeddings	Residue physicochemical properties transformed via Multiscale Local Descriptors[50] and autocorrelation.	Pan dataset[54]	Web server not available
CNN-FSRF[27]	2019	Convolutional Neural Networks + Random Forest	Position-specific scoring matrix.	Guo dataset derived from Data Base of Interactive Proteins (DIP)[55]	Web server not available
Lei et al.[56]	2019	Multimodal Deep Polynomial Networks	Amino acid mutation rate (BLOSUM62), hydrophobicity, and hydrophilicity.	Several PPI datasets from different species.	Web server not available
EnsDNN[52]	2019	Ensemble of 27 Deep Neural Networks	Residue physicochemical properties transformed via Multiscale Local Descriptors[50] and autocorrelation.	DeepPPI datasets[46]	Web server not available
DPPI[47]	2018	Deep Convolutional Networks	Sequence profiles.	Profppkernel benchmark datasets.[48]	https://github.com/hashemifar/DPPI/
DeepPPI[46]	2017	Deep Multi-Layer Perceptron	Amino acid composition, dipeptide composition, composition, transitions, and distributions of residue along the sequence, pseudo-amino acid composition.	A dataset derived from DIP; Eight different PPI datasets for evaluation.	Web server not available
Sun et al.[51]	2017	Stacked autoencoders + softmax classifier	Hydrophobicity, net charge of side chains; polarity, polarizability; solvent accessible area, volume of side chains. Fixed-size vector representation for each protein is obtained by auto-covariance and conjoint triad methods.	A positive dataset of proteins extracted from the human protein reference database. Negative examples are obtained by pairing proteins found in different subcellular compartments.	Web server not available
MLD-RF[50]	2015	Random Forest	Protein sequences are divided into a fixed number of non-overlapping regions. Each region is encoded with descriptors representing composition, transitions, and distributions of residue properties in the region.	Eight different PPI datasets from several organisms derived from DIP.	Web server not available
Profppikernel[48]	2015	Support Vector Machine with profile kernel[57]	Sequence profiles.	Park and Marcotte datasets.[58] A dataset of human PPI derived from the Hippie database[59]; a dataset of Yeast PPI derived from DIP.	https://rostlab.org/owiki/index.php/Profppikernel

TABLE 1 (Continued)

Name	Year	Method details	Input features	Dataset/s	URL
Bock & Gough[49]	2001	Support Vector Machine	Residue charge, hydrophobicity, and surface tension.	A positive dataset of 2664 proteins obtained from the DIP. Negative examples were obtained by random sampling of synthetic pairs from DIP.	Web server not available

different and range from simple averaging[49] to more sophisticated approaches based on autocovariance,[28,51,52] namely indexes considering correlations between residues at a certain distance apart in the sequence, and multiscale local descriptors which segment sequences into fixed-size nonoverlapping regions.[50]

Different machine-learning frameworks have been applied in this area. Early methods were based on traditional learning approaches such as SVMs[48,49] and RFs.[50] Recently, deep-learning approaches have been adopted, including deep fully-connected, multi-layer NNs,[46,52] stacked autoencoders,[51] convolutional NNs,[27,47] attention mechanisms,[53] and graph embeddings[28] (Table 1).

4.2 | ML approaches for PPI site prediction

PPI can be tackled by predicting protein interaction sites, taking as input sequence or structure (Figure 1). Routinely structure-based methods overperform sequence-based ones.

Machine-learning approaches (in particular, traditional ML methods) strongly rely on hand-crafted feature engineering and selection to perform the prediction task they are designed for. In the context of sequence- and structure-based PPI-site prediction methods, descriptors used to encode individual surface residues derive from a broad range of sources (Table 2).

4.2.1 | Input features for sequence-based approaches

Sequence-based methods (Table 2) only extract feature descriptors from the primary protein sequence. Descriptors routinely adopted in this area can be roughly classified into four major categories: (i) residue primary encoding; (ii) evolutionary information; (iii) residue physicochemical properties; and (iv) predicted structural features.

The residue one-hot representation (a vector with 19 zeros except one for the residue at hand) is routinely adopted for the basic encoding of the protein primary sequence in many tasks in computational biology. Despite its simplicity, one-hot encoding is not appropriate to capture important information encoded in protein evolution. To this aim, PPI-site prediction methods adopt richer protein encodings that rely on evolutionary information extracted from multiple sequence alignments (MSA), such as sequence profile and/or position-specific scoring matrices (PSSM) as well as different types of conservation scores.[60–64] Evolutionary descriptors are very informative but require the execution of computationally intensive alignments to find enough related sequences for the target protein. Recently, powerful techniques, traditionally adopted in the field of Natural Language Processing (NLP) to learn embedded representations of words and sentences of natural language, have been imported into the field of computational biology to learn embeddings for protein sequences.[65,66] Some of these approaches have been recently adopted also in the field of PPI site prediction.[60,67] These embeddings represent a trade-off between the simple one-hot encoding and the more informative but computational demanding evolutionary information.

Residue physicochemical properties such as residue hydrophobicity, charge, polarity, volume, and/or conformational propensities have been included in the input of many prediction methods in the last years.[60,62,67,68] These properties are routinely extracted from databases such as the AAindex[69] or obtained using dimensionality reduction procedures of precomputed residue properties.[70]

The absence of structural information is complemented in sequence-based approaches using predicted structural features including relative solvent accessibility,[60–62,64,67,71] secondary structure,[61,62,67] protein flexibility and disorder.[60]

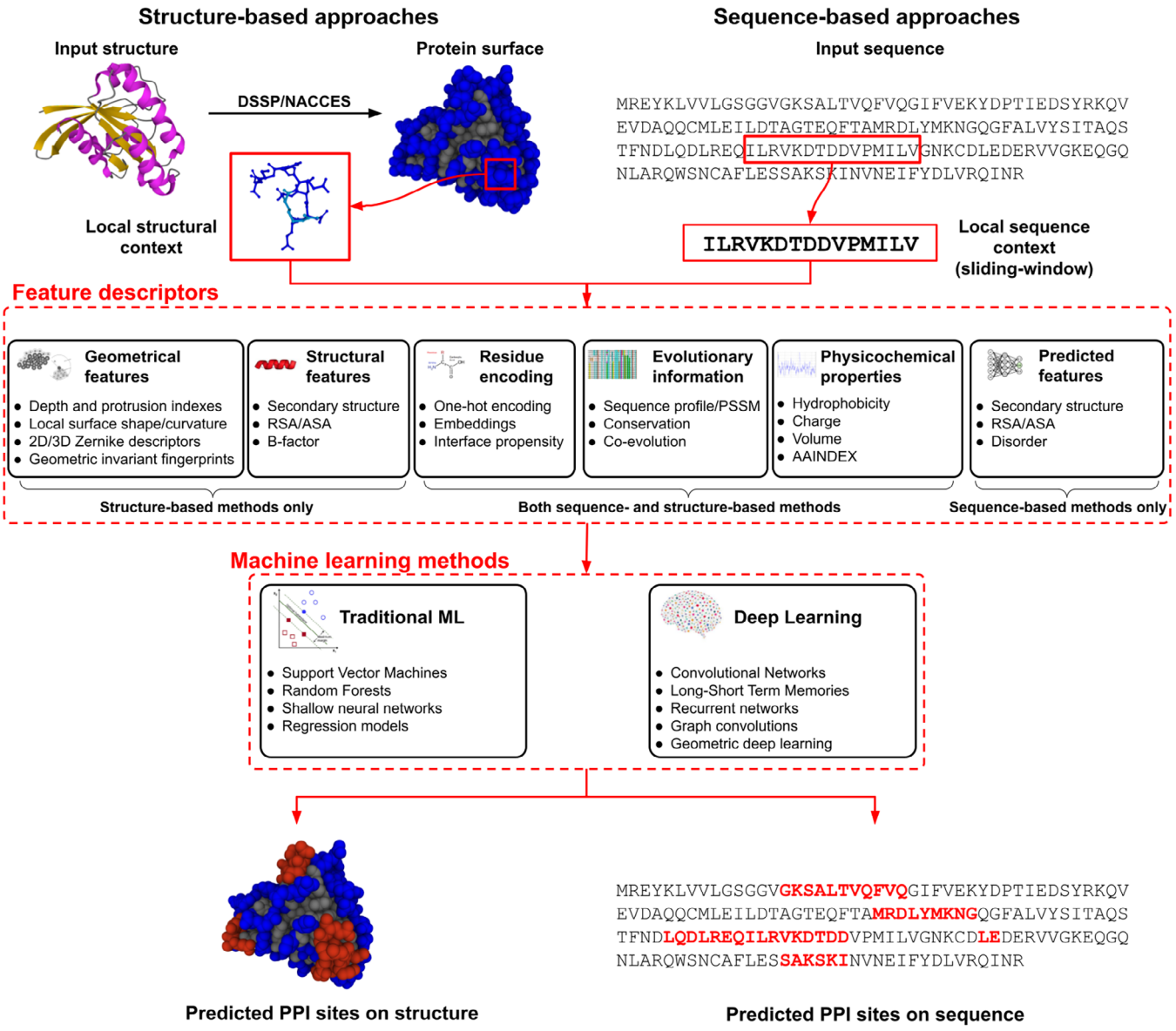

FIGURE 1 Schematic overview of ML methods for PPI-site prediction from structure and sequence

4.2.2 | Sequence-based ML methods

Many sequence-based approaches (Table 2) for the prediction of PPI sites are based on traditional machine-learning methods processing sequence features discussed above. ML techniques routinely applied include SVMs,[68,71] RFs or other tree-based approaches,[61,63,71] shallow NNs,[67] and simple regression algorithms.[62,64]

In some cases, the specific machine-learning approach is accompanied by other ML-based techniques for data preprocessing/balancing[68] and for automatic feature selection.[71]

A subset of sequence-based methods,[72,73] routinely partner-specific approaches that identify pairs of interacting residues between two input partners, are based on the analysis of protein coevolution. This is an unsupervised ML framework that, starting from MSA, attempts to detect putative interchain residue contacts analyzing the pattern of co-variation across protein–protein interfaces.

Recently, deep-learning methods appeared also in the field of PPI-site prediction from the sequence. Specifically, approaches that are well-suited for the analysis of sequence data, such as deep recurrent NNs and long-short term memory networks, have been applied.[60] Moreover, architectures based on convolutional NNs have been implemented for processing both local and global sequence contexts.[74]

TABLE 2 Sequence-based ML methods for PPI-site prediction

Name	Year	Partner specificity	Method details	Input features	Dataset/s (Dset, DB)	URL
DELPHI[60]	2020	No	Ensemble learning of convolutional and gated recurrent unit networks	3-mer amino acid embedding (ProtVec1D), residue position, position-specific scoring matrix, conservation, predicted relative solvent accessibility, interface propensity, predicted disorder, hydrophobicity, number of residue atoms, charge, potential hydrogen bonds, graph-shape index, polarizability, volume, isoelectric point, helix, and sheet probability.	ZhangDataset, Dset_448[52]; Dset_186, Dset_72[64]; Dset_164.[75]	https://delphi.csd.uwo.ca/
DeepPPISP[74]	2020	No	Convolutional Neural Networks	Position-specific scoring matrix, secondary structure, one-hot encoding.	Dset_186, Dset_72[64]; Dset_164.[75]	Web server not available
Wang et al.[68]	2020	No	Dataset balancing + Support Vector Machines	Sequence profile, profile entropy, conservation.	Dataset derived from the Ansari dataset.[76]	No web server available
ProNA2020[67]	2020		Neural Networks	Predicted secondary structure, predicted solvent accessibility, and physicochemical features.	Hamp dataset.[48]	http://www.predictprotein.org
SCRIBER[62]	2019	No	Multi-level Logistic Regression	Propensity for binding, predicted solvent accessibility, conservation, hydrophobicity, polarity, charge, predicted secondary structure, physicochemical properties, residue position.	ZhangDataset, Dset_448[52]	http://biomine.cs.vcu.edu/servers/SCRIBER/
SeRenDIP[61,77]	2019	No	Random Forest	Conservation, residue specificity in homodimers and monomers, sequence length, backbone dynamics, predicted solvent accessibility, and secondary structure.	Hou dataset[78]; Dset_186, Dset_72[64]	http://www.ibi.vu.nl/programs/serendipwww/

(Continues)

TABLE 2 (Continued)

Name	Year	Partner specificity	Method details	Input features	Dataset/s (Dset, DB)	URL
BIPSPI-sequence[63]	2018	Yes	XGBoost, tree bosting	Residue one-hot encoding, sequence profile, position-specific scoring matrix, conservation. Sliding window-based context.	Docking Benchmark v.5 (DBv5)[79]; DBv4[80]; DBv3[81]; CAPRI targets[82]; A dataset of 117 dimers (DImS).	http://bipspi.cnb.csic.es/xgbPredApp/
SSWRF[71]	2016	No	Random Forest + Support Vector Machines	Position-specific scoring matrix, hydrophobicity, predicted relative solvent accessibility.	Dset_186, Dset_72[64]; Dset_164[75]	http://202.119.84.36:3079/SSWRF-PPI/SSWRF-PPI.html
EVComplex[72]	2014	Yes	Direct coupling analysis based on mean field approximation	Multiple sequence alignment.	330 protein complexes extracted from in *E. coli*, literature-curated interactions and PDB[83]	https://evcouplings.org/complex
GREMLIN[73]	2014	Yes	Direct coupling analysis based on maximization of pseudo-likelihoods	Multiple sequence alignment.	18 protein complexes defined in this study.	http://gremlin.bakerlab.org/cplx_submit.php
PSIVER[64]	2010	No	Naïve Bayes classifier with kernel density estimation	Position-specific scoring matrix and predicted accessibility.	Two datasets comprising 186 and 72 heteromeric complexes (Dset_186 and Dset_72)	https://mizuguchilab.org/PSIVER/

4.2.3 | Input features for structure-based approaches

In structure-based approaches (Table 3), the availability of the protein structure allows the extraction of physicochemical and evolutionary features not only for a single surface residue but also considering its local surface structural context.[82,84]

An important class of features adopted by structure-based approaches fall in the category of geometrical descriptors extracted from the input protein structure. These include average depth[85] and protrusion[86] indexes computed over the set of atoms belonging to each surface residue, indexes describing the local surface shape[87] or curvature,[88] 2D/3D Zernike descriptors of voxelized protein surface representations[89,90] and geometric invariant fingerprint descriptors.[91]

Other common features extracted from the protein three-dimensional structure include measures of protein flexibility as derived from crystallographic B-factors, secondary structure motifs, and residue solvent accessibility.

For the same monomer, the value of the features can change if different conformations are considered. This is particularly relevant when addressing the problem of PPIs. Indeed, the structure of the isolated monomer (unbound structure) is in some cases very different from that of the same protein extracted from a complex (bound structure), because of the conformational rearrangements induced by the interaction. Therefore, using features extracted from bound instead of unbound monomers can introduce biases when predicting interaction sites.

TABLE 3 Structure-based ML methods for PPI-site prediction

Name	Year	Partner specificity	Method details	Input features	Dataset/s (Dset, DB)	URL
GraphPPIS[98]	2021	No	Deep Graph Convolutional Networks	Position-specific scoring matrix and hidden Markov model profile from multiple sequence alignment, secondary structure, torsion angles, relative solvent accessibility.	Dset_186, Dset_72[64]; Dset_164[75]; New dataset comprising 315 protein chains from PDB.	https://biomed.nscc-gz.cn/apps/GraphPPIS
MASIF-site[45]	2020	No	Geometric deep learning	Discretized representation of the protein surface into non-overlapping patches with fixed geodesic radius. For each vertex of the patches, geometrical features and chemical features are computed.	Combination of datasets extracted from the PRISM repository,[101] DBv5,[79] the PDBBind database,[102] and the SAbDab database.[103]	https://github.com/LPDI-EPFL/masif
IntPred[94]	2018	no	Random Forest	Interface propensity, hydrophobicity, conservation, disulfide bonds, hydrogen bonds, secondary structure, planarity.	Datasets derived from biological units available in PISA[104] (4345 and 4204 protein chains for training and testing).	http://www.bioinf.org.uk/intpred/
BIPSPI-structure[63]	2018	Yes	XGBoost, tree bosting	One-hot encoding, sequence profile, conservation, hydrophobicity, depth index, protrusion index, secondary structure, half-sphere exposure, and contact number. Structural context computed by means of Voronoi diagrams.[105]	DBv5[79]; DBv4[80]; DBv3[81]; Selection of CAPRI targets[82]; A new dataset of 117 dimers (DImS).	http://bipspi.cnb.csic.es/xgbPredApp/
SVM + 3D Zernike[89]	2018	No	Support Vector Machines + 3D Zernike descriptors[89]	Protein surface is described using 3D Zernike representation.[89] Residue features are extracted from AAindex[69]: alpha helix/beta-strands propensity, propensity for accessibility, volume, and optimized relative partition energies.	DBv5[79]	Web server not available
ISPRED4[82]	2017	No	Support Vector Machines + Conditional Random Fields	Sequence profile, conservation, interface propensity, physicochemical properties, mutual information, and PSICOV[106] coevolution score from multiple sequence alignments, depth, protrusion, secondary structure, B-factor, the difference between predicted and real relative solvent accessibility. Features averaged over a local structural context.	DBv5[79]; New dataset comprising a selection of CAPRI targets.	https://ispred4.biocomp.unibo.it

(Continues)

TABLE 3 (Continued)

Name	Year	Partner specificity	Method details	Input features	Dataset/s (Dset, DB)	URL
INSPiRE[107]	2017	No	Knowledge base of amino acids structural neighborhoods	Residue sequence.	Knowledge base built on top of the complete PDB database (release November 2015).	Web server not available
PrISE[108]	2012	No	Information transfer using local surface structural similarity	Residue identity, absolute solvent accessibility (single residue and structural context), atomic composition of the residue.	DS24Carl[109]; DS188, DS56bound and DS56unbound.[110]	http://ailab1.ist.psu.edu/prise/index.py
PresCont[92]	2012	No	Support Vector Machines	Relative solvent accessibility, interface planarity, hydrophobic patches, and residue conservation.	New datasets derived from PDB (PlaneDimers and Dimers); A subset of chains from DBv4[80]; Test cases from CAPRI.	http://www-bioinf.uni-regensburg.de/
SPPIDER[93]	2007	no	Support Vector Machines + Neural Networks	Residue properties (AAindex[69]): hydrophobicity, expected number of contacts. Features derived from multiple sequence alignments, position-specific scoring matrix, sequence profile, entropy, properties conservation. Structural features: number and distances from spatial surface neighbors, difference between predicted and real relative solvent accessibility.	S435 and S149 training and testing datasets defined in this study.	http://sppider.cchmc.org/

4.2.4 | Structure-based ML methods

In the last decade, the field of structure-based PPI site prediction has been dominated by traditional ML methods. The major difference with sequence-based methods is clearly the availability of the protein three-dimensional structure allowing to extract of very informative geometrical and structural features as described above. These descriptors, routinely computed for the subset of residues placed on the protein molecular surface, are then processed by traditional approaches including SVMs,[82,89,92,93] shallow NNs,[93] RFs,[63,94] and Markovian probabilistic graphical models such as hidden-Markov SVMs[95] and CRFs.[82,96,97] Markov models are well-suited for sequential data like protein sequences, being able to capture the potential relationships among adjacent residues in the protein surface,[82,97] mapped on the protein sequence.

Deep-learning has recently emerged in the field of structure-based PPI site predictors.[45,98] The main direction in this area involves the application of techniques under the umbrella of geometric deep learning approaches.[99] These approaches are useful for modeling data that cannot be easily represented into a standard Euclidean space, that is, data having an underlying non-Euclidean structure such as graphs or networks. The goal of geometric deep learning is to provide key basic operations that are at the basis of successful deep learning on Euclidean data (e.g., convolutional, or recurrent operations) also for the case of non-Euclidean data.

Recently,[45] an approach has been described based on a geometric invariant fingerprint[91] representation of the protein surface and the generalization of the standard convolution operator to protein surfaces described by means of a local geodesic polar system of coordinates. Briefly, a discretized representation of the surface is computed, and features assigned to each vertex of the resulting mesh. Then, for each vertex, a local patch is extracted with predefined geodesic radius. After patch extraction, the position of each vertex is mapped in radial and angular coordinates with respect to the center of the patch, adding information about spatial relationships between features. The canonical convolution operator is then generalized to this geodesic representation using a system of Gaussian kernels whose parameters are learned. These kernels act as the "filters" in the canonical convolutional layer.

Following this trend, another recent work[98] describes the application of graph convolutional networks for the prediction of PPI sites starting from protein structure. The protein molecular surface is represented as a graph where vertices are residues while edges highlight the proximity of two residues within a predefined distance threshold (C_α–C_α distance below 14 Å). On the resulting graph, different graph convolutional layers are applied in cascade, generalizing the basic convolution operator to graphs. After a cascade of N graph convolutions, the final layer is processed by a standard NN and transformed to per-residue interaction probabilities.

4.2.5 | A recent benchmark

A recent benchmark[98] compared sequence-based and structure-based methods on blind test sets, including different number of proteins ranging from 135 and 31, respectively. What is interesting in the benchmarking is the inclusion of SPPIDER,[93] a shallow learning-based method among ones based on deep learning. Scoring performances are lower for sequence-based than structure-based ones. However, among the structure-based ones SPPIDER is performing at the same level of MaSIF-site,[45] the deep learning method recently introduced that, in turn, is slightly outperformed by GraphPPIS[98] (Table 2). However, scoring values, measured as Matthews correlation coefficient and Area Under the Precision Recall Curve are about 0.3 and 0.4, respectively. Notwithstanding all the recent technological advancements, and rather independently of the method adopted, results of the benchmark[98] suggest that there is still large room for improvement, since the theoretical maximum for both scoring indexes is one.

4.3 | Recent advancements

ML generative models can efficiently explore subregions of the protein space to highlight sequence functional properties,[111] while the mathematical representation of biomolecular data can reduce ML dimensionality and simplify structural representation.[112]

With the advent of deep language models, and the concomitant explosive growth of available protein sequences, D-SCRIPT now associates genome to "phenome" with sequence-based, structure aware, and PPI proteome scale prediction.[113]

After AlphaFold2 models, the fraction of the dark structural proteome decreased from 26% to 10%, allowing a coverage increase of the critically important sets of disease-associated genes and mutations.[114]

One method[115] integrates information from three different levels, including protein sequence, distance map, and structure (similarly to AlphaFold2[21,22]) and enables rapid solutions of structural problems, including PPIs. When tested on a set comprising 68 protein complexes from *Escherichia coli*, known with atomic resolutions, the method satisfactory predicts some 43% of the known interfaces, and 82% of the associated protein structures.[115] Apparently, the procedure starting from protein sequence can bypass traditional approaches requiring modeling of individual subunits, followed by docking procedures.[115]

Furthermore, other papers still in bioRxiv, support the notion that the problem of predicting the interfaces of protein complexes, can take advantage of the deep learning-based methods included in AlphaFold2. AlphaFold-Multimer[116] filters 4433 recent protein complexes and produces high accuracy predictions of the interfaces in 23% of cases. Docking of protein models can be improved by adopting AlphaFold2 and a docking method (ClusPro)[117] and heterodimeric protein complex interactions can be better predicted by including AlphaFold2.[118]

All this work suggests that indeed the computational power of deep learning developments can efficiently extract information at different levels of our knowledge of PPI interaction. Presently, improved protein structure predictions seem to pave the way to improve PPI prediction methods.

4.4 | The issue of false positives

When evaluating scores, routinely ML methods compare their computed outputs with the expected ones. What are the expected ones? Based on structure complexes we can define known interactions as those that should be correctly predicted; however, given the scenario that is outlined in the introduction and the models at the proteomic scale of PPIs, it is difficult to estimate the real number of interactions that a protein can make in the crowded milieu of the cell.

For this reason, even benchmarking on very strict blind tests can be biased by wrong assumptions. The problem of false positives and their prediction is barely addressed. Recently when presenting, GraphPPIS,[98] authors discussed the problem by predicting unbound and bound structures for the same complexes, reaching the conclusion that the performance of methods trained on bound structures decreases when tested on unbound monomers. This suggests that long-range contacts are difficult to capture, and that the overall interaction surface is poorly represented, despite the accuracy in generating features for residues in contact.

We recently found that proteins in the Cajal granules have a number of interactors much larger than average in the human interactome (as reported in IntAct and BioGRID). The number of interactors moderately correlates with the number of residues predicted as flexible sites (with MobiDB)[100]; correlation increases when the number of predicted PPIs is considered (with ISPRED4, sequence based) and increases when the PPIs that are also flexible are retained. Apparently, the inherent flexibility of the residues may help in adjusting the interacting surface to multiple proteins.[23] This experiment confirms that flexibility, which is routinely considered associated with nonspecific interactions[5] can also integrate some functional PPIs. However, the property "being flexible" does not necessarily imply "being involved in a functional interaction site."[23]

5 | FINAL REMARKS

Despite large volumes of experimental data, advanced computational resources, and ML algorithms, we like to conclude that our knowledge of PPI is limited. Presently, it is difficult to solve the complexity of PPI in cells, considering the presence of both transient and nontransient aggregates, different compartments, and macromolecular condensates. Therefore, our PPI data at the proteomic level need still improvement to highlight all the possible interactions both in space and time. On the other hand, complexes solved with atomic resolution may represent overall only a limited set of the possible functional interactions that each protein can have in the crowded cell interior. However, it is very difficult to distinguish between functional and nonfunctional interaction surfaces. ML learning and particularly deep learning models are presently extremely successful in different research fields, including protein structure prediction. Still, when benchmarked on the task of PPI predictions, results, although promising, indicate limitations, including recent advancements. In relation to the problem of predicting PPI sites, no significant difference is found among shallow and deep learning,[98] suggesting that our representations of the interacting surfaces are still insufficient to capture all the

details of the binding affinities, which may differ depending on the cell type, requirement, and regulation. More examples, particularly protein complexes with high atomic resolution will eventually fill the gap among the potentiality of the methods and protein–protein interface description. Given this scenario, the problem of distinguishing with computational tools between functional and nonfunctional protein interactions remains open.

AUTHOR CONTRIBUTIONS

Rita Casadio: Conceptualization (equal); data curation (equal); formal analysis (equal); visualization (equal); writing – original draft (equal); writing – review and editing (equal). **Pier Luigi Martelli:** Conceptualization (equal); data curation (equal); formal analysis (equal); visualization (equal); writing – original draft (equal); writing – review and editing (equal). **Castrense Savojardo:** Conceptualization (equal); data curation (equal); formal analysis (equal); visualization (equal); writing – original draft (equal); writing – review and editing (equal).

ACKNOWLEDGMENTS

Open Access Funding provided by Universita degli Studi di Bologna within the CRUI-CARE Agreement. [Correction added on 18 May 2022, after first online publication: CRUI funding statement has been added.]

CONFLICT OF INTEREST

The authors have declared no conflicts of interest for this article.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Rita Casadio https://orcid.org/0000-0002-7462-7039
Pier Luigi Martelli https://orcid.org/0000-0002-0274-5669
Castrense Savojardo https://orcid.org/0000-0002-7359-0633

RELATED WIREs ARTICLES

Computational close up on protein-protein interactions: How to unravel the invisible using molecular dynamics simulations?
Computational prediction of protein-protein binding affinities

REFERENCES

1. Kessel A, Ben-Tal N. Introduction to proteins: structure, function, and motion. Mathematical and computational biology series. 2nd ed. Boca Raton, FL: CRC Press, Taylor & Francis Group, Chapman & Hall/CRC; 2018.p. 932.
2. Gerlich DW. Cell organization by liquid phase separation. Nat Rev Mol Cell Biol. 2017;18(10):593–3.
3. Boeynaems S, Alberti S, Fawzi NL, Mittag T, Polymenidou M, Rousseau F, et al. Protein phase separation: a new phase in cell biology. Trends Cell Biol. 2018;28(6):420–35.
4. Ditlev JA, Case LB, Rosen MK. Who's in and who's out—compositional control of biomolecular condensates. J Mol Biol. 2018;430(23): 4666–84.
5. Lyon AS, Peeples WB, Rosen MK. A framework for understanding the functions of biomolecular condensates across scales. Nat Rev Mol Cell Biol. 2021;22(3):215–35.
6. Alberti S, Hyman AA. Biomolecular condensates at the nexus of cellular stress, protein aggregation disease and ageing. Nat Rev Mol Cell Biol. 2021;22(3):196–213.
7. Lafontaine DLJ, Riback JA, Bascetin R, Brangwynne CP. The nucleolus as a multiphase liquid condensate. Nat Rev Mol Cell Biol. 2021;22(3):165–82.
8. Roden C, Gladfelter AS. RNA contributions to the form and function of biomolecular condensates. Nat Rev Mol Cell Biol. 2021;22(3): 183–95.
9. Lu H, Zhou Q, He J, Jiang Z, Peng C, Tong R, et al. Recent advances in the development of protein–protein interactions modulators: mechanisms and clinical trials. Sig Transduct Target Ther. 2020;5(1):213.
10. Feng Z, Jia B, Zhang M. Liquid–liquid phase separation in biology: specific stoichiometric molecular interactions vs promiscuous interactions mediated by disordered sequences. Biochemistry. 2021;60(31):2397–406.
11. Rattray DG, Foster LJ. Dynamics of protein complex components. Curr Opin Chem Biol. 2019;48:81–5.
12. Walport LJ, Low JKK, Matthews JM, Mackay JP. The characterization of protein interactions—what, how and how much? Chem Soc Rev. 2021;50:12292–307.

13. Baldi P. Deep learning in science. Cambridge: Cambridge University Press; 2021.p. 371.
14. Bishop CM. Pattern recognition and machine learning. Information science and statistics. New York: Springer; 2006.p. 738.
15. Greener JG, Kandathil SM, Moffat L, Jones DT. A guide to machine learning for biologists. Nat Rev Mol Cell Biol. 2022;23(1):40–55.
16. Goodfellow I, Bengio Y, Courville A. Deep learning. Adaptive computation and machine learning. Cambridge, MA: The MIT Press; 2016.p. 775.
17. Jamasb AR, Day B, Cangea C, Liò P, Blundell TL. Deep for protein–protein interaction site prediction. Methods Mol Biol. 2021;2361: 263–88.
18. Jones DT. Setting the standards for machine learning in biology. Nat Rev Mol Cell Biol. 2019;20(11):659–60.
19. Walsh I, Fishman D, Garcia-Gasulla D, Titma T, Pollastri G, ELIXIR Machine Learning Focus Group, et al. DOME: recommendations for supervised machine learning validation in biology. Nat Methods. 2021;18(10):1122–7.
20. Lensink MF, Brysbaert G, Mauri T, Nadzirin N, Velankar S, Chaleil RAG, et al. Prediction of protein assemblies, the next frontier: the CASP14-CAPRI experiment. Proteins. 2021;89:1800–23.
21. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583–9.
22. Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Žídek A, et al. Highly accurate protein structure prediction for the human proteome. Nature. 2021;596(7873):590–6.
23. Savojardo C, Martelli PL, Casadio R. Protein–protein interaction methods and protein phase separation. Annu Rev Biomed Data Sci. 2020;3(1):89–112.
24. Lee J, Hyeon DY, Hwang D. Single-cell multiomics: technologies and data analysis methods. Exp Mol Med. 2020;52(9):1428–42.
25. Shi Y, Yu Y, Zhou Y, Zhao J, Zhang W, Zou D, et al. A single cell interactome of human tooth germ from growing third molar elucidates signaling networks regulating dental development. Cell Biosci. 2021;11(1):178.
26. Gemovic B, Sumonja N, Davidovic R, Perovic V, Veljkovic N. Mapping of protein-protein interactions: web-based resources for revealing interactomes. Curr Med Chem. 2019;26(21):3890–910.
27. Wang L, Wang H-F, Liu S-R, Yan X, Song K-J. Predicting protein-protein interactions from matrix-based protein sequence using convolution neural network and feature-selective rotation Forest. Sci Rep. 2019;9(1):9848.
28. Yang F, Fan K, Song D, Lin H. Graph-based prediction of protein-protein interactions with attributed signed graph embedding. BMC Bioinformatics. 2020;21(1):323.
29. Kovács IA, Luck K, Spirohn K, Wang Y, Pollis C, Schlabach S, et al. Network-based prediction of protein interactions. Nat Commun. 2019;10(1):1240.
30. Chen Y, Wang W, Liu J, Feng J, Gong X. Protein Interface complementarity and gene duplication improve link prediction of protein-protein interaction network. Front Genet. 2020;11:291.
31. Paiano A, Margiotta A, De Luca M, Bucci C. Yeast two-hybrid assay to identify interacting proteins. Curr Protoc Protein Sci. 2019; 95(1):e70.
32. Dunham WH, Mullin M, Gingras A-C. Affinity-purification coupled to mass spectrometry: basic principles and strategies. Proteomics. 2012;12(10):1576–90.
33. McWhite CD, Papoulas O, Drew K, Dang V, Leggere JC, Sae-Lee W, et al. Co-fractionation/mass spectrometry to identify protein complexes. STAR Protocols. 2021;2(1):100370.
34. Low TY, Syafruddin SE, Mohtar MA, Vellaichamy A, Rahman NSA, Pung Y-F, et al. Recent progress in mass spectrometry-based strategies for elucidating protein–protein interactions. Cell Mol Life Sci. 2021;78(13):5325–39.
35. Wodak SJ, Vlasblom J, Turinsky AL, Pu S. Protein–protein interaction networks: the puzzling riches. Curr Opin Struct Biol. 2013;23(6): 941–53.
36. Bai X, McMullan G, Scheres SHW. How cryo-EM is revolutionizing structural biology. Trends Biochem Sci. 2015;40(1):49–57.
37. Schreiber G. CHAPTER 1 Protein–protein interaction interfaces and their functional implications. In: Roy S, Fu H, editors. Protein-protein interaction regulators. Washington, DC: Royal Society of Chemistry; 2020. p. 1–24.
38. Su M, Yang Q, Du Y, Feng G, Liu Z, Li Y, et al. Comparative assessment of scoring functions: the CASF-2016 update. J Chem Inf Model. 2019;59(2):895–913.
39. Jankauskaitė J, Jiménez-García B, Dapkūnas J, Fernández-Recio J, Moal IH. SKEMPI 2.0: an updated benchmark of changes in protein–protein binding energy, kinetics, and thermodynamics upon mutation. Xenarios I, editor. Bioinformatics. 2019;35(3):462–9.
40. Siebenmorgen T, Zacharias M. Computational prediction of protein–protein binding affinities. WIREs Comput Mol Sci. 2020;10:e1448.
41. Bitencourt-Ferreira G, de Azevedo WF. Machine learning to predict binding affinity. Methods Mol Biol. 2019;2053:251–73.
42. Bitencourt-Ferreira G, Rizzotto C, de Azevedo Junior WF. Machine learning-based scoring functions, development and applications with SAnDReS. Curr Med Chem. 2021;28(9):1746–56.
43. Xu Q, Dunbrack RL. ProtCID: a data resource for structural information on protein interactions. Nat Commun. 2020;11(1):711.
44. Janin J, Bahadur RP, Chakrabarti P. Protein–protein interaction and quaternary structure. Q Rev Biophys. 2008;41(2):133–80.
45. Gainza P, Sverrisson F, Monti F, Rodolà E, Boscaini D, Bronstein MM, et al. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. Nat Methods. 2020;17(2):184–92.
46. Du X, Sun S, Hu C, Yao Y, Yan Y, Zhang Y. DeepPPI: boosting prediction of protein–protein interactions with deep neural networks. J Chem Inf Model. 2017;57(6):1499–510.

47. Hashemifar S, Neyshabur B, Khan AA, Xu J. Predicting protein–protein interactions through sequence-based deep learning. Bioinformatics. 2018;34(17):i802–10.
48. Hamp T, Rost B. Evolutionary profiles improve protein–protein interaction prediction from sequence. Bioinformatics. 2015;31(12): 1945–50.
49. Bock JR, Gough DA. Predicting protein–protein interactions from primary structure. Bioinformatics. 2001;17(5):455–60.
50. You Z-H, Chan KCC, Hu P. Predicting protein-protein interactions from primary protein sequences using a novel multi-scale local feature representation scheme, and the random Forest. PLoS One. 2015;10(5):e0125811.
51. Sun T, Zhou B, Lai L, Pei J. Sequence-based prediction of protein protein interaction using a deep-learning algorithm. BMC Bioinformatics. 2017;18(1):277.
52. Zhang L, Yu G, Xia D, Wang J. Protein–protein interactions prediction based on ensemble deep neural networks. Neurocomputing. 2019;324:10–9.
53. Li F, Zhu F, Ling X, Liu Q. Protein interaction network reconstruction through ensemble deep learning with attention mechanism. Front Bioeng Biotechnol. 2020;8:390.
54. Pan X-Y, Zhang Y-N, Shen H-B. Large-scale prediction of human protein−protein interactions from amino acid sequence based on latent topic features. J Proteome Res. 2010;9(10):4992–5001.
55. Guo Y, Yu L, Wen Z, Li M. Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences. Nucleic Acids Res. 2008;36(9):3025–30.
56. Lei H, Wen Y, You Z, Elazab A, Tan E-L, Zhao Y, et al. Protein–protein interactions prediction via multimodal deep polynomial network and regularized extreme learning machine. IEEE J Biomed Health Inform. 2019;23(3):1290–303.
57. Kuang R, Ie E, Wang K, Wang K, Siddiqi M, Freund Y, et al. Profile-based string kernels for remote homology detection and motif extraction. J Bioinform Comput Biol. 2005;03(03):527–50.
58. Park Y, Marcotte EM. Flaws in evaluation schemes for pair-input computational predictions. Nat Methods. 2012;9(12):1134–6.
59. Schaefer MH, Fontaine J-F, Vinayagam A, Porras P, Wanker EE, Andrade-Navarro MA. HIPPIE: integrating protein interaction networks with experiment based quality scores. PLoS One. 2012;7(2):e31826.
60. Li Y, Golding GB, Ilie L. DELPHI: accurate deep ensemble model for protein interaction sites prediction. Bioinformatics. 2021;37(7): 896–904.
61. Hou Q, De Geest PFG, Griffioen CJ, Abeln S, Heringa J, Feenstra KA. SeRenDIP: SEquential REmasteriNg to DerIve profiles for fast and accurate predictions of PPI interface positions. Bioinformatics. 2019;35(22):4794–6.
62. Zhang J, Kurgan L. SCRIBER: accurate and partner type-specific prediction of protein-binding residues from proteins sequences. Bioinformatics. 2019;35(14):i343–53.
63. Sanchez-Garcia R, Sorzano COS, Carazo JM, Segura J. BIPSPI: a method for the prediction of partner-specific protein–protein interfaces. Bioinformatics. 2019;35(3):470–7.
64. Murakami Y, Mizuguchi K. Applying the Naïve Bayes classifier with kernel density estimation to the prediction of protein-protein interaction sites. Bioinformatics. 2010;26(15):1841–8.
65. Heinzinger M, Elnaggar A, Wang Y, Dallago C, Nechaev D, Matthes F, et al. Modeling aspects of the language of life through transfer-learning protein sequences. BMC Bioinformatics. 2019;20(1):723.
66. Yang KK, Wu Z, Bedbrook CN, Arnold FH. Learned protein embeddings for machine learning. Bioinformatics. 2018;34(15):2642–8.
67. Qiu J, Bernhofer M, Heinzinger M, Kemper S, Norambuena T, Melo F, et al. ProNA2020 predicts protein–DNA, protein–RNA, and protein–protein binding proteins and residues from sequence. J Mol Biol. 2020;432(7):2428–43.
68. Wang B, Mei C, Wang Y, Zhou Y, Cheng M-T, Zheng C-H, et al. Imbalance data processing strategy for protein interaction sites prediction. IEEE/ACM Trans Comput Biol and Bioinf. 2021;18(3):985–94.
69. Kawashima S. AAindex: amino acid index database. Nucleic Acids Res. 2000;28(1):374–4.
70. Kidera A, Konishi Y, Oka M, Ooi T, Scheraga HA. Statistical analysis of the physical properties of the 20 naturally occurring amino acids. J Protein Chem. 1985;4(1):23–55.
71. Wei Z-S, Han K, Yang J-Y, Shen H-B, Yu D-J. Protein-protein interaction sites prediction by Ensembling SVM and sample-weighted random forests. Neurocomput. 2016;193(C):201–12.
72. Hopf TA, Schärfe CPI, Rodrigues JPGLM, Green AG, Kohlbacher O, Sander C, et al. Sequence co-evolution gives 3D contacts and structures of protein complexes. Elife. 2014;3:e03430.
73. Ovchinnikov S, Kamisetty H, Baker D. Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. Elife. 2014;3:e02030.
74. Zeng M, Zhang F, Wu F-X, Li Y, Wang J, Li M. Protein–protein interaction site prediction through combining local and global features with deep neural networks. Bioinformatics. 2020;36(4):1114–20.
75. Dhole K, Singh G, Pai PP, Mondal S. Sequence-based prediction of protein-protein interaction sites with L1-logreg classifier. J Theor Biol. 2014;348:47–54.
76. Ansari S, Helms V. Statistical analysis of predominantly transient protein–protein interfaces. Proteins Struct Funct Bioinformatics. 2005;61(2):344–55.
77. Hou Q, De Geest PFG, Vranken WF, Heringa J, Feenstra KA. Seeing the trees through the forest: sequence-based homo- and heteromeric protein-protein interaction sites prediction using random forest. Bioinformatics. 2017;33(10):1479–87.

78. Hou Q, Dutilh BE, Huynen MA, Heringa J, Feenstra KA. Sequence specificity between interacting and non-interacting homologs identifies interface residues—a homodimer and monomer use case. BMC Bioinformatics. 2015;16(1):325.
79. Vreven T, Moal IH, Vangone A, Pierce BG, Kastritis PL, Torchala M, et al. Updates to the integrated protein-protein interaction benchmarks: docking benchmark version 5 and affinity benchmark version 2. J Mol Biol. 2015;427(19):3031–41.
80. Hwang H, Vreven T, Janin J, Weng Z. Protein-protein docking benchmark version 4.0. Proteins. 2010;78(15):3111–4.
81. Hwang H, Pierce B, Mintseris J, Janin J, Weng Z. Protein–protein docking benchmark version 3.0. Proteins Struct Funct Bioinformatics. 2008;73(3):705–9.
82. 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.
83. Rajagopala SV, Sikorski P, Kumar A, Mosca R, Vlasblom J, Arnold R, et al. The binary protein-protein interaction landscape of *Escherichia coli*. Nat Biotechnol. 2014;32(3):285–90.
84. Bendell CJ, Liu S, Aumentado-Armstrong T, Istrate B, Cernek PT, Khan S, et al. Transient protein-protein interface prediction: datasets, features, algorithms, and the RAD-T predictor. BMC Bioinformatics. 2014;15(1):82.
85. Pintar A, Carugo O, Pongor S. Atom depth as a descriptor of the protein interior. Biophys J. 2003;84(4):2553–61.
86. Pintar A, Carugo O, Pongor S. CX, an algorithm that identifies protruding atoms in proteins. Bioinformatics. 2002;18(7):980–4.
87. Koenderink JJ. Solid shape. Artificial intelligence. Cambridge, MA: MIT Press; 1990.p. 699.
88. Duncan BS, Olson AJ. Shape analysis of molecular surfaces. Biopolymers. 1993;33(2):231–8.
89. Daberdaku S, Ferrari C. Exploring the potential of 3D Zernike descriptors and SVM for protein–protein interface prediction. BMC Bioinformatics. 2018;19(1):35.
90. Milanetti E, Miotto M, Di Rienzo L, Monti M, Gosti G, Ruocco G. 2D Zernike polynomial expansion: finding the protein-protein binding regions. Comput Struct Biotechnol J. 2021;19:29–36.
91. Yin S, Proctor EA, Lugovskoy AA, Dokholyan NV. Fast screening of protein surfaces using geometric invariant fingerprints. Proc Natl Acad Sci USA. 2009;106(39):16622–6.
92. Zellner H, Staudigel M, Trenner T, Bittkowski M, Wolowski V, Icking C, et al. Prescont: predicting protein-protein interfaces utilizing four residue properties: predicting protein-protein interfaces. Proteins. 2012;80(1):154–68.
93. Porollo A, Meller J. Prediction-based fingerprints of protein-protein interactions. Proteins. 2007;66(3):630–45.
94. Northey TC, Barešić A, Martin ACR. IntPred: a structure-based predictor of protein–protein interaction sites. Bioinformatics. 2018;34(2):223–9.
95. Liu B, Wang X, Lin L, Tang B, Dong Q, Wang X. Prediction of protein binding sites in protein structures using hidden Markov support vector machine. BMC Bioinformatics. 2009;10(1):381.
96. Dong Z, Wang K, Linh Dang TK, Gültas M, Welter M, Wierschin T, et al. CRF-based models of protein surfaces improve protein-protein interaction site predictions. BMC Bioinformatics. 2014;15(1):277.
97. Li M-H, Lin L, Wang X-L, Liu T. Protein–protein interaction site prediction based on conditional random fields. Bioinformatics. 2007;23(5):597–604.
98. Yuan Q, Chen J, Zhao H, Zhou Y, Yang Y. Structure-aware protein–protein interaction site prediction using deep graph convolutional network. Bioinformatics. 2021;38(1):125–136.
99. Bronstein MM, Bruna J, LeCun Y, Szlam A, Vandergheynst P. Geometric deep learning: going beyond Euclidean data. IEEE Signal Processing Magazine. 2017;34(4):18–42.
100. Piovesan D, Necci M, Escobedo N, Monzon AM, Hatos A, Mičetić I, et al. MobiDB: intrinsically disordered proteins in 2021. Nucleic Acids Res. 2021;49(D1):D361–7.
101. Baspinar A, Cukuroglu E, Nussinov R, Keskin O, Gursoy A. PRISM: a web server and repository for prediction of protein–protein interactions and modeling their 3D complexes. Nucleic Acids Res. 2014;42(W1):W285–9.
102. Liu Z, Li Y, Han L, Li J, Liu J, Zhao Z, et al. PDB-wide collection of binding data: current status of the PDBbind database. Bioinformatics. 2015;31(3):405–12.
103. Dunbar J, Krawczyk K, Leem J, Baker T, Fuchs A, Georges G, et al. SAbDab: the structural antibody database. Nucleic Acids Res. 2014;42(D1):D1140–6.
104. Krissinel E, Henrick K. Inference of macromolecular assemblies from crystalline state. J Mol Biol. 2007;372(3):774–97.
105. Segura J, Jones PF, Fernandez-Fuentes N. Improving the prediction of protein binding sites by combining heterogeneous data and Voronoi diagrams. BMC Bioinformatics. 2011;12(1):352.
106. 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(2):184–90.
107. Jelínek J, Škoda P, Hoksza D. Utilizing knowledge base of amino acids structural neighborhoods to predict protein-protein interaction sites. BMC Bioinformatics. 2017;18(S15):492.
108. Jordan RA, El-Manzalawy Y, Dobbs D, Honavar V. Predicting protein-protein interface residues using local surface structural similarity. BMC Bioinformatics. 2012;13:41.
109. Carl N, Konc J, Janežič D. Protein surface conservation in binding sites. J Chem Inf Model. 2008;48(6):1279–86.
110. Zhang QC, Petrey D, Norel R, Honig BH. Protein interface conservation across structure space. Proc Natl Acad Sci USA. 2010;107(24):10896–901.

111. Wittmann BJ, Johnston K, Wu Z, Arnold FH. Advances in machine learning for directed evolution. Curr Opin Struct Biol. 2021;69: 11–8.
112. Nguyen DD, Cang Z, Wei G-W. A review of mathematical representations of biomolecular data. Phys Chem Chem Phys. 2020;22: 4343–67.
113. Sledzieski S, Singh R, Cowen L, Berger B. D-SCRIPT translates genome to phenome with sequence-based, structure-aware, genome-scale predictions of protein-protein interactions. Cell Syst. 2021;12:969–982.e6.
114. Porta-Pardo E, Ruiz-Serra V, Valentini S, Valencia A. The structural coverage of the human proteome before and after AlphaFold. PLoS Comput Biol. 2022;18:e1009818.
115. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science. 2021;373:871–6.
116. Evans R, O'Neill M, Pritzel A, Antropova N, Senior A, Green T, et al. Protein complex prediction with AlphaFold-multimer. bioRxiv. 2021. https://doi.org/10.1101/2021.10.04.463034v1
117. Ghani U, Desta I, Jindal A, Khan O, Jones G, Kotelnikov S, et al. Improved docking of protein models by a combination of Alphafold2 and ClusPro. bioRxiv. 2021. https://doi.org/10.1101/2021.09.07.459290v1
118. Bryant P, Pozzati G, Elofsson A. Improved prediction of protein-protein interactions using AlphaFold2. Nat Commun. 2021;13(1):1265.

How to cite this article: Casadio R, Martelli PL, Savojardo C. Machine learning solutions for predicting protein–protein interactions. WIREs Comput Mol Sci. 2022;12:e1618. https://doi.org/10.1002/wcms.1618