Computational and Structural Biotechnology Journal 20 (2022) 5316–5341

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

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

Review

Protein–protein interaction prediction with deep learning: A comprehensive review

Farzan Soleymani [a], Eric Paquet [b,*], Herna Viktor [c], Wojtek Michalowski [d], Davide Spinello [a]

[a] Department of Mechanical Engineering, University of Ottawa, Ottawa, ON, Canada
[b] National Research Council, 1200 Montreal Road, Ottawa, ON K1A 0R6, Canada
[c] School of Electrical Engineering and Computer Science, University of Ottawa, ON, Canada
[d] Telfer School of Management, University of Ottawa, ON, K1N 6N5, Canada

ARTICLE INFO

Article history:
Received 22 June 2022
Received in revised form 29 August 2022
Accepted 30 August 2022
Available online 19 September 2022

Keywords:
Protein–protein interaction
Deep learning
Protein design
Sequence-based
Structure-based

ABSTRACT

Most proteins perform their biological function by interacting with themselves or other molecules. Thus, one may obtain biological insights into protein functions, disease prevalence, and therapy development by identifying protein–protein interactions (PPI). However, finding the interacting and non-interacting protein pairs through experimental approaches is labour-intensive and time-consuming, owing to the variety of proteins. Hence, protein–protein interaction and protein–ligand binding problems have drawn attention in the fields of bioinformatics and computer-aided drug discovery. Deep learning methods paved the way for scientists to predict the 3-D structure of proteins from genomes, predict the functions and attributes of a protein, and modify and design new proteins to provide desired functions. This review focuses on recent deep learning methods applied to problems including predicting protein functions, protein–protein interaction and their sites, protein–ligand binding, and protein design.
Crown Copyright © 2022 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-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Contents

* Corresponding author.
E-mail addresses: fsoleyma@uottawa.ca (F. Soleymani), Eric.Paquet@nrc-cnrc.gc.ca (E. Paquet), hviktor@uottawa.ca (H. Viktor), wojtek@telfer.uottawa.ca (W. Michalowski), dspinell@uottawa.ca (D. Spinello).

https://doi.org/10.1016/j.csbj.2022.08.070
2001-0370/Crown Copyright © 2022 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-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Proteins are organic molecules abundant in living systems and conduct a wide range of unique functions such as transport, storage, membrane composition, and enzymatic action [1], among others. Proteins may interact with DNA, RNA, ligands, and other proteins to carry out cellular and biological functions [2]. The latter occurs by physical interaction between two or more proteins [3,4]. These interactions ought to comply with two conditions: first, the interaction must be by design, i.e. the result of a specific biomolecular event; second, the interaction has evolved to serve a certain non-generic function [3–5]. Thus, one may obtain biological insights into protein functions, disease prevalence, and therapy development by identifying interaction amongst protein pairs [6–8]. Hence, protein–protein interaction (PPI) and protein–ligand binding problems have drawn attention in bioinformatics and computer-aided drug discovery [7,9,10]. Computational methods paved the way for scientists to predict the 3-D structures of proteins from genomes and, hence, to predict their functions and attributes, allowing them to modify proteins and design new ones to target desired functions. However, experimental validation benchmarking remains challenging [11].

Protein–protein interactions compose complexes to conduct numerous biological processes and functions such as metabolic cycles, signal transduction, DNA transcription and replication, catalysis, and immune response [12–18]. The activities of cells and their functions are affected by abnormalities in protein interactions, leading to numerous diseases such as cancer and chronic degenerative diseases [19]. Comprehensive identification of PPIs can help to decode the molecular mechanisms of the specific biological functions involved [19]. The proximity of proteins in PPI is of paramount importance for specific functionality. Despite significant efforts in molecular biology and genomics, the functions of most proteins are not yet established [20–22]. It has been demonstrated by Jansen et al. [23] that the interaction between known and unknown functional proteins can significantly contribute toward deciphering many protein functions. Therefore, predicting PPIs has become a crucial challenge in the field of bioinformatics [19].

PPIs may help in decoding the functionality of unannotated proteins [19,24]. Therefore, many experimental studies have been conducted to identify PPI, among which the yeast two–hybrid [25–27], mass spectrometry [28–32], protein microarrays [33–36] are often used [37,38]. However, these approaches are laborious and time-consuming, which makes them difficult to employ for all protein pairs [9,39,40]. Moreover, the validity of the experimental techniques is highly dependent on how well one implements the assay protocols in target organisms [41]. Therefore, one may use computational methods as pre-treatment in advance of the experimental methods, aiming to reduce false-positive and false-negative results [24,41,42].

A protein comprises a unique linear sequence of amino acids called its primary structure, which determines the folded shape or conformation. The local secondary structure elements, such as strands, helices, and random coils, are created as a result of interactions between the protein backbone, the side-chains, and the environment and extended to the ultimate 3-D structure of the protein [43]. The large number of possible configurations of the peptide backbone, and the desirable chemical bonding geometry and interactions, make the problem of modelling protein structures challenging [17]. This paper reviews recent advances in deep learning methods developed and/or applied to problems, including predicting protein functions, protein–protein interactions and their sites, protein–ligand binding, and protein design.

This review is structured as follows. We first outline protein structure architectures in Section 2. Next, we describe protein shapes in Section 3. Then, we present some of the main resources for protein structures and sequences in Section 4. Section 5 briefly explains some of the most commonly used deep learning methods, and Section 6 provides an overview of PPI prediction. Section 6.2 and Section 6.3 discuss structure-based PPI prediction methods and their computational solutions, respectively. Sequence-based PPI prediction methods and their associated computational solutions are described in Section 6.4 and Section 6.5. Section 7 reviews deep learning methods addressing protein design problem and Section 8 concludes the review.

2. Protein Structure

Proteins are a broad class of biomolecules forming more than 50% of the dry weight of cells [44]. Their diverse functionality and abundance determine the function and structure of cells, with each protein being an agent performing a specific biological role [44]. Genes are the basic physical and functional units of inheritance and act as instructions to create the proteins that are the agents of biological function. In fact, a unique protein structure is encoded by each gene in cellular DNA, which leads to numerous possible structures [1]. The uniqueness of proteins originates in the amino acid sequences and the bonds that hold them together.

The interaction between proteins is mainly non-covalent [43] except for covalent disulfide bonds (formed by the coupling of two thiol (–SH) groups), between the cysteine amino acid residues of the interacting partner proteins. Hydrogen bonding between proteins in a specific PPI is the most important type of non-covalent interaction.

The main and side-chain atoms of the different amino acid residues are involved in the hydrogen bonding between interacting protein partners. The ion pairs, which form mainly between an acidic and a basic amino acid in the proteins, form the second most important non-covalent interaction between protein partners [45]. The stability of protein structures is also affected by long-range interactions. The impact of short, medium and long-range interactions on various structural classes of proteins are discussed in [46–48].

As stated in [47], the all-α protein class, i.e., the proteins whose secondary structure is completely formed by α-helices apart from a few β-sheets on the edges[49], are governed by medium-range interactions. In contrast, long-range interactions dominate in all-β proteins, in which the secondary structure is mainly composed of β-sheets aside from some α-helices on the edges [49].

The primary to quaternary protein structures are examined in more detail in the following sections.

2.1. Primary structure

The primary structure, as shown in Fig. 1, is a unique, linear, amino acid sequence that forms the backbone of protein. Intramolecular bonding and folding of the linear amino acid chain

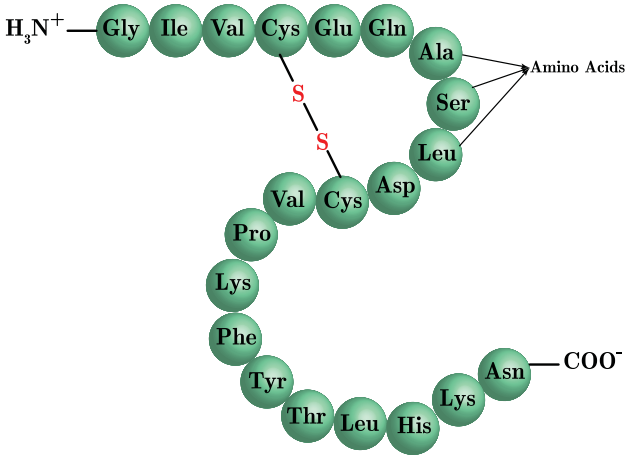

Fig. 1. primary structure.

eventually establish the protein's three-dimensional shape. The sequence of protein is determined by the gene encoding it, so changing the gene's DNA sequence may alter the protein's amino acid sequence and, thus, the protein's overall structure and function.

Amino acids, the building blocks of proteins, are small organic molecules composed of a central carbon atom, called the α-carbon, attached to an amino group (-NH_2), a carboxyl group (-COOH), and a hydrogen atom [50]. The carboxyl group is typically deprotonated and carries a negative charge at physiological pH (7.2–7.4) [51], whereas the amino group is typically protonated and shows a positive charge. The identity of each amino acid depends on its R group, which is an atom or group attached to the central atom. For example, the R group of glycine, as shown in Fig. 2 is a hydrogen atom, while the R group of alanine is a methyl group (-CH_3). Fig. 2 illustrates the twenty common amino acids, each of which has a unique side chain. The side chains govern each acid's chemical behaviour, e.g. whether it is acidic, basic, polar, or nonpolar. Nonpolar amino acids contain aliphatic (hydrocarbon) chains, while polar neutral amino acids contain a hydroxyl (-OH), sulfur, or amide in the R group. Polar acidic amino acids have a carboxylic acid group in the side chain, in addition to the one in the backbone. Polar basic amino acids contain an amine group (which may be neutral or charged) in the side chain, in addition to that in the backbone.

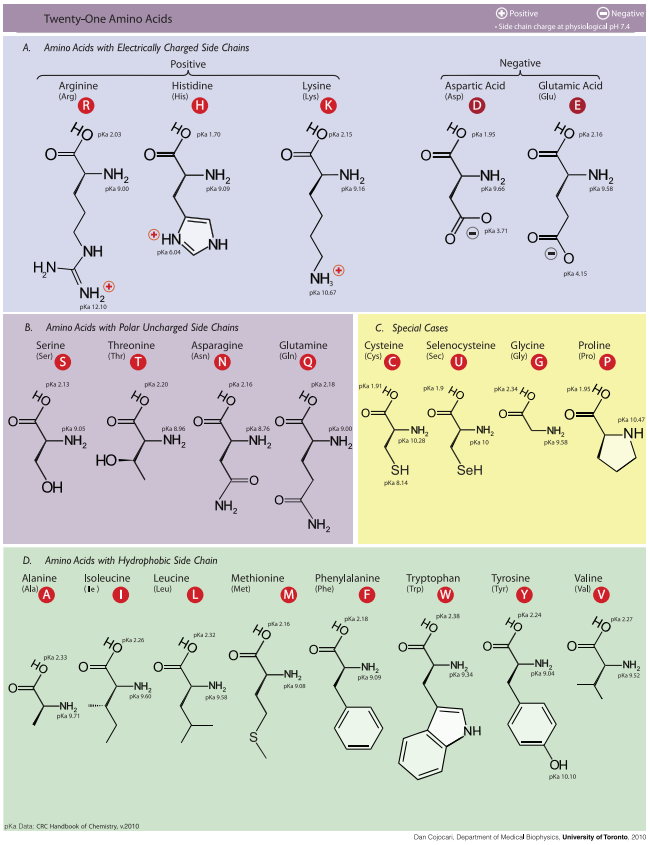

Fig. 2. Amino acids.

The physicochemical properties of 20 common amino acids are reported in Table 1. SASA represents the solvent-accessible surface area, and the side-chain net charge number is given by NCN [52]. These properties help determine the feasibility of protein's interac-

Table 1
Physicochemical properties of 20 amino acids. Column (a) steric parameters (graph shape index) [54,57]; (b) volume; (c) isoelectric point; (d) helix probability [58]; (e) sheet probability [58]; (f) hydrophobicity [59]; (g) hydrophilicity [59]; (h) side-chain residue size [52,54,60]; (i) polarity [52]; (j) polarizability [52]); (SASA) solvent-accessible surface area; (NCN) net charge number [52].

Amino acid	Symbol	a	b	c	d	e	f	g	h	i	j	SASA	NCN
Alanine	A	1.28	1.00	6.11	0.42	0.23	0.62	−0.50	27.50	8.10	0.046	1.181	0.007187
Cysteine	C	1.77	2.43	6.35	0.17	0.41	0.29	−1.00	44.60	5.50	0.128	1.461	−0.03661
Aspartate	D	1.60	2.78	2.95	0.25	0.20	−0.90	3.00	40.00	13.00	0.105	1.587	−0.02382
Glutamate	E	1.56	3.78	3.09	0.42	0.21	−0.74	3.00	62.00	12.30	0.151	1.862	0.006802
Phenylalanine	F	2.94	5.89	5.67	0.30	0.38	1.19	−2.50	115.50	5.20	0.29	2.228	0.037552
Glycine	G	0.00	0.00	6.07	0.13	0.15	0.48	0.00	0.00	9.00	0.00	0.881	0.179052
Histidine	H	2.99	4.66	7.69	0.27	0.30	−0.40	−0.50	79.00	10.40	0.23	2.025	−0.01069
Isoleucine	I	4.19	4.00	6.04	0.30	0.45	1.38	−1.80	93.50	5.20	0.186	1.81	0.021631
Lysine	K	1.89	4.77	9.99	0.32	0.27	−1.50	3.00	100.00	11.30	0.219	2.258	0.017708
Leucine	L	2.59	4.00	6.04	0.39	0.31	1.06	−1.80	93.50	4.90	0.186	1.931	0.051672
Methionine	M	2.35	4.43	5.71	0.38	0.32	0.64	−1.30	94.10	5.70	0.221	2.034	0.002683
Asparagine	N	1.60	2.95	6.52	0.21	0.22	−0.78	2.00	58.70	11.60	0.134	1.655	0.005392
Proline	P	2.67	2.72	6.80	0.13	0.34	0.12	0.00	41.90	8.00	0.131	1.468	0.23953
Glutamine	Q	1.56	3.95	5.65	0.36	0.25	−0.85	0.20	80.70	10.50	0.18	1.932	0.049211
Arginine	R	2.34	6.13	10.74	0.36	0.25	−2.53	3.00	105.00	10.50	0.291	2.56	0.043587
Serine	S	1.31	1.60	5.70	0.20	0.28	−0.18	0.30	29.30	9.20	0.062	1.298	0.004627
Threonine	T	3.03	2.60	5.60	0.21	0.36	−0.05	−0.40	51.30	8.60	0.108	1.525	0.003352
Valine	V	3.67	3.00	6.02	0.27	0.49	1.08	−1.50	71.50	5.90	0.14	1.645	0.057004
Tryptophan	W	3.21	8.08	5.94	0.32	0.42	0.81	−3.40	145.50	5.40	0.409	2.663	0.037977
Tyrosine	Y	2.94	6.47	5.66	0.25	0.41	0.26	−2.30	117.30	6.20	0.298	2.368	0.023599

Table 2
The nonstandard amino acids [65–67].

Name	Symbol	Abbr
Aspartic acid or Asparagine	B	Asx
Leucine or Isoleucine	J	Xle
Pyrrolysine	O	Pyl
Selenocysteine	U	Sec
Glutamic acid or Glutamine	Z	Glx
unknown amino acid	X	Unk

Table 3
amino acids classified by side chain properties.

Charge	Positive	H, K, R
	Negative	D, E
	Neutral	A, C, N, P, Q, S, F, G, I, L, M, T, V, W
Polarity	Polar	Y
	Nonpolar	C, D, E, H, K, N, Q, R, S, T
Aromaticity	Aliphatic	A, F, G, I, L, M, P, V, W
	Aromatic	I, L, V
	Neutral	F, H, W, Y
Size	Small	A, C, D, E, Q, R, S, G, K, M, N, P, T
	Medium	A, G, P, S
	Large	D, N, T

tion [52]. The physicochemical attributes of amino acids, such as hydropathy [53], isoelectric, the pH at which the molecule carries no net charge [54,55], and charge, play crucial roles in identifying the interaction between protein sequences [56].

Beyond the common amino acids shown in Table 1, there are also nonstandard amino acids [61]. These are also known as biosynthetic amino acids, and require complex synthetic and translational mechanisms that differ from the canonical enzymatic system used for the 20 standard amino acids [62]) namely, pyrrolysine [63] and selenocysteine [64]. These nonstandard amino acids are presented in Table 2. Sometimes it is not possible to differentiate two closely related amino acids. Therefore, we have the indeterminate residues in protein sequences as represented by symbols B, J, Z and X.

One way to classify amino acids is based on the side chains, as shown in Table 3 [68,69], in which case.

Multiple amino acids are linked together by peptide bonds, forming a long chain called the polypeptide. The order of the amino acids determines the polypeptide's functionality. Polypeptides are classified by the number of amino acid units in the chain. Each amino acid is linked covalently to its neighbours by peptide bonds, in a dehydration synthesis (condensation) reaction. Each protein is composed of one or more polypeptide chains. During protein synthesis, the carboxyl group (-COOH) of the amino acid at the end of the growing polypeptide chain reacts with the amino group of an incoming amino acid, forging a peptide bond and releasing a water molecule. Peptide bonds connect the carbon of the carboxyl group of one amino acid to the nitrogen of the amino group of the next, as shown in Fig. 3.

Table 4
Databases for PPI prediction

Type	Database	Description	Last update	URL
Protein–Protein Interactions	STRING [109]	Functional associations between protein pairs, which contains 67,592,464 proteins from 14094 organisms; 20,052,394,042 interactions.	2021	https://string-db.org/
	IntAct [116]	Contains manually curated datasets (topical), interactomes (for 16 different species) and annotations of experimental evidence.	2021	https://www.ebi.ac.uk/intact/home
	Biogrid [111]	Contains 2,467,140 protein and genetic interactions, 29,417 chemical interactions and 1,128,339 PTMs from major model organism species.	2020	http://www.thebiogrid.org/
	DIP [112]	Experimentally determined PPI database including biological information of proteins, PPIs and experimental techniques for identifying interactions.	2020	https://dip.doe-mbi.ucla.edu/dip/Main.cgi
	Negatome 2.0 [108]	Contains 21,795 interactions, with scores of zero and one, using text mining from literature and analysing protein complexes from PDB.	2014	http://mips.helmholtz-muenchen.de/proj/ppi/negatome/
	MINT [114]	Experimentally curated PPI database that includes approximately 117001 PPIs from 607 different species.	2012	https://mint.bio.uniroma2.it/
	HPRD [115]	Consists of 41,327 PPIs, 93,710 PTMs, 22,490 Subcellular Localizations and 112,158 Protein Expressions.	2010	http://www.hprd.org
	BIND [113]	PPIs collected from of humans, yeasts, nematodes, etc.	2005	http://download.baderlab.org/BINDTranslation
Protein sequences	UniProt [106]	A collection of protein sequence and functional information, including UniProtKB, UniParc, UniRef and Proteomes. UniProtKB contains 567,483 reviewed (Swiss-Prot)—manually annotated, and 231,354,261 unreviewed (TrEMBL)—computationally analysed, protein sequences.	2020	http://www.uniprot.org
	SWISS-MODEL [107]	A web-based integrated service providing information for protein structure homology modelling. The repository contains 2,217,470 models from SWISS-MODEL for UniProtKB targets, as well as 180,107 structures from PDB with mapping to UniProtKB.	2020	https://swissmodel.expasy.org/
	PIR [119]	Integrated protein resources, including protein sequences and high-quality annotations by integrating more than 90 biological databases.	2022	http://pir.georgetown.edu/
Higher-level structures	RCSB PDB [110]	Information about the 3-D structure of proteins, nucleic acids, and complex assemblies. 191144 structures, 57349 human sequence structures, and 14406 nucleic acid-containing Structures	2021	https://www.rcsb.org/
	SCOP [122]	Classification of known proteins and a comprehensive description of the structural and evolutionary relationships between them. As of 2022-05-30, this dataset contains 72,448 non-redundant domains, representing 858,316 protein structures.	2022	http://scop.mrc-lmb.cam.ac.uk/scop
Genomic information	CGD [124]	A resource for genomic sequence data, genes and protein information for *Candida albicans* and related species.	2022	http://www.candidagenome.org/

Polypeptide chains are directional, i.e. its ends are chemically distinct from one another. The end with a free amino group is called the amino terminus or N-terminus, while the other end has a free carboxyl group, and is known as the carboxyl terminus or C-terminus (see Fig. 3).

Most of the side chains are nonpolar, several are positively or negatively charged, some are polar but not charged. These features, and their consequent bonds, are responsible for protein structure and functionality by maintaining the protein in a specific shape or conformation. The polar side chains can form hydrogen bonds, while the charged side chains can form ionic bonds. Hydrophobic side chains interact via van der Waals interactions [1]. Consequently, protein folding is directed by the side-chain interactions, the sequence and the location of amino acids in that protein. The order of the acids, i.e. the primary structure, determines which bond types can form at each location along the polypeptide, and thus governs the protein's tertiary structures [70].

2.2. Secondary structure

Secondary structures result from interactions between parts of the polypeptide chain. The most common folding patterns are α-helices and β-pleated sheets [44].

In an α-helix, the hydrogen bonding occurs between the carbonyl group (C = O) of one amino acid and the hydrogen atom of the amino acid four places further along the chain. This bonding pattern draws the polypeptide chain into a helix, with each turn of containing 3.6 amino acids. The R groups stick outwards from the α-helix, and are free to interact. In a β-pleated sheet, segments of a polypeptide chain align next to each other, making a sheet structure coupled by hydrogen bonds between carbonyl and amino groups of backbone, while the R groups extend above and below the plane of the sheet.

The strands of a β-pleated sheet may be parallel (i.e. their N- and C-termini match up), or anti-parallel (i.e. the N-terminus of one strand alongside the C-terminus of the next). In certain cases, the amino acids are not found in α-helices or β-pleated sheets. For instance, proline is known as a "helix breaker" owing to its unusual R group, which bonds to the amino group to form a ring creating a bend in the chain that prevents helix formation. Proline is generally found in bends, unstructured regions between secondary structures. Proteins can contain α-helices, β-pleated sheets or both, or may form neither type.

2.3. Tertiary structure

The tertiary structure, as shown in Fig. 5, is formed as the polypeptide chains of protein molecules fold into a more compact shape with a low surface-to-volume ratio.. The tertiary structure results mainly from electrostatic forces between the R groups. For instance, oppositely charged R groups bond ionically, while similarly charged R groups repel one another. Similarly, polar R groups may form hydrogen bonds and other dipole–dipole interactions.

A cluster of amino acids with nonpolar, hydrophobic R groups on the inside of the protein leaves the hydrophilic amino acids on the outside to interact with nearby water molecules.

The tertiary structure can also be produced by disulfide bonds. Disulfide bonds are covalent and hence keep parts of the polypeptide firmly attached to each other [44]. A synthesis of a tertiary structure is portrayed in Fig. 6.

2.4. Quaternary structure

Many proteins comprise two or more polypeptide chains that interact to form a stable folded structure, known as a subunit of

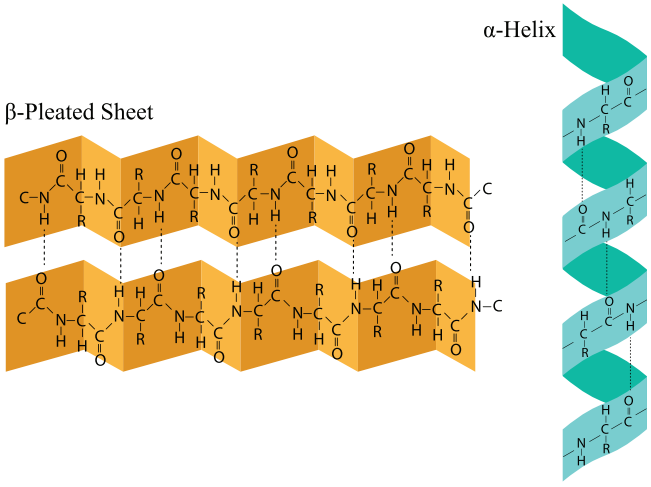

Fig. 4. secondary structure.

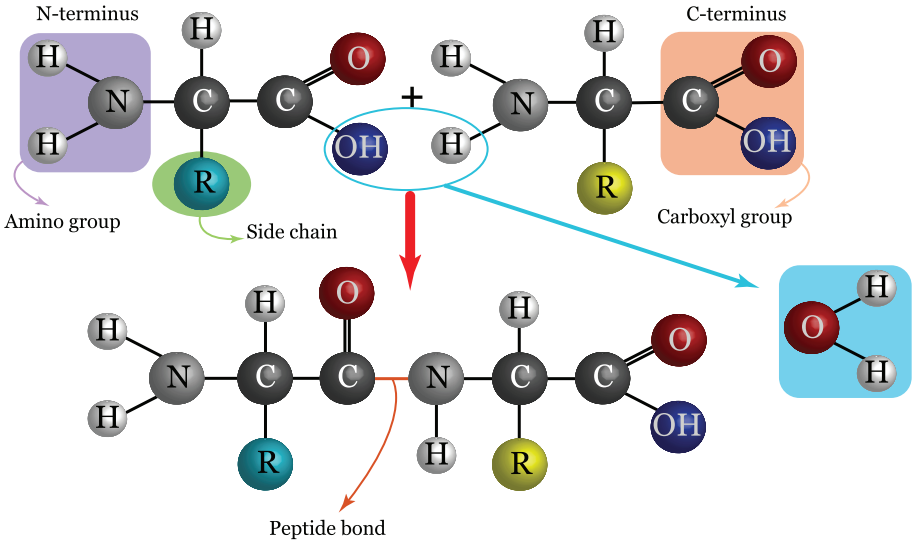

Fig. 3. peptide bond formation. The N-terminus is on the left, and the C-terminus is on the right.

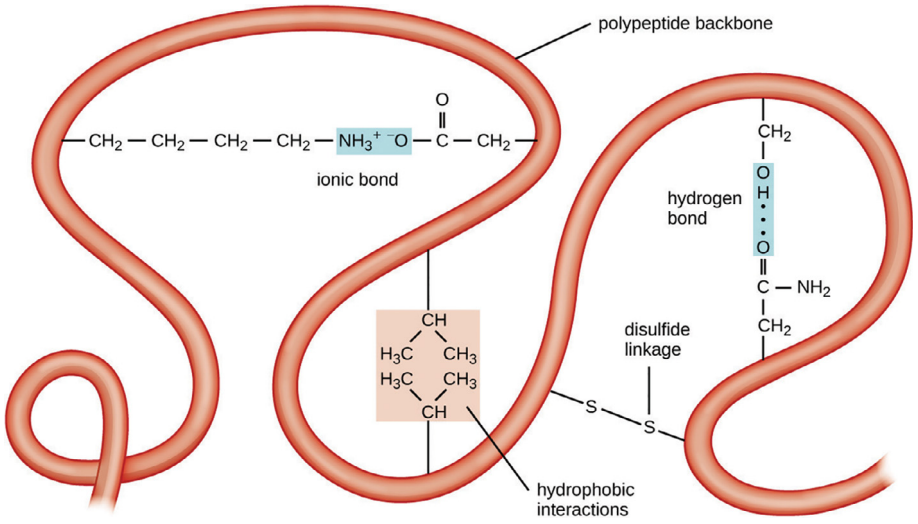

Fig. 5. tertiary structure [70].

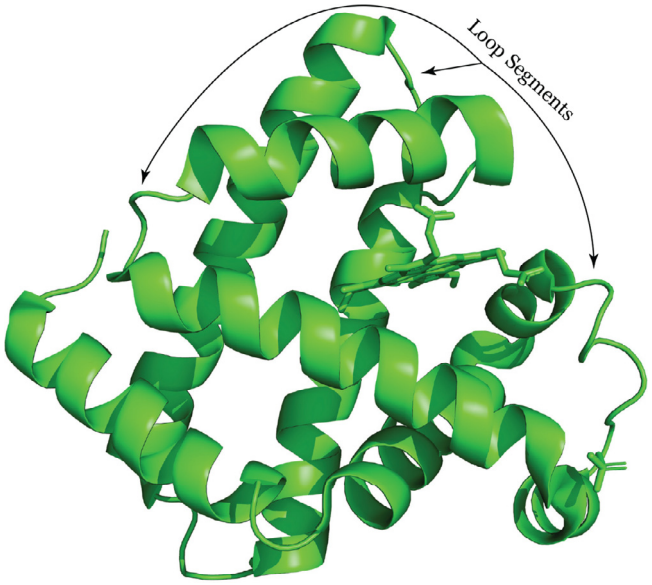

Fig. 6. myoglobin illustrates a type of tertiary structure consisting of αhelices connected by loop segments.

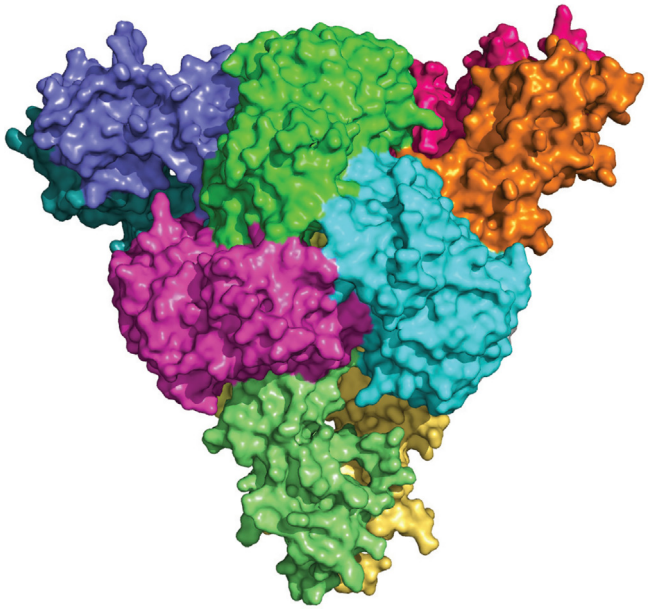

Fig. 7. aspartate transcarbamoylase is an enzyme at the beginning of the pathway for pyrimidine synthesis, presents a remarkable example of quaternary structure.

the protein. The amino acid sequences of each subunit can either be identical (as in tobacco mosaic virus protein), similar (as in the α and β chains of hemoglobin), or entirely different (as in aspartate transcarbamoylase see Fig. 7). Subunit arrangement establishes the protein's quaternary structure.

In the following section, protein shapes are discussed.

3. Protein Shapes

Proteins may be classified on shape and solubility into three global classes: fibrous (Fig. 8), globular (Fig. 9), or membrane (Fig. 10).

In general, fibrous proteins have relatively simple, regular linear structures, and often provide cells with structural functions.

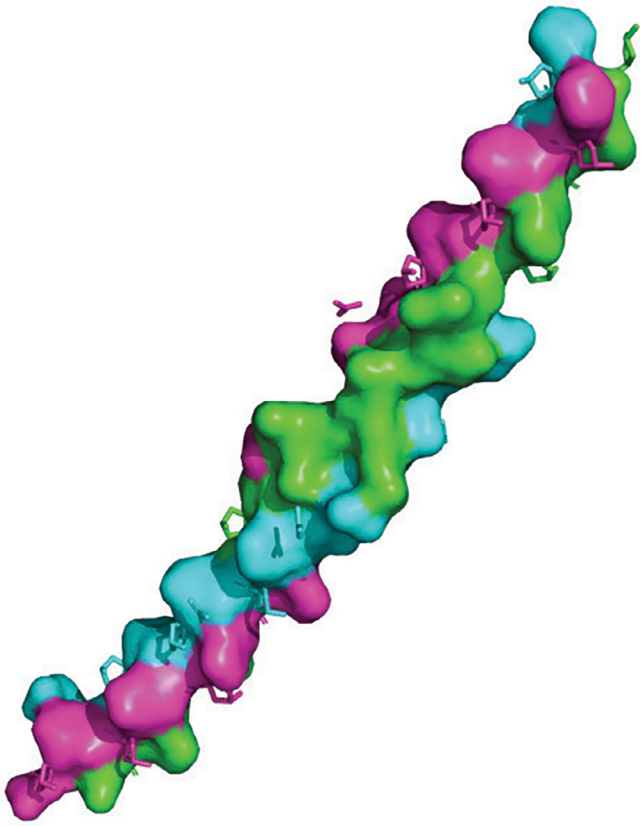

Fig. 8. a small part of collagen separated by chains.

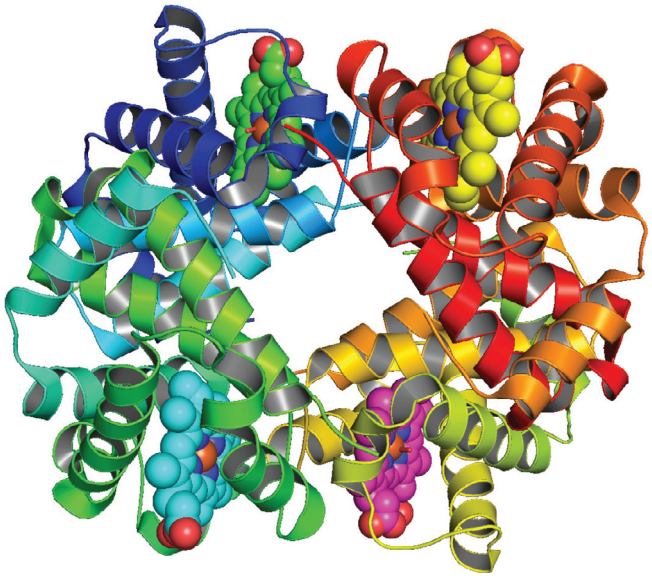

Fig. 9. haemoglobin, a globular protein.

Fibrous proteins are usually insoluble in water and dilute salt solutions. A well-known example of such proteins is collagen, abundant in all animals [71]. As illustrated in Fig. 8, collagen is composed of three chains, each containing 1400 amino acids, twisted together into a triple helix. Glycine appears in every third position along each chain, and, due to its small size, it perfectly fits inside the helix. Proline and hydroxyproline [72] fill numerous positions on a chain. There are numerous types of collagen, all comprising a long stretch of triple helix attached to different ends.

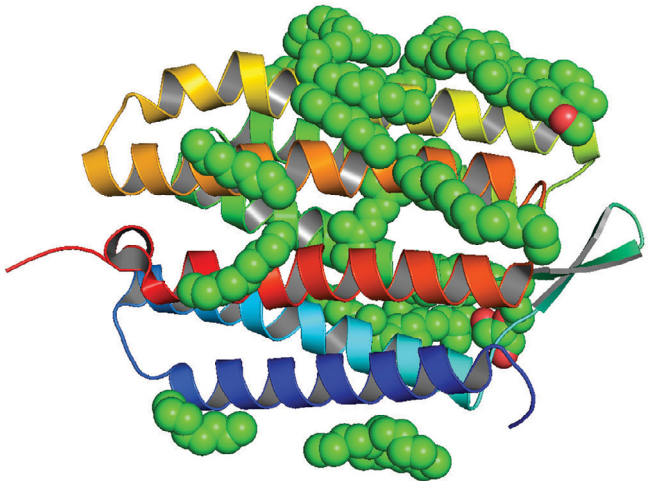

Fig. 10. bacteriorhodopsin, a membrane protein.

On the other hand, globular proteins are nearly spherical, as shown in Fig. 9 and very soluble in aqueous solutions. Examples include haemoglobin, in the red blood cells, that binds to oxygen.

Membrane proteins have hydrophobic side chains directed outwards, and interact with the nonpolar phase within membranes. Therefore, membrane proteins are insoluble in aqueous solutions but can be solubilised in solutions of detergents. Bacteriorhodopsin (Fig. 10) represents an example of such proteins which is made by halophilic (salt-loving) bacteria. This protein pumps protons across cell membranes, powered by sunlight [44].

PPI prediction and protein design may benefit from classifying deformable protein shapes. A novel classification method for protein shapes, based on their macromolecular surfaces, is introduced in [73]. They proposed a novel description, based on bifractional Fokker–Planck and Dirac–Kähler equations for deformable shapes.

3.1. Protein folding

Over the past two decades, considerable efforts have been made in the protein design field, which has further expanded due to the evolution of computational methods and machine learning algorithms. Some of the successful examples include novel folds in protein design [74,75], enzymes [76,77], antibodies [78–80], vaccines [78,81], ligand-binding proteins [82,83], protein assemblies [84–88], and membrane proteins [89–91]. Some of the most recent comprehensive reviews in this field are presented in [92–95]. Generally, the backbone structure of a target protein forms the input for computational protein design. An optimal sequence can be generated using computational sampling methods, seeking potential folding into the desired structure for experimental validation.

A vital component of the solution process involves the scoring function, which can distinguish folds that are or are not physically compatible with a given amino acid sequence [96]. One approach to defining the scoring function considers van der Waals and electrostatic energy along with knowledge-based terms such as backbone dihedral preference statistics about protein structures [97,98], and side-chain rotamers [99]. There is a gap between automated protein design and current approaches, which mostly depend on human experience. This is due to restrictions on artificially created sequences which must comply with various factors such as *in silico* folding free-energy landscape [100,101] and shape complementarity [87].

Despite the rapidly growing number of known protein structures, the number of unique protein folds is converging, suggesting

that statistical learning based on existing structures leads to progress in design methods [102–104]. This statistical potential enables machine learning, especially deep-learning neural networks, to be used for accurate prediction and feature extraction [105].

Some of the commonly used resources for the structure of proteins and sequence are discussed in the following section.

4. PPI Databases

There are several known PPI databases, such as Uniprot[106], SWISS-MODEL [107], Negatome 2.0 [108], STRING [109], RCSB PDB [110], BioGRID [111], DIP [112], BIND [113], MINT [114], HPRD [115] and IntAct [116]. However, among these databases, some are not currently being maintained, such as BIND and HPRD, and are thus rarely used [117]. STRING, IntAct and MINT provide interaction scores from different sources to indicate their reliability. The Negatome 2.0 dataset comprises the manually curated interacting protein pairs from literature and analysed protein complexes from PDB, with scores of zero and one to indicate non-interacting and interacting pairs [108].

Computational methods often use the proteins' biological information, including protein sequences and protein structures. The biological characteristics and high-level structure of proteins are affected significantly by their primary structure. Therefore, one may use the knowledge extracted from protein sequences to estimate the interaction likelihood between protein pairs [118]. Protein sequences can be obtained from the STRING [109], PDB [110], UniProt [106], PIR [119], SWISS-MODEL [107], and TrEMBL [120] databases. Information on higher-level protein structures can be acquired from PDB [121] and SCOP [122]. Dandekar et al. have asserted that proteins encoded by conserved gene pairs physically interact [123]. That basis is used, in genomic-based computational methods, for prediction. Genomic information can be found in The Candida Genome Database (CGD) [124].

In the following section, some of deep learning methods are briefly explained.

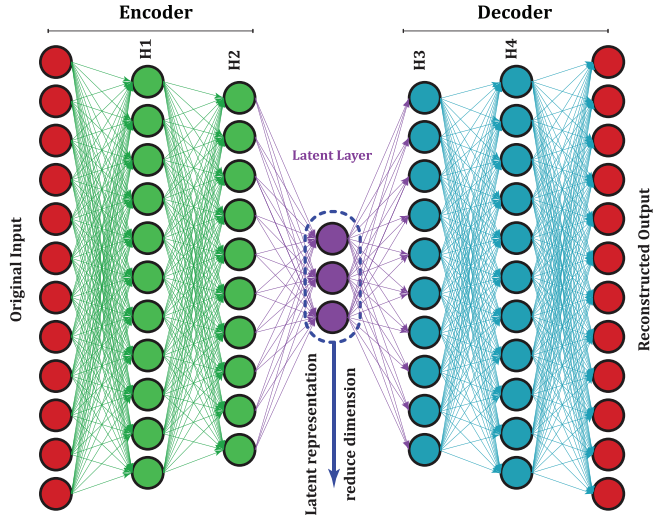

Fig. 11. Autoencoder architecture.

5. Deep Learning Models

Autoencoders, as illustrated in Fig. 11, are a type of unsupervised feedforward neural network reconstructing the output from the high-dimensional and possibly correlated input feature space. [125]. It consists of two parts, the encoder and the decoder. The encoder maps the input data into a low-dimensional and uncorrelated features space, called the latent layer, while the decoder reconstructs the input data from the latent layer. Autoencoders remove redundancies and correlations while extracting highly informative features [126,127].

Recurrent neural networks (RNNs) can capture contextual information when mapping input to output sequences. However, RNNs often suffer from vanishing gradients, limiting the context range they can access [128,129]. To address this problem, long-short term memory (LSTM) architecture was introduced [130], as illustrated in Fig. 12.

The LSTM architecture consists of recurrently connected memory blocks and corresponding control gates, the forget gate f_t, the input gate i_t, and the output gate o_t, which update and control the cell states [131]. The input and forget gates control current network memory and the flow of new information. Specifically, as new information flows into the network, the forget gate manages the information that needs to be removed from cell states, while the input gate controls the information that needs to be stored in cell states. Finally, the output gate determines the encoded information that needs to be forwarded as the input for the next step.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{1a}$$

$$h_t = o_t \otimes \tanh(S_t) \tag{1b}$$

where σ is the sigmoid activation function, W is the weight matrix, b is the bias vector, and $\otimes$ is the point-wise product. The initial operation is performed by the forget gate f_t, Eq. 1a, which determines whether the information should be kept or removed. The LSTM

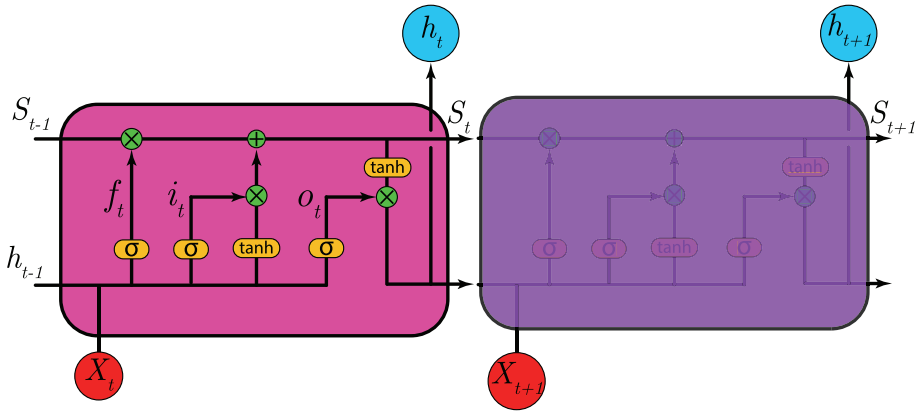

Fig. 12. LSTM architecture.

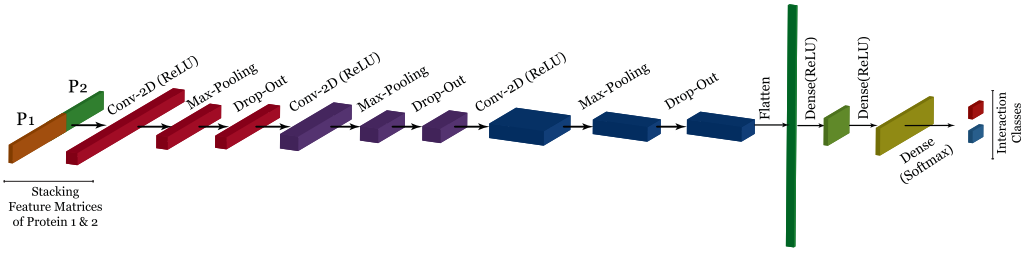

Fig. 13. convolutional neural network architecture. The input is the feature matrices of two proteins. The output predicts the interaction score between two proteins.

architecture contains a hidden state h_t, Eq. 1b that is formed by sequential information.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2a}$$

$$\widehat{S}_t = \sigma(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{2b}$$

The next step involves storing the new input information in the cell state via the input gate i_t, Eq. 2a. Therefore, the cell state can be modified through candidate values $\widehat{S}_t$, Eq. 2b, and Eq. 3a. Finally, the LSTM determines the output of each unit as Eq. 3b.

$$S_t = f_t \otimes S_{t-1} + i_t \otimes \widehat{S}_t \tag{3a}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{3b}$$

Despite the many advantages of LSTM, it is a computationally demanding architecture and slow to train.

The convolutional neural network (CNN) architecture is illustrated in Fig. 13. Its input is a matrix of the encoded representation of two proteins stacked in two columns. In this example, the CNN architecture comprises three 2-D convolutional layers and three dense layers. The convolutional and max-pooling layers reduce the size of the input tensor. The dropout layers are used to reduce overfitting and improve generalisation error [132]. The flatten layer reduces the dimensionality of the input. In addition, three densely connected layers reduce the features to the desired size. Finally, the output is obtained from a densely connected layer with the softmax activation function, classifying interactions into interacting and noninteracting pairs.

Most data used in deep learning can be readily represented in Euclidean space [133], where the convolution operation is properly defined [134]. However, when data cannot be represented on a regular grid due to the complex nature of their correlations [135,136], standard convolution cannot be directly applied to non-Euclidean geometries, limiting the applicability of CNNs [137,138].

However, the convolution theorem states that [137] convolution may be evaluated using Fourier transform. The Fourier transform is first performed for both the input and the filter. Then, both transformations are multiplied by the Hadamard product. Finally, the inverse Fourier transform of the Hadamard product is evaluated. If the Fourier transform is defined correctly, the convolution theorem remains valid under non-Euclidean geometry [137,134], allowing the application of CNNs to non-Euclidean geometries [138]. The spectral graph convolution in the non-Euclidean domain can be obtained by applying the Fourier transform graph and convolution theorem to both the input signal and the convolving filter [139].

Graph convolution extracts underlying local information by collecting node information in the local neighbourhood. Localisation can be achieved by expressing the filters in terms of Chebyshev polynomials of the first kind [140,141]. Fig. 14 illustrates a graph convolutional network (GCN) with stacked layers to extract multi-scale substructure features [138]. The propagation rule for the multi-layer GCN is given by:

$$\mathbf{f}_i^{l+1} = \sigma\left(\sum_{j=1}^{p} \Phi \widehat{\mathbf{G}}_{i,j} \Phi^{\mathrm{T}} \mathbf{f}_i^l\right),\ i = 1, \cdots, q,\ j = 1, \cdots, p \tag{4}$$

Where σ is the nonlinear activation function (l) and $\hat{\mathbf{G}} = diag(g_\theta(\lambda))$. The number of features is denoted by (q), and the number of assets is denoted by (p).

Locality is assumed for all nodes in GCN. As the size of the neighbourhood increases, algorithmic time and space complexity also increase [142]. This issue violates the purpose of using deep

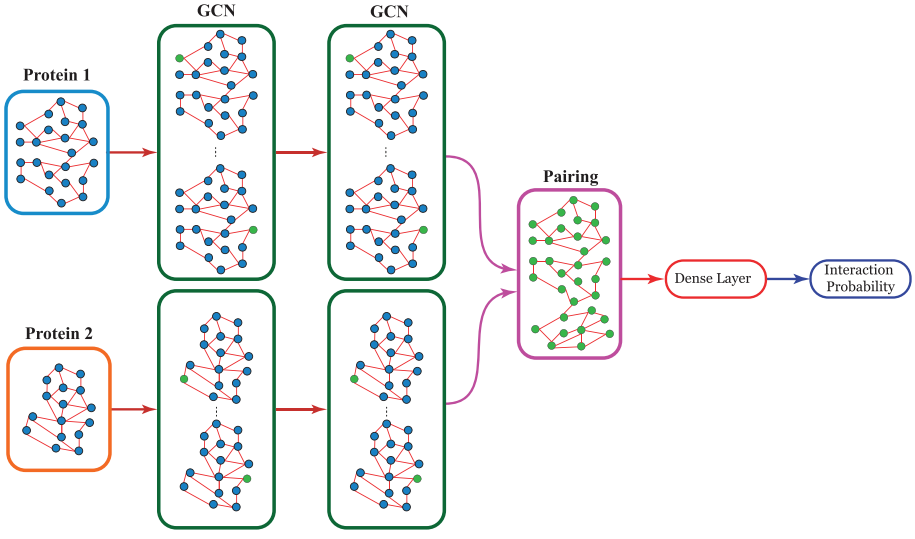

Fig. 14. graph convolutional network architecture for PPI prediction.

models. While few studies have addressed this issue (e.g., skip connection-based models), how to construct a deep architecture that can better adaptively exploit deeper structural graph patterns is still an open challenge [136].

Generative models aim to model the underlying distribution of the data, enabling the generation of new samples with comparable properties to those on which the model was trained [143,144]. Numerous generative models have been developed on the basis of deep neural networks, such as Variational Autoencoder (VAE) [145–147], Generative adversarial Network (GAN) [148], and deep autoregressive models [149–151].

In their original form, GAN algorithms are composed of two components, namely, generator and discriminator, with the generator producing synthetic data while the discriminator evaluates the discrepancy between the generated data and the real data. Each network attempts to improve its performance until an equilibrium is reached, where the discriminator is unable to detect the fake samples and the generator fails to produce better samples [148,152–154].

As illustrated in Fig. 15, given a data distribution, $\mathbf{x} \sim p_x$, $\mathbf{x} \in \mathscr{X}$, the generator learns the distribution p_G which maps the latent variable drawn from a prior distribution $\mathbf{z} \sim p_z$, $\mathbf{z} \in \mathscr{Z}$ to the sample space as $G : \mathscr{Z} \rightarrow \mathscr{X}$, while the discriminator is trained to distinguish between fake and real samples via a score $D : \mathscr{X} \rightarrow \mathbb{R}$ [148].

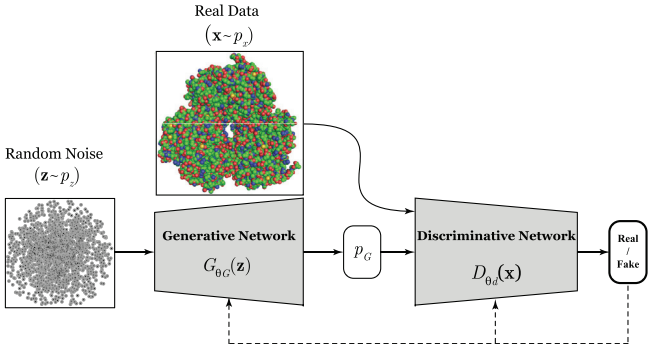

Fig. 15. GAN architecture.

The VAE architecture, as shown in Fig. 16, is a class of generative models based on variational Bayesian inference with multivariate prior distribution [155–157], initially introduced in [145]. The VAEs comprise two linked models that are individually parameterised, namely the encoder or recognition model and the decoder or generative model. Unlike autoencoders, in which the encoder compresses the input features into real-valued latent features, the encoder in a VAE stochastically maps the observed variables' x-space to a probabilistic latent z-space (latent variable) [158].

Fractionally strided convolutions, also known as transposed convolutions, perform a reverse spatial transformation by switching the forward and backward pass [159]. Fractionally strided convolutions may allow for recovering the shape of the initial feature map but do not guarantee retrieving the input itself [159]. This allows the network to learn its own spatial downsampling and upsampling. An extension of the 2-D GAN framework, called conditional GAN, has been proposed in [160] that applies conditions on class labels for both the generator and the discriminator networks. Multimodal data generation is better represented using conditional GANs.

Both generator and discriminator are trained based on an additional information placed as condition $\mathbf{y}$ in the input layer, as depicted in Fig. 17. The adversarial training framework allows for flexible-joint hidden representations composed from input noise $p_z(\mathbf{z})$ and $\mathbf{y}$ in the generator [160]. The fake samples are generated as $G(\mathbf{x}, \mathbf{z}) = \mathbf{x}^*|\mathbf{y}$ ($\mathbf{x}^*$ is synthetic sample given $\mathbf{y}$ as a condition) aiming to resemble real samples as well as possible. The discriminator receives real samples with labels $(\mathbf{x}, \mathbf{y})$ and fake samples from generator $(\mathbf{x}^*|\mathbf{y}, \mathbf{y})$. The discriminator outputs a single probability through a sigmoid activation function (σ) indicating its decision on fake and real inputs.

The following section represents PPI prediction methods.

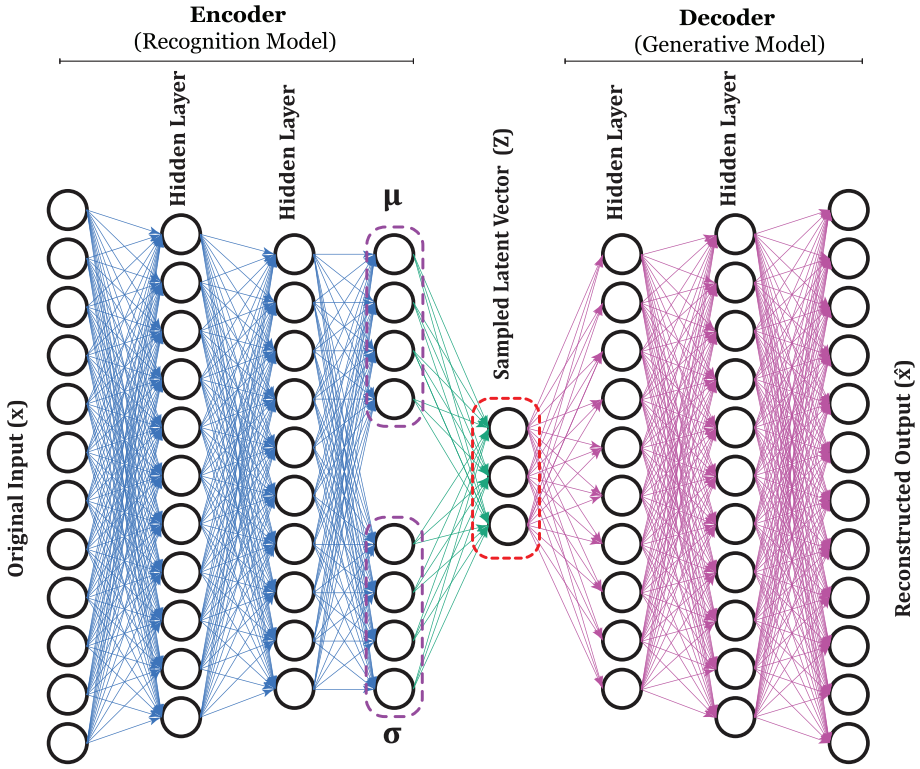

Fig. 16. variational autoencoder architecture.

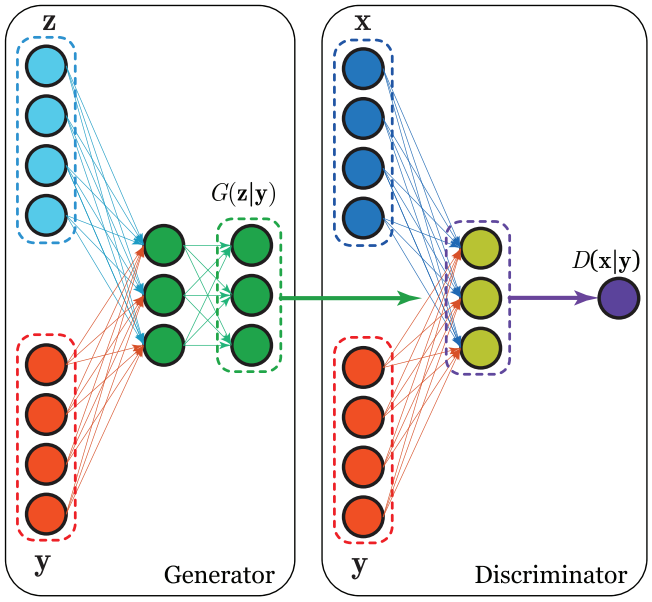

Fig. 17. the conditional GAN architecture.

6. PPI Prediction Methods

High-throughput experimental methods have produced PPIs at an ever greater rate, but these acquired data are noisy with both false positives and false negatives. For instance, mass spectrometry methods may not be able to detect transient or weak interactions [161–165]. The noise levels of different PPI-identifying technologies are studied in [161], showing that high-throughput methods such as two–hybrid system, mass spectrometry, protein chip and phage display have relatively high noise levels.

From a practical perspective, studying PPIs provides the foundation for diagnostic and therapeutic medical applications, thus facilitating the design of novel drugs [117,166–168]. The development of computational methods for the PPI prediction problem is motivated by such shortcomings.

Recent advances in computational modelling methods have brought about exceptional findings in protein design, including enzymes [76,77,169], the development of new therapies [170,171], biosensors [172], and small-molecule binders [82]. However, these methods are mainly suited to modifying naturally found proteins [173]. On the other hand, creating proteins *de novo provides full control over their structure and function [92,174]. Hence, a new objective is to discover new, non-native folds or structural elements as building blocks for novel proteins [173].*

Computational protein design mainly aims to automate the fabrication of proteins with specific structural and functional properties [9,73]. This field has gained traction in the past two decades, such as in the design of novel 3-D folds [74], protein complexes [87], and enzymes [169]. Even though these methods have shown great achievements, current approaches are unreliable as initial designs frequently fail, entailing multiple trial-and-error cycles [175,176]. Since these approaches are highly dependent on the accuracy of complex energy functions for protein physics and the performance of sampling algorithms for jointly exploring the protein sequence, it is difficult to determine the source of the poor reliability [177–180]. Nevertheless, computational methods have facilitated the generation of synthetic protein domains which mimic natural folds using sequences unlike those in nature [181–183].

Quick computational testing of many possible outcomes, potentially narrowing the set of necessary experiments, would ultimately save time. Among the computational methods addressing the PPI prediction problem, some use extracted features as inputs to learn the model [184], while others extract new protein information [185–187]. These methods are further explained in Section 6.5. The information extracted from a tertiary structure of proteins may be used to predict PPI. There exist several experimental techniques for determining a tertiary structure of proteins, including X-ray crystallography and NMR spectroscopy [188]. It is suggested in [189] that locations of protein–protein binding sites are engraved in the proteins' structures. Although experimentally determined 3-D protein structures may facilitate the detection of interaction sites and the understanding of protein functions, experimental biological methods are laborious and time-consuming, and consequently, the geometries of only a small fraction of known proteins have been determined [189–194]. To address this shortcoming, various studies use deep learning to predict, from protein structure and other protein features, potential PPI [185,186,195–197]. Some of these methods are discussed further in Section 6.3.

6.1. PPI Site Prediction

Identifying PPI sites is crucial for understanding the mechanisms of disease and for novel drug design. PPI binding sites consist of amino acid residues forming chemical bonds with a part of another molecule [40]. Identifying interaction domains in sequences helps in understanding cell regulatory mechanisms, locating drug targets and predicting protein functions [198]. Yuan et al. have addressed PPI site prediction as a graph node classification problem, modelling proteins as undirected graphs. They developed GraphPPIS [199] to predict PPI sites.

PPI site predictions are roughly categorised into three categories: protein–protein docking, structure-based, and sequence-based methods. Docking methods aim to generate structures of the resulting protein complex [200], as proposed in [201], by defining a scoring function for novel shape complementarity at the initial docking stage. Some of the recent sequence-based methods for predicting protein–protein interaction sites include: attention-based convolutional neural networks [197], simplified LSTM [191], the DeepPPISP method which uses a combination of local contextual and global sequence features [196], CNN with a residue binding propensity to address data imbalance [202], and the DELPHI method which comprises an ensemble structure as a combination of CNN and recurrent neural network (RNN) [203]. CNN and LSTM architectures are illustrated in Fig. 13 and 12 respectively.

Structure-based methods for PPI prediction are addressed in the next subsection.

6.2. Structure-based PPI Prediction

Proteins adopt complex 3-D structures to perform biological functions via physical contact between effectors and regulators. The effectors may be characterised as the molecules that activate or suppress the regulator's function and alter gene expression as a result [204,205]. Therefore, predicting which residues are involved in PPIs may help structure-based drug discovery, improve the accuracy of protein–protein docking, and obtain richer annotation of protein function [206,207]. A protein may interact with multiple partners over different or overlapping sections of its surface. These interactions may occur at different times or, when the interaction site is large, simultaneously.

Structure-based methods exploit information such as similarity in protein structure to predict PPIs [208]. For instance, two proteins, A' and B', structured similarly to two interacting proteins A and B, respectively, can be assumed also interact with each other [209]. Structure-based techniques often employ empirical scoring functions, physics-based methods, knowledge-based approaches,

or quantitative structure–activity relationship methods to determine both the binding affinities and structural orientations of PPIs [210]. Protein–protein docking techniques can model the orientation of two interacting proteins and their binding affinity and identify key residues in PPIs [210].

Docking-based methods use the structures of individual proteins to predict the structure of the complex. Generally, the only information available is the structure of these individual proteins. The docking method includes two steps. Firstly, the binding orientations of two interacting proteins are identified. Secondly, the binding free energy between the interacting proteins are estimated [211,212]. A global search is conducted by holding the target protein (receptor) stationary while moving the ligand around it. After modelling all possible orientations, the interactions between the two proteins are determined [210]. The global search method demands an unlimited number of translations and rotations, making it a computationally expensive approach. To address this issue, a fast Fourier transform (FFT) approximation has been used in [213].

The local docking technique may improve solutions found by the global docking approach. In the global docking scenario, the sampling starts from a random point, whereas the local techniques assume a known starting point (binding mode) and restrict the sampling search around it [214,215]. The ZDOCK server is among the commonly adopted docking resources which employ FFT-based global search [216]. RosettaDock is a local protein–protein docking algorithm based on a Monte Carlo search. It allows for user-defined initial poses or random orientation of the two proteins. RosettaDock aims to find the system with the lowest energy, initially through a low-resolution optimisation, followed by a high-resolution refinement [217]. Finally, the docking score is estimated by an all-atom energy function [218–220]. The structural features and physicochemical properties are used for showing the models of unknown PPIs. MEGADOCK is a template-free docking methods [221] identify the most promising interactions from a large set of potential interaction sites by assessing the unbound protein components. This method investigates a protein docking approach based on the tertiary structures of the target proteins and physicochemical properties. The docking calculation is accelerated using a novel scoring function called the real Pairwise Shape Complementarity (rPSC) score. Although docking methods have proven successful for some proteins, they fail to deliver the same performance for proteins that sustain conformational changes during interaction [222].

Homologous proteins, i.e. proteins exhibiting similarity through common ancestors sequences [223], are apt to adopt the same binding interfaces [224]. However, the PPI interfaces may be structurally similar, even though their global structures differ [225]. Template-based docking techniques predict PPIs by comparing a protein—protein complex under examination against templates, i.e. other, experimentally determined, protein–protein complex structures [168,211,226–229]. In general, these techniques operate in five steps i) developing the template library, (ii) selecting the target set, (iii) searching for the similarities between target and template, (iv) refinement and (v) scoring. Developing the template library is the most crucial step. In the last decade, the number of experimentally determined structures has grown exponentially, thus improving the performance of template-based techniques [210].

The search for similarity between target and template is performed globally and locally, and can be conducted through sequence alignment [230], structural alignment, and threading [231–234]. Alignments can be obtained from sequences, structures, or feature information from both sequences and structures. The structure framework in the aligned regions of the template with the highest alignment score is selected as the basis of the target protein structure [222,235] (See Fig. 18).

The next subsection will review computational methods predicting PPI using protein 3-D structure.

6.3. Structure-based PPI Prediction Using Computational Methods

In order to address the high-dimensionality problem of protein structure, several dimensional reduction techniques have been applied, such as random forests [237] and the support vector machine (SVM) and its derivatives [238,239]. Northey et al. introduced a multi-layer perception network (MLP)-based method called IntPred [240] to predict interaction by splitting proteins into a group of patches that integrates 3-D structural information into a feature set.

In recent years, many graph convolutional network (GCN) variants (see Fig. 14) [134,241] have been successfully employed in a variety of tasks with graph-structured data [242], such as protein solubility prediction [243], genomic analysis [244] and drug discovery [245]. A GCN-based approach is proposed in [246] to acquire positional information in PPIs. Their representation method combines the information from the amino acid sequence and the protein positions. In order to determine the amino acids of an interacting protein interface, Fout et al. integrated 3-D structures into a GCN [247]. To accurately predict interactions between query proteins entirely from 3-D structural data, Baranwal et al. proposed a GCN-based mutual attention classifier called Struct2Graph [248]. The generative model proposed in [177] is a graph-based model which captures the joint distribution of the full protein sequence, which is founded on long-range interactions resulting from the protein structure. A multimodal approach based on LSTM is proposed in [187], which predicts PPI by integrating structural and sequential information about proteins into the input feature set.

The advantages and disadvantages of these methods are listed in Table 5. Moreover, Table 6 lists the datasets used by each method.

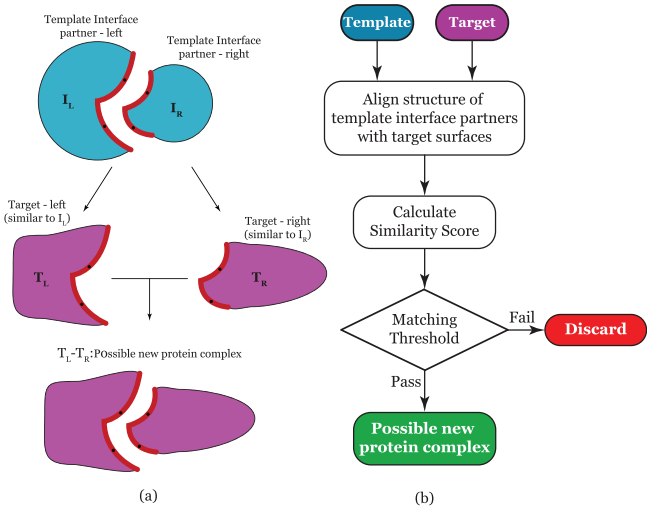

Fig. 18. schematic illustration of PRISM algorithm as an example of template-based docking method for PPI prediction [236] (a) If the template interface on complementary partners (IL and IR) are similar to any two targets surfaces (TL and TR), these two targets may interact and create a protein complex. The black points illustrate hot spot residues. (b) The algorithm flowchart includes the template data set and the target data set. The surface of each partner of the template interface is aligned with the target surfaces. If the matching threshold for hot spot residues passes, the target proteins may form an interacting pair [236,224].

Table 5
Summary of advantages and disadvantages of structure-based deep learning methods for PPI prediction.

Framework	Description	Advantage	Disadvantage
GCN-based [2017] [247]	This study proposed a pairwise classification architecture in which one or more graph convolution layers process the neighbourhood of a residue in each protein. Then, the representation of two residues is paired and passed through a dense layer for classification. This study analysed several GCN-based methods, concluding that neighbourhood-based convolution methods outperform diffusion-based convolution and SVM-based methods.	The proposed convolution operators and obtained features may be helpful for other applications, including protein function, catalytic and other functional residues, and protein interactions with DNA and RNA.	The accuracy of this approach is examined based on a limited number of labelled training examples.
IntPred [2018] [240]	This method uses a random forest to predict protein–protein interface sites at both the surface patch and residue levels.	The performance of a binary classifier can be evaluated using different measurements, such as the Matthews' correlation coefficient (MCC), sensitivity, precision, and specificity [240]. IntPred outperformed the methods ProMate [189], PIER [249], PINUP [250], and meta-PPISP [251], but not SPPIDER [252], based on MCC.	The performance of this method depends on the application. For instance, IntPred was better suited to cases in which false positives are less well tolerated than false negatives.
Graph-based generative model [2019] [177]	This method uses a graph transformer model for designing protein sequences given graph representations of 3-D protein structures, leveraging the spatial locality of dependencies in molecular structures.	This method uses a self-attention mechanism to capture higher-order, interaction-based dependencies between sequence and structure. The graph-based model offers computational efficiency due to the representation of long-range sequence dependencies by short-range sequence dependencies in 3-D space [253–255]. Additionally, they achieved linear computational scaling concerning the sequence length and representational flexibility for coarse and fine-grained structure descriptions.	The evaluation dataset only contained chains up to a length of 500, limiting the applicability of this approach.
Struct2Graph [2020] [248]	In this method, graph embeddings of each protein are obtained using an assigned GCN. Next, relevant geometric features associated with query protein pairs are extracted using a mutual attention network. Finally, a feedforward neural network performs a binary classification between interacting and noninteracting pairs.	Struct2Graph only uses 3-D structural information to predict the PPI. They have reported state-of-the-art performance on both balanced and unbalanced datasets.	Limited availability of 3-D structural information may restrict the applicability of this method.
LSTM-based [2020] [187]	The proposed method integrates the 3-D structure and sequence-based information of proteins to predict PPIs. The 3-D coordinate information, hydropathy index, isoelectric point, and amino acid charges of each protein are fed into a pre-trained ResNet50 model to extract features from these attributes. A stacked autoencoder obtains the compact form of encoded proteins using autocovariance and conjoint triad. The structural features from ResNet50 are passed through LSTM and concatenated with features from the stacked autoencoder. The merged features are then fed into the classifier to predict protein pair labels.	This method performs well despite being trained on a low number of instances.	Limited availability of 3-D structural information may restrict the applicability of this method. Additionally, LSTM models are computationally demanding and slow.

The sequence-based methods for PPI prediction are reviewed in the next section.

6.4. Sequence-based PPI Prediction

Traditional methods often analyse protein sequences based on multiple sequence alignments. This leads to a simple inference of functional and structural constraints from sequence data [265]. While protein design and engineering have benefited from evolutionary information of alignments [266–268], adding distant proteins will induce large and unreliable alignments [269], restricting the diversity of sequences.

Unlike docking and structure-based methods, sequence-based methods do not require structural data and instead leverage the abundance of existing protein sequence data from sequencing technology, especially since the introduction of metagenomics [106,270,271]. One may predict PPIs from amino acid sequence similarity in the known interactions, depending on interactions already identified in one species to infer interaction in different species [24,272]. The sequence-based method hence focuses on primary structure, disregarding the protein's 3-D shape [273].

In domain-based technique, specific sequences in the protein structure are represented by conserved domains. Conserved Domains may be defined by local multiple sequence alignments, including a wide range of organisms to display sequence regions

Table 6
Datasets of structure-based methods.

Framework	Data Processing
GCN-based [2017] [247]	Version 5 of the docking benchmark dataset was used by this study [256], comprising a selected subset of structures generated from X-ray crystallography or nuclear magnetic resonance experiments and containing the atomic coordinates of each amino acid residue in the protein from the Protein Data Bank (PDB). Proteins with 29 to 1,979 residues are included. Since proteins may change their shape upon binding, the features are computed from the unbound form of the protein in the complex. The labels are acquired from the structure of the proteins in the complex.
IntPred [2018] [240]	The training dataset comprised 58,397 biological units from protein, interfaces, structures and assemblies (PISA), including transient and obligate interfaces [257]. Structures with a resolution below 3 A° or Rfactor above 30%, Viral capsids, NMR entries and proteins with fewer than 30 amino acids are removed. Any structure with more than one chain is retained, resulting in 25,876 structures constructed from 87,738 chains. The chains were clustered at 25% sequence similarity using PISCES [258] to remove redundancy. The final training set contained 4,345 chains. For the test set, no clustering was performed, resulting in 4,204 chains. The NOXclass [259] is used to construct a dataset of obligate and transient interfaces. This method predicts protein interactions as either obligate or non-obligate (transient) with or without crystal packing contacts.
Graph-based generative model [2019] [177]	The dataset was obtained from the CATH (version.4.2) [260]. The training, validation, and testing sets were divided into 80/10/10 sets by randomly assigning their CATH topology classifications (CAT code). The resulting dataset included 18,024 chains in the training set, 608 in the validation set, and 1,120 in the test set, with zero CAT overlap.
Struct2Graph [2020] [248]	The database was generated based only on direct/physical protein interactions. Therefore, IntAct [116] and STRING [261] were selected, and only concordant matches between these two databases were chosen as true interactions. The organisms included in this dataset were S.cerevisiae, H.sapiens, E.coli, C.elegans, and S. aureus, resulting in 427,503 pairs from IntAct and 852,327 pairs from STRING. Only "direct association/interactions" from IntAct and "binding" from STRING were regarded as physical interactions. Only extracting concordant, physical interaction data reduced the interactions to 12,676 pairs for IntAct and 446,548 pairs for STRING. Negative PPI was retrieved from [262]. Structure information for this method was acquired from PDB files, which reduced the total number of pairs to 117,933 (5,580 positive and 112,353 negative). All proteins were matched with PDB files using UniProt accession numbers (UniProt Acc) and mapped PDB files [263]. Finally, PDB files were curated based on the length of their chain ID and the highest resolution within each PDB file, resulting in 5,580 negative pairs for a balanced dataset.
LSTM-based [2020] [187]	This study used two PPI datasets, Pan's PPI dataset [264] and S. cerevisiae PPI data obtained from the Database of Interacting Proteins (DIP; version 20160731; see Stacked Autoencoder (SAE) and DeepPPI for further details). Structure information was only available for 10,359 protein sequences in the Pan's PPI dataset and 1,308 proteins in the S. cerevisiae PPI dataset. Therefore, Pan's PPI dataset contained 25,493 pairs (18,025 positive and 7468 negative), while the S. cerevisiae PPI dataset contained 4,314 positive and 6,265 negative pairs.

containing the same, or comparable, patterns of amino acids [274]. This idea can be used to predict the subcellular location of the protein and the class and subclass of the enzyme, to find functional interactions, and to identify the membrane protein [275]. Computational approaches have been developed to predict PPIs based on the information this technique provides [275]. A domain-based method to estimate the interaction map of E.coli, for example, is developed in [276]. Another domain-based method is introduced by Kim et al. [277] which estimates the probability of the interaction between interacted domains. Using the relevant vector machine (RVR) and domain features with support vector regression (SVR), Kamada et al. identified PPIs [278]. DomainGA is a multiparameter optimisation approach which is developed to predict the score of PPI [279].

The ortholog-based techniques also use similarities between amino acid sequences [210]. The annotations to a functionally determined protein sequence are transferred to similar parts of a target sequence. This work relies on databases of annotated proteins to construct the homologous model of the studied protein [280]. Significant sequence similarities may be shared among multiple proteins from an organism in systems in diverse organisms. Thus, if a significant similarity is found between an input protein and an annotated protein (with known functions), the input protein may be hypothesised to possess similar properties or functions. In order to identify these functions, paralog and ortholog approaches are employed. Orthologs are homologous genes that evolved by vertical descent from a single ancestral gene; In contrast, paralogs evolved by duplication [223,281]. For instance, the orthologs of two interacting proteins, A and B, can interact similarly in different species [210].

Computational methods predicting PPI using protein sequence are addressed in the next subsection.

6.5. Sequence-based PPI Prediction Using Computational Methods

The Interface Weighted RAPtor (iWRAP) integrates a boosting classifier with a novel linear programming formulation for interface alignment to predict interacting proteins encoded by the entire yeast genome. The interface profiles are constructed using SCOPPI [282], based on the sequence and structural similarity of the interface [283]. The Universal In Silico Predictor of Protein–protein Interactions (UNISPPI) uses the primary structure to classify protein pairs as interacting or non-interacting [284]. A matrix-based representation of protein sequence coupled with the SVM algorithm is proposed in [285], using the order of primary structure and its dipeptide information. The sequence-based methods may be split into two distinctive techniques: domain-based and ortholog-based [273]. PPI prediction is conducted based on sequences in [286] by defining units of three adjacent amino acids and measuring the frequency of those units in a protein sequence. Other methods such as amino acid index distribution [287], conjoint triad method (CT) [286] and autocovariance (AC) [288] are developed to extract features such as locations of amino acids, frequencies and physicochemical properties, with the aim of representing a protein sequence.

An SVM based approach, ACT-SVM, is developed in [289] to extract features from protein sequences as the input vector for the classifier. A sequence-based human PPI prediction is developed in [18] based on a Stacked autoencoder (SAE). Another sequence-based PPI prediction approach is introduced in [165] called D-

Table 7
Summary of advantages and disadvantages of sequence-based Deep Learning methods regarding PPI prediction.

Framework	Description	Advantage	Disadvantage
SVM-conjoint triad [2007] [286]	Each protein sequence was represented in this study by a vector of amino acid features. The model was developed based on a support vector machine (SVM) integrated with a kernel function and a conjoint triad feature for describing amino acids. This method mapped different types of PPI networks using only sequence information, which could be applied to explore networks for any newly discovered protein with unknown biological relationships. They suggested that methods without local environments for amino acids are often unreliable, so a conjoint triad method was used.	The 20 standard amino acids were clustered into several classes based on their dipoles and side chain volumes to achieve dimensionality reduction of the vector space. This method might predict PPI networks created by pairwise PPIs.	The limited available information on protein pairs restricts the applicability of this method. Additionally, it mainly considers the properties of two nearby amino acids, overlooking long-range interactions.
SVM-autocovariance [2008] [288]	This method combined a new feature representation using autocovariance (AC) and a support vector machine (SVM). AC considers the interactions between more distant amino acids in the protein sequence, specifically long-range interactions. This is an improvement over the method proposed in [286]. This model was evaluated using an independent dataset of 11,474 yeast PPIs.	The conjoint triad (CT) method only considered the attributes of an amino acid and its two neighbouring amino acids [286], while long-range interactions are accounted for by the AC method. In this study, AC variables represented information on interactions between one amino acid and its 30 neighbouring amino acids in the protein sequence.	The model achieved a low prediction accuracy of 58.42% in a negative dataset created using the Prcp method [286].
UNISPPI [2013] [284]	This method used a decision tree model, predicting PPIs using only 20 amino acid frequency combinations from interacting and noninteracting proteins as learning features. This study indicated that asparagine, cysteine, and isoleucine frequencies are important features for discerning between interacting and noninteracting protein pairs.	This method was scalable due to using a limited number of attributes. Moreover, this method was based on experimentally validated instances from various species, covering many species.	Instances with a classification score of 0.50 were classified as neither PPIs nor non-PPIs, limiting the applicability of this method. Additionally, the obtained accuracy of 79.4% for interacting and 72.6% for noninteracting pairs are relatively low.
SVM-based method [2015] [285]	PPI prediction was addressed by integrating a support vector machine (SVM) and a novel matrix-based representation of the sequence order and dipeptide information of the primary protein sequence, extracting more information than amino acid dipeptide composition. The SVM classified the interaction between protein pairs using these feature vectors.	This method extracts more information hidden in protein primary sequences than amino acid dipeptide composition.	SVM algorithms performed relatively poorly with noisy data and are unsuitable for large datasets since training time may increase significantly [303]. Moreover, finding a proper kernel function was difficult.
DeepPPI [2017] [292]	The DeepPPI method used a deep neural network architecture network for each protein to extract high-level discriminative features from common protein descriptors. The interaction between two proteins was determined using the one-hot encoding label. This method comprises two different architectures: DeepPPI-Sep, which uses two separate networks as input for each protein, and DeepPPI-Con, which directly links two proteins in a single network.	This method can capture informative features of protein pairs by a layer-wise abstraction. In addition, DeepPPI can automatically learn an internal distributed feature representation from the data.	The accuracy of DeepPPI for All Human/ Yeast dataset are relatively low, and the accuracy of methods proposed in [304] exceeds that of DeepPPI.
Stacked Autoencoder (SAE) [2017] [18]	This method used a stacked autoencoder to predict PPI. The feature extraction from protein sequences was performed using autocovariance (AC) and the conjoint triad (CT).	SAE can learn hidden interaction features of protein sequences.	They used a synthetic negative interaction dataset, and the accuracy of this model for negative interactions is relatively low.
DPPI [2018] [298]	This method performed sequence-based PPI prediction using a deep, Siamese-like convolutional neural network combined with random projection and data augmentation. This method captured the composition information, sequential order of amino acids, and co-occurrence of interacting sequence motifs in a protein pair. Each protein was characterised as a probabilistic sequence profile generated by PASI-BLAST. The patterns in each sequence were identified using the convolutional module, comprising multiple layers. The representations learned by the convolutional module were projected to two different spaces using the random projection module, allowing DPPI to explore the combination of protein motifs.	DPPI addresses interactions for both homodimeric and heterodimeric proteins. Moreover, this method could model binding affinities.	This method yields lower PPI prediction accuracy on the S.cerevisiae core dataset from PIPR based on 5-fold cross-validation compared to PIPR [299] and DeepTrio [302].

Table 7 (*continued*)

Framework	Description	Advantage	Disadvantage
PIPR [2019] [299]	This study proposed an end-to-end framework for PPI prediction based on amino acid sequences using a deep residual recurrent convolutional neural network in the Siamese architecture. This method leveraged an automatic multi-granular feature selection to capture local significant and sequential features from protein sequences.	The Siamese-based learning architecture captured the mutual influence of protein pairs and allowed for generalising to address different PPI prediction tasks without needing predefined features.	RCNN was built using bidirectional gated recurrent units (bidirectional-GRU). However, GRUs suffer from slow convergence and low learning efficiency [305].
S-VGAE [2020] [300]	This model proposed a signed variational graph autoencoder (S-VGAE) that combined sequence information and graph structure. In this method, the PPI network was regarded as an undirected graph. This framework comprised three parts. First, coding the raw protein sequences. Second, the S-VGAE model extracted vector embedding for each protein with sequence information and graph structure. Finally, a simple three-layer softmax classifier. This model was inspired by the variational graph autoencoder (VGAE) [306] that uses latent variables to learn interpretable representations for undirected graphs.	In this method, the cost function was modified only to consider highly confident interactions, making it more robust to noise.	This model used the conjoint triad (CT) method [286] to encode amino acids. However, CT does not account for long-range interactions in the protein sequence.
ACT-SVM [2020] [289]	This method performed feature extraction on the protein sequence to obtain a vector, composition, and transition descriptor and integrated them into a vector. Then, the feature vector was fed into the SVM classifier. The performance of this method was evaluated using 5-fold, 8-fold, and 10-fold cross-validation on H. pylori and human datasets.	They have observed that SVM method outperforms K-Nearest Neighbour (KNN), ANN, RFM, Naive Bayes, Logistic Regression, s for the H. pylori protein pairs.	Finding the proper kernel and hyperparameters was challenging, and training time for SVM classifiers increases with dataset size [307].
D-SCRIPT [2021] [165]	Deep sequence contact residue interaction prediction transfer (D-script) is an interpretable deep learning method generating structurally informative features given protein sequences using a pre-trained language model from [290]. This method used projection modules to reduce the dimension of features, including the residue-contact map of the protein. Finally, the interaction probability was predicted based on the contact maps.	D-SCRIPT generalised to new species considering the sparsity of training data for most model organisms (i.e., it was relatively accurate for cross-species PPI prediction).	Despite its performance for cross-species PPI prediction, D-SCRIPT underperformed on within-species evaluations. The training dataset only included proteins with 50–800 amino acids, limiting the applicability of this method.
SPNet [2021] [9]	The Siamese pyramid network (SPNet) architecture used self-binding and folding amino acid sequences to predict the binding probability for two proteins based solely on their amino acid sequences. Subsequent screening through potential candidates was performed based on binding probabilities.	This architecture consisted of a multilevel pyramid feature structure encompassing various PPI mechanisms to reduce gradient explosion and disappearance, a multilevel Siamese neural network with an attention mechanism, and a multilevel, trainable binding probability prediction network.	
BiLSTM-RF [2021] [291]	The BiLSTM-RF model extracted features of protein pairs in the human database. BiLSTM comprises forward and backwards LSTMs and is capable of bidirectional encoding (i.e., encoding front-to-back and back-to-front information). A random forest classifier (RF) was built with 100 trees and used a voting strategy to integrate these results to predict the interaction.	BiLSTM extracted the sequence and position of the biological information in the protein sequence.	LSTM models are computationally demanding and slow. Moreover, a large number of trees in the random forest leads to a longer training time.
Heterogeneous Network [2021] [296]	PPI prediction was performed using a computational sequence and network representation learning-based model. Local features were extracted from the protein sequence using the *k*-mer method (k = 3), while global features were extracted from the heterogeneous network. The latter captured network structure and obtained potential linked information. This method integrated local features with global features to represent protein nodes.	The protein node contained protein attribute and network structure information by integrating local and global features.	Model accuracy is relatively low compared to other deep learning methods such as DPPI [298].

(*continued on next page*)

Table 7 (*continued*)

Framework	Description	Advantage	Disadvantage
OR-RCNN [2021] [301]	This method was called ordinal regression and recurrent convolutional neural network (OR-RCNN), which predicted PPIs based on their confidence score. The architecture comprised two recurrent convolutional neural networks (RCNNs) encoders, which shared the same parameters, to extract robust local features and sequential information from protein pairs. Then, one novel embedding vector was obtained by element-wise multiplication of the two embedding vectors from RCNNs. The second part of the architecture performed an ordinal regression model via multiple sub-classifiers that use the ordinal information behind the confidence score. Finally, the confidence score determined the existence of PPI with a threshold.	This method offered better accuracy compared to some existing models, such as autocovariance [288] and composition transition distribution (CTD) descriptor [308] for feature description, and random forest (RF) [309], extreme gradient boosting (XGBoost) [310], and support vector machine (SVM) [311] for the prediction.	The RCNN was built using bidirectional gated recurrent units (bidirectional-GRU). However, GRUs suffer from slow convergence and low learning efficiency [305].
DeepTrio [2022] [302]	The DeepTrio method used a deep-learning framework based on a mask multiscale CNN architecture that performed binary PPI prediction by capturing multiscale contextual information of protein sequences using multiple parallel filters. This method used a single-protein class, allowing it to distinguish relative and intrinsic properties. This method was also made available as an online tool to address cross-platform usage and dependency-related issues.	DeepTrio is available both online and offline.	DeepTrio yields lower PPI prediction accuracy on the S.cerevisiae core dataset from PIPR based on 5-fold cross-validation compared to PIPR [299]. DeepTrio achieves lower PPI prediction accuracy on the S.cerevisiae core dataset from DeepFE-PPI based on 5-fold cross-validation compared to DeepFE-PPI [312].

SCRIPT (Deep Sequence Contact Residue Interaction Prediction Transfer), which models protein structure using a pre-trained language model from [290]. A novel deep learning approach called Siamese Pyramid Network (SPNet) architecture is proposed in [9], which predicts the binding probability of two proteins based on their amino acid sequences. This method is employed to discover the proteins that potentially bind with the 2019-nCov spike, in order to find future vaccines.

Learning protein pair representation is tackled by deep learning methods such as BiLSTM-RF, which uses LSTM to extract the features of protein pair sequences and a random forest classifier [291], and DeepPPI [292] that uses a separate network for each protein and learns high-level features from raw protein features. The EnsDNN approach extracts the interaction information of proteins from amino acid sequences, using AC descriptor [293], local descriptor (LD) [294,295] and multi-scale continuous and discontinuous local descriptor (MCD) [15,237]. A heterogeneous network for PPI prediction is presented in [296] which uses the concatenation of local and global features to present protein node. The local features are extracted from protein sequence by the *k*-mer method (k = 3)[1], while the global features are extracted from the heterogeneous network, and heterogeneous networks by LINE (Large-scale Information Network Embedding), respectively, and random forests to classify and predict potential protein pairs.

DPPI [298] performs sequence-based PPI prediction using a deep, Siamese-like convolutional neural network combined with random projection and data augmentation. This method captures the composition information, sequential order of amino acids, and co-occurrence of interacting sequence motifs in a protein pair. The PIPR method [299] predicts PPI by integrating a deep residual recurrent convolutional neural network (residual-RCNN) in the Siamese architecture, leveraging both robust local features and contextualised information on the protein sequence. The signed variational graph auto-encoder (S-VGAE) graph-based method [300] considers the PPI network as an undirected graph, combining sequence information and graph structure to predict PPI. A deep learning method called OR-RCNN is developed to predict PPI [301], which is composed of two recurrent convolutional neural networks (RCNNs) to extract local features and sequential information from the protein pairs and ordinal regression to construct multiple sub-classifiers. The sequence-based PPI prediction approach DeepTrio [302] uses mask multiple parallel convolutional neural networks.

Table 7 represents further analysis of the sequence-based methods. Additionally, the datasets of each of these methods are reported in Table 8.

Some of the methods addressing protein design problems, including protein function, sequence, and structure, are discussed in the following section.

7. Protein Design

Most protein design problems require profound knowledge and subjective expertise to analyse obstacles and obtain optimal design strategy. However, with the emergence of deep neural networks, computational capacity and the available historical data, new computational methods have shown advantageous in many cases, such as RNNs' successful application in generating SMILES (Simplified Molecular Input Line Entry System) sequences for *de novo drug discovery [328,329] and optimising the RNN output to obtain specific properties through transfer learning and fine-tuning on desired sequences [330]. Recently, numerous studies have been conducted on predicting protein properties and generating new molecules and DNA sequences. These include, for example, graph neural networks for molecule representation [331–334], prediction of amino acid sequence for a particular structure using deep neural networks [104], using GAN to generate DNA sequence [335], and structure prediction methods using neural networks [180,336,337].*

7.1. Protein Function

In several instances, functional folded proteins have been acquired from random-sequence libraries; however, this process

[1] *k*-mers are the substrings with length *k* within a biological sequence [297]

Table 8
Datasets of sequence-based methods.

Framework	Data Processing
SVM-conjoint triad [2007] [286]	A dataset comprising 16,443 nonredundant entries of experimentally verified PPI was extracted from the Human Protein Reference Database (HPRD; version 2005–0913; www.hprd.org). These interactions are primarily based on individual in vivo (e.g., coimmunoprecipitation) or in vitro (e.g., GST pull-down) experiments [286]. The negative dataset was created by excluding pairs that appeared in the positive dataset. For example, if AB and IJ are interacting pairs, AI, AJ, BI, and BJ may be noninteracting pairs. Additional conditions were applied, including an equal number of negative and positive pairs (16,443 in this study) and harmonious contributions of proteins forming the negative set. Therefore, the training set is equally distributed, comprising 32,486 protein pairs. The test set contained another 400 protein pairs. Both positive and negative pairs were randomly selected.
SVM-autocovariance [2008] [288]	The PPI data was extracted from the S. cerevisiae core subset of the Database of Interacting Proteins (DIP; version.20070219) [112], containing 5,966 interaction pairs. The expression profile reliability (EPR) and paralogous verification method (PVM) were used to test the reliability of this core subset [313]. By removing proteins with fewer than 50 amino acids, 5,943 protein pairs formed the final positive data set. The CD-HIT program was used to obtain a nonredundant subset with a sequence identity of 40%. The negative dataset was created using the Psub method, assuming that proteins located in different subcellular localisations do not interact. The subcellular location information was extracted from Swiss-Prot (http://www.expasy.org/sprot/). This method excluded proteins without subcellular localisation information and those marked as 'putative' or 'hypothetical', while proteins localised to the cytoplasm, nucleus, endoplasmic reticulum, Golgi apparatus, lysosome, and mitochondrion remained. The noninteracting pairs were generated by pairing proteins from one subset with those from the other. This strategy must satisfy the following conditions: (1) The DIP yeast interacting pairs do not include any noninteracting pairs, (2) there is an equal number of negative and positive pairs, and (3) the negative set should have a harmonious contribution.
SVM-based [2015] [285]	This method was evaluated using S. cerevisiae and H. pylori PPI datasets. The former was obtained from the S. cerevisiae core subset of the Database of Interacting Proteins (DIP). The non-redundant and negative pairs were obtained according to Guo et al. [288]. Therefore, the PPI dataset included 11,188 interacting and noninteracting pairs. The H. pylori dataset contained 2,916 protein pairs (1,458 interacting and 1,458 noninteracting) following [314].
DeepPPI [2017] [292]	The dataset evaluating the DeepPPI comprised 11,188 negative and positive protein pairs from S. cerevisiae obtained from the Database of Interacting Proteins (DIP; version 20160731), 1,458 interacting and 1,458 noninteracting pairs from H.pylori, 3,899 interacting and 4,262 noninteracting pairs from humans, 4,013 interacting pairs from C.elegans, 6,954 interacting pairs from E. coli, 1,412 interacting pairs from H. sapiens, 313 interacting pairs from M. musculus, and one additional H. pylori data set of 1,420 interacting pairs used in [315]. The negative dataset was created by pairing proteins from one subcellular location information extracted from Swiss-Prot (http://www.expasy. org/sprot/) with proteins from other locations.
Stacked Autoencoder (SAE) [2017] [18]	Pan's PPI dataset was acquired from [264], comprising 36,630 positive PPIs from the human protein reference database (HPRD, version 2007). Negative PPIs were generated by pairing proteins discovered in different subcellular locations from the Swiss-Prot database (version 57.3). After removing proteins with fewer than 50 amino acid residues, 2,184 unique proteins from six subcellular locations (cytoplasm, nucleus, endoplasmic reticulum, Golgi apparatus, lysosome, and mitochondrion) remained. The addition of negative pairs from the Negatome dataset [108] provided 36,480 total negative pairs. Protein pairs with nonstandard amino acids such as U and X were removed, resulting in a benchmark dataset of 36,545 positive and 36,323 negative pairs. The pre-training set contained 33,052 positive and 32,816 negative pairs, while 7,000 randomly selected pairs (3,493 positive and 3,507 negative) formed the test set. Pre-training and testing used 10-fold cross-validation. The external test sets used in this study included the 2010 version of the HPRD dataset, the 2010 HPRD NR dataset, the DIP dataset, and the HIPPIE dataset.
DPPI [2018] [298]	This study used human and yeast datasets from [316]. The human PPI dataset was created by taking the 10% top-scoring interactions from the Hippie database v1.2 [317]. The yeast PPI dataset was retrieved from DIP database [318]. Negative pairs were generated by randomly sampling from all proteins, where a 10:1 negative-to-positive ratio was considered [316]. Additionally, data redundancy regarding sequence similarity in PPI (>40%) was removed following the strategy of [316]. Finally, a 10-fold cross-validation was performed.
PIPR [2019] [299]	Guo's dataset [288] comprised 2,497 proteins forming 11,188 PPI pairs, half representing positive pairs and half representing negative pairs. Interaction pairs for H. sapiens were obtained from the STRING database (version 10.5) [319]. Three thousand randomly selected proteins and 8,000 proteins that shared $< 40\%$ sequence identity formed two subsets. Finally, protein binding affinity data were obtained from the SKEMPI dataset [320], comprising 3,047 binding affinity changes after mutation of protein subunits within a protein complex for use in the affinity estimation task.
ACT-SVM [2020] [289]	Following [321] a nonredundant dataset including H. pylori and human PPI was created. The H. pylori dataset comprised 1,458 interacting and 1,457 noninteracting protein pairs, while the human dataset comprised 3,899 interacting and 4,262 noninteracting protein pairs.
S-VGAE [2020] [300]	This study used data from the human protein reference database (HPRD) and the Database of Interacting Protein (DIP) for humans, Drosophila, E.coli, and C.elegans.

(*continued on next page*)

Table 8 (*continued*)

Framework	Data Processing
D-SCRIPT [2021] [165]	This study used a dataset from the STRING database (version 11) [261]. The positive pairs were limited to interactions associated with a positive experimental-evidence score. Only proteins containing 50–800 amino acids were retained. Additionally, proteins meeting the 40% similarity threshold were clustered using CD-HIT [322,323], removing redundant PPI from the dataset and preventing the model from memorising interactions based only on sequence similarity. Negative pairs were generated by randomly pairing proteins from the nonredundant set with a 10:1 negative-to-positive ratio [316]. The human PPI dataset comprised 47,932 positive and 479,320 negative protein interactions. Training and validation sets comprised 80% (38,345) and 20% (9,587) of pairs, respectively.
SPNet [2021] [9]	This study used the dataset of [324] comprising all the amino acids retrieved from the UniProt repository on 18 June 2019 and all proteins from the H. sapiens. The dataset comprised 16,210 unique proteins with a maximum length of 1,166 amino acids creating 104,262 total pairs. The training and validation sets contained 91,036 and 12,506 pairs, respectively, of which 33,318 and 6,094 pairs belonged to binding proteins (i.e., proteins with the potential to construct either transient or long-lived complexes). Two test sets were used in this study. Test-460 was a balanced strict set with 230 true positive and 230 true negative instances. Test-720 contained 260 true positive and 460 true negative instances. A 24-bin one-hot indicator represented each of the 20 standard amino acids and two stop codons, with the last two bins representing unknown or ambiguous amino acids.
BiLSTM-RF [2021] [291]	A nonredundant human dataset was retrieved from the DIP database. Sequences were clustered using the CD-HIT tool based on sequence similarity to remove redundancy and establish a nonredundant human PPI dataset [325,321]. This dataset included 4,262 interacting protein pairs and 3,899 noninteracting protein pairs.
Heterogeneous Network [2021] [296]	The 20 amino acids are divided into four groups based on their side chain polarity [286]: Ala, Val, Leu, Ile, Met, Phe, Trp, and Pro; Gly, Ser, Thr, Cys, Asn, Gln, and Tyr; Arg, Lys, and His; Asp and Glu. The protein sequences were simplified to a $4 \times 4 \times 4$ dimensional vector using the 3-mer method. Each vector dimension indicated the frequency of the amino acid sequence in the original protein sequence. Each dimension was initialised at zero. With a sliding window of length three, the whole protein sequence was scanned in steps of one. The amino acid sequence in the window was attached to the corresponding vector position in each step. Then, the vector was normalised. Finally, the vector obtained using 3-mers was an attribute feature.
OR-RCNN [2021] [301]	This study used datasets derived from the STRING database [319] for S. cerevisiae and H. sapiens. Each interaction was associated with a confidence score between zero (for noninteracting) and one (for interacting with the highest confidence). The confidence score interval was separated into K sub-intervals of equal length, where K = 20, (0, 0.05), [0.05, 0.1),·, [0.95, 1). The retrieved data was limited to protein sequences of length 50–2000. For S. cerevisiae, they randomly selected 5400 data points from each sub-interval, while H. sapiens included 5000 randomly selected data points for each sub-interval. The training dataset contained 90% of the data in each sub-interval, and the testing dataset contained the remaining 10%.
DeepTrio [2022] [302]	The training and testing datasets were obtained from the Biological General Repository for Interaction Datasets (BioGRID) [326] and the Database of Interacting Proteins (DIP) [318,112]. The BioGRID database contains PPIs derived from multiple major species based on the criteria that interacting pairs must be validated by at least two different experimental systems or published sources. The S. cerevisiae and H. sapiens benchmark datasets from BioGRID were used for training. Protein sequences were obtained from the UniProt [327] and restricted to lengths of 150–1,500 amino acids. The S. cerevisiae dataset contained 255 pairs after removing proteins >1,500 amino acids. The PIPR's dataset [299] comprised 231 pairs after proteins longer than 2,000 amino acids were removed.

is often laborious and limited in the types of protein they can model [338–340,92].

Machine learning algorithms offer an alternative and possibly complementary approach capable of using the information available in protein sequence and structure databases. The information about the structural and biophysical constraints on the amino acid sequence within functional proteins can be found in natural sequence variation. However, these data are not labelled, which presents a challenge for straightforward supervised learning techniques. This is where generative modelling methods show promise due to their capability to exploit these data unsupervised.

A GAN-based data augmentation approach, FFPred-GAN, has been proposed in [341] to tackle the protein function prediction problem by learning the distribution of protein amino acid sequence-based biophysical features and producing high-quality artificial protein feature samples. In the presence of auxiliary information, generative models can conduct the generative process by modelling the data distribution conditioned on the auxiliary variables. In particular, designing a protein may entail preserving a special function while modifying a property such as stability or solubility. An example of these models is conditional GAN [160]. Such a generative framework is proposed in [177], which learns a conditional generative model for protein sequences by considering a certain target structure represented by a graph over the R-group of amino acids. A VAE-based method is developed in [342] to generate novel variants of bacterial luciferase, an enzyme that emits light through the oxidation of flavin mononucleotide (FMNH2). A combination of reinforcement learning and RNNs is proposed in [343] to generate optimal molecules for biological activity.

7.2. Structure Design

Generative models for protein structures and modelling have been studied in [177,344], among which [104,345,346] have employed neural network-based models for sequences given their 3-D structure, modelling the amino acids independently from each other. Deep generative models have enabled new and viable protein structures [347]. Predicting the missing segments of corrupted protein structures can also be achieved using GAN, as represented by [173], in which the training data is restricted to structural information about the distances between adjacent α-carbons on the protein backbone.

7.3. Sequence Design

In addition to developing structure-based models, deep generative models have gained considerable attention for analysing protein sequences in individual protein families [348–351]. Even

though these approaches have proven effective, they require that a large number of sequences from a particular family are already available. This assumption cannot be met when designing novel proteins that diverge significantly from natural sequences, owing to an unbalanced dataset, non-interacting proteins [177]. ProteinGAN [352] is a GAN-based method with a customized temporal convolutional network [353] and self-attention mechanism [354] that aims to learn vital long-range inter-residue interactions and sequence motifs, as well as focusing on functional areas [355].

A conditional variational autoencoder (CVAE) model was developed in [356] to design protein sequences conditioned on a 1-D, context-free, grammar-based specification for folding topology. In [357,358], the conditional distributions of single amino acids are modelled, considering the encompassing structure and sequence context of the given protein, using convolutional neural networks. The generative model proposed in [177] is a graph-based model that captures the joint distribution of the full protein sequence, established on long-range interactions resulting from the protein 3-D structure.

Building on several recent successful studies using deep learning methods in modelling protein sequences such as contact prediction [359], prediction of secondary structure [360,361], and prediction of the fitness effects of mutations [348], generative modelling methods have begun to show potential for designing new sequences [177,349,362,356,363,350,364,365]. A LSTM is used in [366] as a generative approach in terms of amino acid sequences and peptide *de novo design*.

8. Conclusion

In the last decade, advancements in deep learning algorithms and GPUs as accelerators for high-performance computing have facilitated resolving intricate problems [367] concerning protein–protein and protein–ligand interaction, and drug discovery. This review offers an outline of protein structures and how they interact with other proteins, towards understanding their wide range of functionalities. Additionally, we outlined several deep learning methods and their applications to predicting protein–protein interaction, new drug delivery methods, and the improvement of existing solutions.

The available datasets for deep learning methods can be divided into structure-based and sequence-based. However, there is more sequential information available for proteins than there is 3-D structural information, thus driving progress in the development of sequence-based methods [368]. In fact, all the vital information required to identify PPIs is encoded in the proteins' amino acid sequence [369]. Several studies have been conducted by combining structural and sequential information.

Nevertheless, the viability of these techniques is yet to be verified experimentally. Continuing work is needed, such as analysing the strengths and limitations of different methods and the possibilities for incorporation into existing engineering operations. First and foremost, representation of the proteins to the network is a matter of importance [342]. This issue has been tackled using graph-based representations to model protein sequences and their 3-D structures [177,365]. Additionally, we may bolster our limited knowledge of protein folding mechanisms using deep reinforcement learning methods, aiming to find possible trajectories from extended protein chains to well-folded protein structures [355].

Following the Covid-19 pandemic and the dire need for rapid and reliable methods to create vaccines, one may see the potential of deep learning methods for solving such problems [370]. Additionally, a range of neurodegenerative diseases, infectious diseases, and cancers are closely related to abnormal protein–protein interactions [371–373]. Therefore, identifying protein–protein interactions using deep learning methods helps pave the way towards developing new drugs and targeted therapeutic approaches [8,14,374].

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.

Acknowledgement

All persons who have made substantial contributions to the work reported in the manuscript (e.g., technical help, writing and editing assistance, general support), but who do not meet the criteria for authorship, are named in the Acknowledgements and have given us their written permission to be named. If we have not included an Acknowledgements, then that indicates that we have not received substantial contributions from non-authors.

References

[1] M.A. Clark, J. Choi, Biology (2018).
[2] Zhang F, Shi W, Zhang J, Zeng M, Li M, Kurgan L. Probselect: accurate prediction of protein-binding residues from proteins sequences via dynamic predictor selection. Bioinformatics 2020;36:i735–44.
[3] Chatr-Aryamontri A, Ceol A, Licata L, Cesareni G. Protein interactions: integration leads to belief. Trends in Biochemical Sciences 2008;33(6). 241–241.
[4] De Las Rivas J, Fontanillo C. Protein–protein interactions essentials: key concepts to building and analyzing interactome networks. PLoS Computational Biology 2010;6(6):e1000807.
[5] Mackay JP, Sunde M, Lowry JA, Crossley M, Matthews JM. Protein interactions: is seeing believing? Trends in Biochemical Sciences 2007;32(12):530–1.
[6] Zhao Z, Gong X. Protein-protein interaction interface residue pair prediction based on deep learning architecture. IEEE/ACM Transactions on Computational Biology and Bioinformatics 2017;16:1753–9.
[7] Zhao J, Cao Y, Zhang L. Exploring the computational methods for protein-ligand binding site prediction. Computational and structural biotechnology journal 2020;18:417–26.
[8] Lu H, Zhou Q, He J, Jiang Z, Peng C, Tong R, Shi J. Recent advances in the development of protein–protein interactions modulators: mechanisms and clinical trials. Signal transduction and targeted therapy 2020;5(1):1–23.
[9] Wu J, Paquet E, Viktor HL, Michalowski W. Paying attention: Using a siamese pyramid network for the prediction of protein-protein interactions with folding and self-binding primary sequences. International Joint Conference on Neural Networks (IJCNN) 2021;2021:1–8. https://doi.org/10.1109/IJCNN52387.2021.9534212.
[10] Jiang P, Huang S, Fu Z, Sun Z, Lakowski TM, Hu P. Deep graph embedding for prioritizing synergistic anticancer drug combinations, Computational and Structural. Biotechnology Journal 2020;18:427–38.
[11] Gao W, Mahajan SP, Sulam J, Gray JJ. Deep learning in protein structural modeling and design. Patterns 2020:100142.
[12] Jones S, Thornton JM. Principles of protein-protein interactions. Proceedings of the National Academy of Sciences 1996;93(1):13–20.
[13] Zhou H-X, Shan Y. Prediction of protein interaction sites from sequence profile and residue neighbor list, Proteins: Structure. Function, and Bioinformatics 2001;44(3):336–43.
[14] Skrabanek L, Saini HK, Bader GD, Enright AJ. Computational prediction of protein–protein interactions. Molecular biotechnology 2008;38(1):1–17.
[15] Z.-H. You, L. Zhu, C.-H. Zheng, H.-J. Yu, S.-P. Deng, Z. Ji, Prediction of protein-protein interactions from amino acid sequences using a novel multi-scale continuous and discontinuous feature set, in: BMC bioinformatics, Vol. 15, Springer, 2014, pp. 1–9.
[16] Sandhya S, Mudgal R, Kumar G, Sowdhamini R, Srinivasan N. Protein sequence design and its applications. Current Opinion in Structural Biology 2016;37:71–80.
[17] S. Nivedha, S. Bhavani, A survey on prediction of protein-protein interactions, in: Journal of Physics: Conference Series, Vol. 1937, IOP Publishing, 2021, p. 012011.
[18] 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):1–8.
[19] Wang Y, You Z, Li L, Chen Z. A survey of current trends in computational predictions of protein-protein interactions. Frontiers of Computer Science 2020;14(4):1–12.
[20] Phizicky EM, Fields S. Protein-protein interactions: methods for detection and analysis. Microbiological reviews 1995;59(1):94–123.

[21] Zhang C, Zheng W, Cheng M, Omenn GS, Freddolino PL, Zhang Y. Functions of essential genes and a scale-free protein interaction network revealed by structure-based function and interaction prediction for a minimal genome. Journal of proteome research 2021;20(2):1178–89.
[22] Chirgadze Y, Boshkova E, Kargatov A, Chirgadze N. Functional identification of 'hypothetical protein-structures with unknown function. Journal of Biomolecular Structure and Dynamics 2022:1–5.
[23] Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M. bayesian networks approach for predicting protein-protein interactions from genomic data. Science 2003;302:449–53.
[24] Wang T, Li L, Huang Y-A, Zhang H, Ma Y, Zhou X. Prediction of protein-protein interactions from amino acid sequences based on continuous and discrete wavelet transform features. Molecules 2018;23(4):823.
[25] Schwikowski B, Uetz P, Fields S. A network of protein–protein interactions in yeast. Nature biotechnology 2000;18(12):1257–61.
[26] Ito T, Tashiro K, Muta S, Ozawa R, Chiba T, Nishizawa M, Yamamoto K, Kuhara S, Sakaki Y. Toward a protein–protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. Proceedings of the National Academy of Sciences 2000;97(3):1143–7.
[27] Gavin A-C, Bösche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon A-M, Cruciat C-M, et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 2002;415 (6868):141–7.
[28] Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams S-L, Millar A, Taylor P, Bennett K, Boutilier K, et al. Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry. Nature 2002;415 (6868):180–3.
[29] Pandey A, Mann M. Proteomics to study genes and genomes. Nature 2000;405(6788):837–46.
[30] Figeys D. Novel approaches to map protein interactions. Current opinion in biotechnology 2003;14(1):119–25.
[31] Noor Z, Ahn SB, Baker MS, Ranganathan S, Mohamedali A. Mass spectrometry–based protein identification in proteomics–a review. Briefings in bioinformatics 2021;22(2):1620–38.
[32] Garza KY, Feider CL, Klein DR, Rosenberg JA, Brodbelt JS, Eberlin LS. Desorption electrospray ionization mass spectrometry imaging of proteins directly from biological tissue sections. Analytical chemistry 2018;90 (13):7785–9.
[33] MacBeath G, Schreiber SL. Printing proteins as microarrays for high-throughput function determination. Science 2000;289(5485):1760–3.
[34] Büssow K, Nordhoff E, Lübbert C, Lehrach H, Walter G. A human cdna library for high-throughput protein expression screening. Genomics 2000;65(1):1–8. https://doi.org/10.1006/geno.2000.6141, https://www.sciencedirect.com/science/article/pii/S088875430096141X.
[35] Brizuela L, Braun P, LaBaer J. Flexgene repository: from sequenced genomes to gene repositories for high-throughput functional biology and proteomics. Molecular and biochemical parasitology 2001;118(2):155–65.
[36] Brizuela L, Richardson A, Marsischky G, Labaer J. The flexgene repository: exploiting the fruits of the genome projects by creating a needed resource to face the challenges of the post-genomic era. Archives of medical research 2002;33(4):318–24.
[37] A. Droit, G.G. Poirier, J.M. Hunter, Experimental and bioinformatic approaches for interrogating protein–protein interactions to determine protein function, Journal of Molecular Endocrinology 34 (2) (01 Apr. 2005) 263–280.
[38] Zhou M, Li Q, Wang R. Current experimental methods for characterizing protein–protein interactions. ChemMedChem 2016;11(8):738.
[39] Piehler J. New methodologies for measuring protein interactions in vivo and in vitro. Current Opinion in Structural Biology 2005;15(1):4–14.
[40] Rao VS, Srinivas K, Sujini G, Kumar G. Protein-protein interaction detection: methods and analysis. International journal of proteomics 2014.
[41] Ding Z, Kihara D. Computational identification of protein-protein interactions in model plant proteomes. Scientific reports 2019;9(1):1–13.
[42] S. Tsukiyama, M.M. Hasan, S. Fujii, H. Kurata, Lstm-phv: Prediction of human-virus protein-protein interactions by lstm with word2vec, bioRxiv.
[43] Lodish H, Zipursky SL. Molecular cell biology. Biochem Mol Biol Educ 2001;29:126–33.
[44] R.H. Garrett, Biochemistry, Cengage Learning Canada Inc, 2015.
[45] A. Bagchi, Protein-protein interactions: Basics, characteristics, and predictions, in: Soft Computing for Biological Systems, Springer, 2018, pp. 111–120.
[46] Go N, Taketomi H. Respective roles of short-and long-range interactions in protein folding. Proceedings of the National Academy of Sciences 1978;75 (2):559–63.
[47] Gromiha MM, Selvaraj S. Importance of long-range interactions in protein folding. Biophysical chemistry 1999;77(1):49–68.
[48] Melkikh AV, Meijer DK. On a generalized levinthal's paradox: The role of long- and short range interactions in complex bio-molecular reactions, including protein and dna folding. Progress in Biophysics and Molecular Biology 2018;132:57–79.
[49] Gromiha MM, Selvaraj S. Influence of medium and long range interactions in different structural classes of globular proteins. Journal of Biological Physics 1997;23(3):151–62.
[50] Buxbaum E. Fundamentals of protein structure and function, Vol. 31. Springer; 2007.
[51] Maloy S. Amino acids. In: Maloy S, Hughes K, editors. Brenner's Encyclopedia of Genetics (Second Edition). Edition: Academic Press, San Diego; 2013. p. 108–10.
[52] Yang L, Han Y, Zhang H, Li W, Dai Y. Prediction of protein-protein interactions with local weight-sharing mechanism in deep learning. BioMed Research International 2020.
[53] Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. Journal of molecular biology 1982;157(1):105–32.
[54] Meiler J, Müller M, Zeidler A, Schmäschke F. Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks. Molecular modeling annual 2001;7(9):360–9.
[55] Pogliani L. Molecular connectivity model for determination of isoelectric point of amino acids. Journal of pharmaceutical sciences 1992;81(4):334–6.
[56] Biro J. Amino acid size, charge, hydropathy indices and matrices for protein structure analysis. Theoretical Biology and Medical Modelling 2006;3 (1):1–12.
[57] Fauchère J-L, Charton M, Kier LB, Verloop A, Pliska V. Amino acid side chain parameters for correlation studies in biology and pharmacology. International journal of peptide and protein research 1988;32(4):269–78.
[58] Holm L, Sander C. Mapping the protein universe. Science 1996;273 (5275):595–602.
[59] K.-Y. Law, Definitions for hydrophilicity, hydrophobicity, and superhydrophobicity: getting the basics right (2014).
[60] Jha K, Saha S, Tanveer M. Prediction of protein-protein interactions using stacked auto-encoder. Transactions on Emerging Telecommunications Technologies 2021:e4256.
[61] Lu Y, Freeland S. On the evolution of the standard amino-acid alphabet. Genome biology 2006;7(1):1–6.
[62] Schmidt RL, Simonović M. Synthesis and decoding of selenocysteine and human health. Croatian medical journal 2012;53(6):535–50.
[63] Zhang Y, Baranov PV, Atkins JF, Gladyshev VN. Pyrrolysine and selenocysteine use dissimilar decoding strategies. Journal of Biological Chemistry 2005;280 (21):20740–51.
[64] Hatfield DL, Gladyshev VN. How selenium has altered our understanding of the genetic code. Molecular and cellular biology 2002;22(11):3565–76.
[65] Turanov AA, Xu X-M, Carlson BA, Yoo M-H, Gladyshev VN, Hatfield DL. Biosynthesis of selenocysteine, the 21st amino acid in the genetic code, and a novel pathway for cysteine biosynthesis. Advances in nutrition 2011;2 (2):122–8.
[66] Gdr HB, Sharon N, Australia EW. Nomenclature and symbolism for amino acids and peptides. Pure and Applied Chemistry 1984;56:595–624.
[67] Biochemistry Human. Elsevier 2018. https://doi.org/10.1016/c2009-0-63992-1.
[68] Alhazmi HAM. Mobility shift-affinity capillary electrophoresis for investigation of protein-metal ion interactions: aspects of method development, validation and high throughput screening. Ph.D. thesis 2015. https://doi.org/10.24355/dbbs.084-201506241157-0.
[69] Tahir M, Khan F, Hayat M, Alshehri MD. An effective machine learning-based model for the prediction of protein–protein interaction sites in health systems. Neural Computing and Applications 2022:1–11.
[70] Keenleyside W. Microbiology: Canadian Edition, Pressbooks 2019.
[71] Shoulders MD, Raines RT. Collagen structure and stability. Annual review of biochemistry 2009;78:929–58.
[72] Stegemann H, Stalder K. Determination of hydroxyproline. Clinica chimica acta 1967;18(2):267–73.
[73] E. Paquet, H. Viktor, K. Madi, J. Wu, Deformable protein shape classification based on deep learning, and the fractional fokkerplanck and khlerdirac equations., IEEE Transactions on Pattern Analysis and Machine Intelligence.
[74] 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 (5649):1364–8.
[75] Fernandez-Fuentes N, Dybas JM, Fiser A. Structural characteristics of novel protein folds. PLoS Computational Biology 2010;6(4):e1000750.
[76] Jiang L, Althoff EA, Clemente FR, Doyle L, Röthlisberger D, Zanghellini A, Gallaher JL, Betker JL, Tanaka F, Barbas CF, et al. De novo computational design of retro-aldol enzymes. science 2008;319(5868):1387–91.
[77] Röthlisberger D, Khersonsky O, Wollacott AM, Jiang L, DeChancie J, Betker J, Gallaher JL, Althoff EA, Zanghellini A, Dym O, et al. Kemp elimination catalysts by computational enzyme design. Nature 2008;453:190–5.
[78] Correia BE, Ban Y-EA, Holmes MA, Xu H, Ellingson K, Kraft Z, Carrico C, Boni E, Sather DN, Zenobia C, et al. Computational design of epitope-scaffolds allows induction of antibodies specific for a poorly immunogenic hiv vaccine epitope. Structure 2010;18(9):1116–26.
[79] Leaver-Fay A, Froning KJ, Atwell S, Aldaz H, Pustilnik A, Lu F, Huang F, Yuan R, Hassanali S, Chamberlain AK, et al. Computationally designed bispecific antibodies using negative state repertoires. Structure 2016;24(4):641–51.
[80] Lewis SM, Wu X, Pustilnik A, Sereno A, Huang F, Rick HL, Guntas G, Leaver-Fay A, Smith EM, Ho C, et al. Generation of bispecific igg antibodies by structure-based design of an orthogonal fab interface. Nature biotechnology 2014;32 (2):191–8.
[81] Correia BE, Bates JT, Loomis RJ, Baneyx G, Carrico C, Jardine JG, Rupert P, Correnti C, Kalyuzhniy O, Vittal V, et al. Proof of principle for epitope-focused vaccine design. Nature 2014;507(7491):201–6.
[82] Tinberg CE, Khare SD, Dou J, Doyle L, Nelson JW, Schena A, Jankowski W, Kalodimos CG, Johnsson K, Stoddard BL, et al. Computational design of ligand-

binding proteins with high affinity and selectivity. Nature 2013;501 (7466):212–6.
[83] Zhou L, Bosscher M, Zhang C, Özçubukçu S, Zhang L, Zhang W, Li CJ, Liu J, Jensen MP, Lai L, et al. A protein engineered to bind uranyl selectively and with femtomolar affinity. Nature chemistry 2014;6(3):236–41.
[84] King NP, Sheffler W, Sawaya MR, Vollmar BS, Sumida JP, André I, Gonen T, Yeates TO, Baker D. Computational design of self-assembling protein nanomaterials with atomic level accuracy. Science 2012;336(6085):1171–4.
[85] King NP, Bale JB, Sheffler W, McNamara DE, Gonen S, Gonen T, Yeates TO, Baker D. Accurate design of co-assembling multi-component protein nanomaterials. Nature 2014;510(7503):103–8.
[86] Gonen S, DiMaio F, Gonen T, Baker D. Design of ordered two-dimensional arrays mediated by noncovalent protein-protein interfaces. Science 2015;348 (6241):1365–8.
[87] Bale JB, Gonen S, Liu Y, Sheffler W, Ellis D, Thomas C, Cascio D, Yeates TO, Gonen T, King NP, et al. Accurate design of megadalton-scale two-component icosahedral protein complexes. Science 2016;353(6297):389–94.
[88] Samish I. Computational protein design. Springer; 2017.
[89] Korendovych IV, Senes A, Kim YH, Lear JD, Fry HC, Therien MJ, Blasie JK, Walker FA, DeGrado WF. De novo design and molecular assembly of a transmembrane diporphyrin-binding protein complex. Journal of the American Chemical Society 2010;132(44):15516–8.
[90] Joh NH, Wang T, Bhate MP, Acharya R, Wu Y, Grabe M, Hong M, Grigoryan G, DeGrado WF. De novo design of a transmembrane zn2+-transporting four-helix bundle. Science 2014;346(6216):1520–4.
[91] Zhang Y, Bartz R, Grigoryan G, Bryant M, Aaronson J, Beck S, Innocent N, Klein L, Procopio W, Tucker T, et al. Computational design and experimental characterization of peptides intended for ph-dependent membrane insertion and pore formation. ACS chemical biology 2015;10(4):1082–93.
[92] Huang P-S, Boyken SE, Baker D. The coming of age of de novo protein design. Nature 2016;537(7620):320–7.
[93] Norn CH, André I. Computational design of protein self-assembly. Current Opinion in Structural Biology 2016;39:39–45.
[94] Liu H, Chen Q. Computational protein design for given backbone: recent progresses in general method-related aspects. Current Opinion in Structural Biology 2016;39:89–95.
[95] Yang W, Lai L. Computational design of ligand-binding proteins. Current Opinion in Structural Biology 2017;45:67–73.
[96] Dima RI, Banavar JR, Maritan A. Scoring functions in protein folding and design. Protein Science 2000;9(4):812–9.
[97] Li Z, Yang Y, Zhan J, Dai L, Zhou Y. Energy functions in de novo protein design: current challenges and future prospects. Annual review of biophysics 2013;42:315–35.
[98] Boas FE, Harbury PB. Potential energy functions for protein design. Current Opinion in Structural Biology 2007;17(2):199–204.
[99] Shapovalov MV, Dunbrack Jr RL. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. Structure 2011;19(6):844–58.
[100] Doyle L, Hallinan J, Bolduc J, Parmeggiani F, Baker D, Stoddard BL, Bradley P. Rational design of α)helical tandem repeat proteins with closed architectures. Nature 2015;528(7583):585–8.
[101] Bhardwaj G, Mulligan VK, Bahl CD, Gilmore JM, Harvey PJ, Cheneval O, Buchko GW, Pulavarti SV, Kaas Q, Eletsky A, et al. Accurate de novo design of hyperstable constrained peptides. Nature 2016;538(7625):329–35.
[102] Broom A, Trainor K, MacKenzie DW, Meiering EM. Using natural sequences and modularity to design common and novel protein topologies. Current Opinion in Structural Biology 2016;38:26–36.
[103] Khersonsky O, Fleishman SJ. Why reinvent the wheel? building new proteins based on ready-made parts. Protein Science 2016;25(7):1179–87.
[104] Wang J, Cao H, Zhang JZ, Qi Y. Computational protein design with deep learning neural networks. Scientific reports 2018;8(1):1–9.
[105] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521 (7553):436–44.
[106] T.U. Consortium, UniProt: the universal protein knowledgebase in 2021, Nucleic Acids Research 49 (D1) D480–D489. doi:10.1093/nar/gkaa1100.
[107] Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, Heer FT, de Beer TAP, Rempfer C, Bordoli L, et al. Swiss-model: homology modelling of protein structures and complexes. Nucleic acids research 2018;46(W1):W296–303.
[108] Blohm P, Frishman G, Smialowski P, Goebels F, Wachinger B, Ruepp A, Frishman D. Negatome 2.0: a database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis. Nucleic acids research 2014;42(D1):D396–400.
[109] Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, Doncheva NT, Legeay M, Fang T, Bork P, et al. The string database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic acids research 2021;49(D1):D605–12.
[110] Bittrich S, Rose Y, Segura J, Lowe R, Westbrook JD, Duarte JM, Burley SK. Rcsb protein data bank: improved annotation, search and visualization of membrane protein structures archived in the pdb. Bioinformatics 2022;38 (5):1452–4.
[111] Oughtred R, Rust J, Chang C, Breitkreutz B-J, Stark C, Willems A, Boucher L, Leung G, Kolas N, Zhang F, et al. The biogrid database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. Protein Science 2021;30(1):187–200.
[112] Xenarios I, Salwinski L, Duan XJ, Higney P, Kim S-M, Eisenberg D. Dip, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. Nucleic acids research 2002;30(1):303–5.
[113] Bader GD, Betel D, Hogue CW. Bind: the biomolecular interaction network database. Nucleic acids research 2003;31(1):248–50.
[114] Calderone A, Iannuccelli M, Peluso D, Licata L. Using the mint database to search protein interactions. Current Protocols in Bioinformatics 2020;69(1): e93.
[115] Keshava Prasad T, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, et al. Human protein reference database–2009 update. Nucleic acids research 2009;37(suppl_1):D767–72.
[116] Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, Campbell NH, Chavali G, Chen C, Del-Toro N, et al. The mintact project–intact as a common curation platform for 11 molecular interaction databases. Nucleic acids research 2014;42(D1):D358–63.
[117] Hu L, Wang X, Huang Y-A, Hu P, You Z-H. A survey on computational models for predicting protein–protein interactions. Briefings in Bioinformatics 2021;22(5):bbab036.
[118] Anfinsen CB. Principles that govern the folding of protein chains. Science 1973;181(4096):223–30.
[119] Panchal GK, Das S, Sakure A, Singh BP, Hati S. Production and characterization of antioxidative peptides during lactic fermentation of goat milk. Journal of Food Processing and Preservation 2021;45(12):e15992.
[120] V.K. Chaturvedi, D. Mishra, A. Tiwari, V. Snijesh, N.A. Shaik, M. Singh, Sequence databases, in: Essentials of Bioinformatics, Volume I, Springer, 2019, pp. 29–46.
[121] Rose Y, Duarte JM, Lowe R, Segura J, Bi C, Bhikadiya C, Chen L, Rose AS, Bittrich S, Burley SK, et al. Rcsb protein data bank: architectural advances towards integrated searching and efficient access to macromolecular structure data from the pdb archive. Journal of molecular biology 2021;433(11):166704.
[122] Andreeva A, Kulesha E, Gough J, Murzin AG. The scop database in 2020: expanded classification of representative family and superfamily domains of known protein structures. Nucleic acids research 2020;48(D1): D376–82.
[123] Dandekar T, Snel B, Huynen M, Bork P. Conservation of gene order: a fingerprint of proteins that physically interact. Trends in Biochemical Sciences 1998;23(9):324–8.
[124] Nishimoto AT, Sharma C, Rogers PD. Molecular and genetic basis of azole antifungal resistance in the opportunistic pathogenic fungus candida albicans. Journal of Antimicrobial Chemotherapy 2020;75(2):257–70.
[125] Y. Bengio, L. Yao, G. Alain, P. Vincent, Generalized denoising auto-encoders as generative models, arXiv preprint arXiv:1305.6663.
[126] Bengio Y, Goodfellow IJ, Courville A. Deep learning. Nature 2015;521 (7553):436–44.
[127] Soleymani F, Paquet E. Financial portfolio optimization with online deep reinforcement learning and restricted stacked autoencoder–deepbreath. Expert Systems with Applications 2020;156:113456.
[128] Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult. IEEE transactions on neural networks 1994;5 (2):157–66.
[129] Hochreiter S, Bengio Y, Frasconi P, Schmidhuber J, et al. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies, 2001.
[130] Hochreiter S, Schmidhuber J. Long short-term memory. Neural computation 1997;9(8):1735–80.
[131] James J, Lam AY, Hill DJ, Li VO. Delay aware intelligent transient stability assessment system. IEEE Access 2017;5:17230–9.
[132] Srivastava N, Mansimov E, Salakhudinov R. Unsupervised learning of video representations using lstms. In: in: International conference on machine learning PMLR. p. 843–52.
[133] 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.
[134] M. Henaff, J. Bruna, Y. LeCun, Deep convolutional networks on graph-structured data, arXiv preprint arXiv:1506.05163.
[135] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, P.S. Yu, A comprehensive survey on graph neural networks, arXiv preprint arXiv:1901.00596.
[136] Zhang S, Tong H, Xu J, Maciejewski R. Graph convolutional networks: a comprehensive review. Computational Social Networks 2019;6(1):11.
[137] Shuman DI, Narang SK, Frossard P, Ortega A, Vandergheynst P. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. IEEE signal processing magazine 2013;30(3):83–98.
[138] Soleymani F, Paquet E. Deep graph convolutional reinforcement learn-ing for financial portfolio management–deeppocket. Expert Systems withApplications 2021;182:115127.
[139] J. Bruna, W. Zaremba, A. Szlam, Y. LeCun, Spectral networks and locally connected networks on graphs, arXiv preprint arXiv:1312.6203.
[140] Shuman DI, Vandergheynst P, Frossard P. Chebyshev polynomial approximation for distributed signal processing. In: in: 2011 International Conference on Distributed Computing in Sensor Systems and Workshops (DCOSS) IEEE. p. 1–8.
[141] M. Defferrard, X. Bresson, P. Vandergheynst, Convolutional neural networks on graphs with fast localized spectral filtering, in: Advances in neural information processing systems, 2016, pp. 3844–3852.

[142] Ullah I, Manzo M, Shah M, Madden MG. Graph convolutional networks: analysis, improvements and results. Applied Intelligence 2022;52 (8):9033–44.
[143] I. Goodfellow, Nips 2016 tutorial: Generative adversarial networks, arXiv preprint arXiv:1701.00160.
[144] Soleymani F, Paquet E. Long-term financial predictions based on feynman–dirac path integrals, deep bayesian networks and temporal generative adversarial networks. Machine Learning with Applications 2022;7:100255.
[145] D.P. Kingma, M. Welling, Auto-encoding variational bayes, arXiv preprint arXiv:1312.6114.
[146] Rezende DJ, Mohamed S, Wierstra D. Stochastic backpropagation and approximate inference in deep generative models. In: in: International conference on machine learning PMLR. p. 1278–86.
[147] C. Doersch, Tutorial on variational autoencoders, arXiv preprint arXiv:1606.05908.
[148] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, arXiv preprint arXiv:1406.2661.
[149] Bengio Y, Ducharme R, Vincent P, Janvin C. A neural probabilistic language model. The journal of machine learning research 2003;3:1137–55.
[150] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, K. Kavukcuoglu, Wavenet: A generative model for raw audio, arXiv preprint arXiv:1609.03499.
[151] A. Van Oord, N. Kalchbrenner, K. Kavukcuoglu, Pixel recurrent neural networks, in: International Conference on Machine Learning, PMLR, 2016, pp. 1747–1756.
[152] Choi Y, Choi M, Kim M, Ha J-W, Kim S, Choo J. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: in: Proceedings of the IEEE conference on computer vision and pattern recognition. p. 8789–97.
[153] Yang G, Yu S, Dong H, Slabaugh G, Dragotti PL, Ye X, Liu F, Arridge S, Keegan J, Guo Y, et al. Dagan: Deep de-aliasing generative adversarial networks for fast compressed sensing mri reconstruction. IEEE transactions on medical imaging 2017;37(6):1310–21.
[154] Seeliger K, Güçlü U, Ambrogioni L, Güçlütürk Y, van Gerven MA. Generative adversarial networks for reconstructing natural images from brain activity. NeuroImage 2018;181:775–85.
[155] D.P. Kingma, M. Welling, An introduction to variational autoencoders, arXiv preprint arXiv:1906.02691.
[156] S. Khobahi, M. Soltanalian, Model-aware deep architectures for one-bit compressive variational autoencoding, arXiv preprint arXiv:1911.12410.
[157] An J, Cho S. Variational autoencoder based anomaly detection using reconstruction probability. Special Lecture on IE 2015;2(1):1–18.
[158] Kingma DP, Salimans T, Jozefowicz R, Chen X, Sutskever I, Welling M. Improved variational inference with inverse autoregressive flow. Advances in neural information processing systems 2016;29:4743–51.
[159] V. Dumoulin, F. Visin, A guide to convolution arithmetic for deep learning, arXiv preprint arXiv:1603.07285.
[160] M. Mirza, S. Osindero, Conditional generative adversarial nets, arXiv preprint arXiv:1411.1784.
[161] Chen Y, Xu D. Computational analyses of high-throughput protein-protein interaction data. Current Protein and Peptide Science 2003;4(3):159–80.
[162] Yakubu RR, Nieves E, Weiss LM. The methods employed in mass spectrometric analysis of posttranslational modifications (ptms) and protein–protein interactions (ppis). Advancements of mass spectrometry in biomedical research 2019:169–98.
[163] Lenz S, Sinn LR, O'Reilly FJ, Fischer L, Wegner F, Rappsilber J. Reliable identification of protein-protein interactions by crosslinking mass spectrometry. Nature communications 2021;12(1):1–11.
[164] Yugandhar K, Gupta S, Yu H. Inferring protein-protein interaction networks from mass spectrometry-based proteomic approaches: a mini-review, Computational and Structural. Biotechnology Journal 2019;17:805–11.
[165] S. Sledzieski, R. Singh, L. Cowen, B. Berger, Sequence-based prediction of protein-protein interactions: a structure-aware interpretable deep learning model, bioRxiv.
[166] Bakail M, Ochsenbein F. Targeting protein–protein interactions, a wide open field for drug design. Comptes Rendus Chimie 2016;19(1–2):19–27.
[167] Murakami Y, Tripathi LP, Prathipati P, Mizuguchi K. Network analysis and in silico prediction of protein–protein interactions with applications in drug discovery. Current Opinion in Structural Biology 2017;44:134–42.
[168] Marchand A, Van Hall-Beauvais AK, Correia BE. Computational design of novel protein–protein interactions–an overview on methodological approaches and applications. Current Opinion in Structural Biology 2022;74:102370.
[169] Siegel JB, Zanghellini A, Lovick HM, Kiss G, Lambert AR, Clair JLS, Gallaher JL, Hilvert D, Gelb MH, Stoddard BL, et al. Computational design of an enzyme catalyst for a stereoselective bimolecular diels-alder reaction. Science 2010;329(5989):309–13.
[170] Whitehead TA, Chevalier A, Song Y, Dreyfus C, Fleishman SJ, De Mattos C, Myers CA, Kamisetty H, Blair P, Wilson IA, et al. Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. Nature biotechnology 2012;30:543–8.
[171] Strauch E-M, Bernard SM, La D, Bohn AJ, Lee PS, Anderson CE, Nieusma T, Holstein CA, Garcia NK, Hooper KA, et al. Computational design of trimeric influenza-neutralizing proteins targeting the hemagglutinin receptor binding site. Nature biotechnology 2017;35:667–71.
[172] Smart AD, Pache RA, Thomsen ND, Kortemme T, Davis GW, Wells JA. Engineering a light-activated caspase-3 for precise ablation of neurons in vivo. Proceedings of the National Academy of Sciences 2017;114(39): E8174–83.
[173] Anand N, Huang P-S. Generative modeling for protein structures. In: in: Proceedings of the 32nd International Conference on Neural Information Processing Systems. p. 7505–16.
[174] Sormani G, Harteveld Z, Rosset S, Correia B, Laio A. A rosetta-based protein design protocol converging to natural sequences. The Journal of Chemical Physics 2021;154(7):074114.
[175] Koga N, Tatsumi-Koga R, Liu G, Xiao R, Acton TB, Montelione GT, Baker D. Principles for designing ideal protein structures. Nature 2012;491 (7423):222–7.
[176] Rocklin GJ, Chidyausiku TM, Goreshnik I, Ford A, Houliston S, Lemak A, Carter L, Ravichandran R, Mulligan VK, Chevalier A, et al. Global analysis of protein folding using massively parallel design, synthesis, and testing. Science 2017;357(6347):168–75.
[177] J. Ingraham, V. Garg, R. Barzilay, T. Jaakkola, Generative models for graph-based protein design, Advances in Neural Information Processing Systems 32.
[178] Namrata A, Raphael E, Po-Ssu H. Fully differentiable full-atom protein backbone generation. In: Proceedings of the International Conference on Learning Representations (ICLR) Workshops.
[179] J. Ingraham, A. Riesselman, C. Sander, D. Marks, Learning protein structure with a differentiable simulator, in: International Conference on Learning Representations, 2018.
[180] AlQuraishi M. End-to-end differentiable learning of protein structure. Cell systems 2019;8(4):292–301.
[181] Dahiyat BI, Mayo SL. De novo protein design: fully automated sequence selection. Science 1997;278(5335):82–7.
[182] Kraemer-Pecore CM, Lecomte JT, Desjarlais JR. A de novo redesign of the ww domain. Protein Science 2003;12(10):2194–205.
[183] Russ WP, Lowery DM, Mishra P, Yaffe MB, Ranganathan R. Natural-like function in artificial ww domains. Nature 2005;437(7058):579–83.
[184] Ding Z, Kihara D. Computational methods for predicting protein-protein interactions using various protein features. Current protocols in protein science 2018;93(1):e62.
[185] Hosur R, Peng J, Vinayagam A, Stelzl U, Xu J, Perrimon N, Bienkowska J, Berger B. A computational framework for boosting confidence in high-throughput protein-protein interaction datasets. Genome biology 2012;13(8):1–14.
[186] Mirabello C, Wallner B. Interpred: a pipeline to identify and model protein–protein interactions, Proteins: Structure. Function, and Bioinformatics 2017;85(6):1159–70.
[187] Jha K, Saha S. Amalgamation of 3d structure and sequence information for protein–protein interaction prediction. Scientific Reports 2020;10 (1):1–14.
[188] Scarff CA, Thalassinos K, Hilton GR, Scrivens JH. Travelling wave ion mobility mass spectrometry studies of protein structure: biological significance and comparison with x-ray crystallography and nuclear magnetic resonance spectroscopy measurements. Rapid Communications in Mass Spectrometry: An International Journal Devoted to the Rapid Dissemination of Up-to-the-Minute Research in Mass Spectrometry 2008;22(20):3297–304.
[189] Neuvirth H, Raz R, Schreiber G. Promate: a structure based prediction program to identify the location of protein–protein binding sites. Journal of molecular biology 2004;338(1):181–99.
[190] Drewes G, Bouwmeester T. Global approaches to protein–protein interactions. Current opinion in cell biology 2003;15(2):199–205.
[191] Zhang B, Li J, Quan L, Chen Y, Lü Q. Sequence-based prediction of protein-protein interaction sites by simplified long short-term memory network. Neurocomputing 2019;357:86–100.
[192] Terentiev A, Moldogazieva N, Shaitan K. Dynamic proteomics in modeling of the living cell. protein-protein interactions. Biochemistry (Moscow) 2009;74 (13):1586–607.
[193] Brettner LM, Masel J. Protein stickiness, rather than number of functional protein-protein interactions, predicts expression noise and plasticity in yeast. BMC systems biology 2012;6(1):1–10.
[194] Wodak SJ, Vlasblom J, Turinsky AL, Pu S. Protein–protein interaction networks: the puzzling riches. Current Opinion in Structural Biology 2013;23(6):941–53.
[195] Hou Q, De Geest PF, 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.
[196] 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.
[197] Lu S, Li Y, Nan X, Zhang S. Attention-based convolutional neural networks for protein-protein interaction site prediction. In: 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE; 2021. p. 141–4.
[198] Bonetta L. Interactome under construction. Nature 2010;468(7325):851–2.
[199] Yuan Q, Chen J, Zhao H, Zhou Y, Yang Y. Structure-aware protein–protein interaction site prediction using deep graph convolutional network. Bioinformatics 2022;38(1):125–32.
[200] Hou Q, Lensink MF, Heringa J, Feenstra KA. Club-martini: selecting favourable interactions amongst available candidates, a coarse-grained simulation approach to scoring docking decoys. PloS one 2016;11(5):e0155251.

[201] B. Huang, M. Schroeder, Using protein binding site prediction to improve protein docking, Gene 422 (1) (2008) 14–21, physical and Chemical Foundations of Bioinformatics Methods. doi: 10.1016/j.gene.2008.06.014.
[202] Xie Z, Deng X, Shu K. Prediction of protein–protein interaction sites using convolutional neural network and improved data sets. International journal of molecular sciences 2020;21(2):467.
[203] Li Y, Golding GB, Ilie L. Delphi: accurate deep ensemble model for protein interaction sites prediction. Bioinformatics 2021;37(7):896–904.
[204] Landete J. Effector molecules and regulatory proteins: Applications. Trends in Biotechnology 2016;34(10):777–80.
[205] Jamasb AR, Day B, Cangea C, Liò P, Blundell TL. Deep learning for protein–protein interaction site prediction. Proteomics Data Analysis, Springer 2021:263–88.
[206] Zhang C, Freddolino PL, Zhang Y. Cofactor: improved protein function prediction by combining structure, sequence and protein–protein interaction information. Nucleic acids research 2017;45(W1):W291–9.
[207] Jubb H, Higueruelo AP, Winter A, Blundell TL. Structural biology and drug discovery for protein–protein interactions. Trends in pharmacological sciences 2012;33(5):241–8.
[208] Hoskins J, Lovell S, Blundell TL. An algorithm for predicting protein–protein interaction sites: abnormally exposed amino acid residues and secondary structure elements. Protein Science 2006;15(5):1017–29.
[209] Shi T-L, Li Y-X, Cai Y-D, Chou K-C. Computational methods for protein-protein interaction and their application. Current Protein and Peptide Science 2005;6(5):443–9.
[210] K. Al-Khafaji, T. Taskin-Tok, Computational techniques for studying protein-protein interactions, in: Advances in Protein Molecular and Structural Biology Methods, Elsevier, 2022, pp. 125–135.
[211] Vakser IA. Protein-protein docking: From interaction to interactome. Biophysical journal 2014;107(8):1785–93.
[212] Siebenmorgen T, Zacharias M. Computational prediction of protein–protein binding affinities, Wiley Interdisciplinary Reviews: Computational Molecular. Science 2020;10(3):e1448.
[213] H.J. Nussbaumer, The fast fourier transform, in: Fast Fourier Transform and Convolution Algorithms, Springer, 1981, pp. 80–111.
[214] Chaudhury S, Lyskov S, Gray JJ. Pyrosetta: a script-based interface for implementing molecular modeling algorithms using rosetta. Bioinformatics 2010;26(5):689–91.
[215] D. Varela, I. André, A memetic algorithm enables global all-atom protein-protein docking with sidechain flexibility, bioRxiv.
[216] Pierce BG, Wiehe K, Hwang H, Kim B-H, Vreven T, Weng Z. Zdock server: interactive docking prediction of protein–protein complexes and symmetric multimers. Bioinformatics 2014;30(12):1771–3.
[217] Mashiach E, Nussinov R, Wolfson HJ. Fiberdock: flexible induced-fit backbone refinement in molecular docking, Proteins: Structure. Function, and Bioinformatics 2010;78(6):1503–19.
[218] Gray JJ, Moughon S, Wang C, Schueler-Furman O, Kuhlman B, Rohl CA, Baker D. Protein–protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. Journal of molecular biology 2003;331(1):281–99.
[219] Wang C, Bradley P, Baker D. Protein–protein docking with backbone flexibility. Journal of molecular biology 2007;373(2):503–19.
[220] Chaudhury S, Berrondo M, Weitzner BD, Muthu P, Bergman H, Gray JJ. Benchmarking and analysis of protein docking performance in rosetta v3. 2, PloS one 2011;6(8):e22477.
[221] Ohue M, Matsuzaki Y, Uchikoga N, Ishida T, Akiyama Y. Megadock: an all-to-all protein-protein interaction prediction system using tertiary structure data. Protein and peptide letters 2014;21(8):766–78.
[222] Szilagyi A, Zhang Y. Template-based structure modeling of protein–protein interactions. Current Opinion in Structural Biology 2014;24:10–23.
[223] Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. Science 1997;278(5338):631–7.
[224] Tuncbag N, Gursoy A, Nussinov R, Keskin O. Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using prism. Nature protocols 2011;6(9):1341–54.
[225] Keskin O, Tsai C-J, Wolfson H, Nussinov R. A new, structurally nonredundant, diverse data set of protein–protein interfaces and its implications. Protein Science 2004;13(4):1043–55.
[226] Tuncbag N, Gursoy A, Keskin O. Prediction of protein–protein interactions: unifying evolution and structure at protein interfaces. Physical biology 2011;8(3):035006.
[227] Kundrotas PJ, Zhu Z, Janin J, Vakser IA. Templates are available to model nearly all complexes of structurally characterized proteins. Proceedings of the National Academy of Sciences 2012;109(24):9438–41.
[228] Fukuhara N, Kawabata T. Homcos: a server to predict interacting protein pairs and interacting sites by homology modeling of complex structures. Nucleic acids research 2008;36(suppl_2):W185–9.
[229] Ghoorah AW, Devignes M-D, Smaïl-Tabbone M, Ritchie DW. Spatial clustering of protein binding sites for template based protein docking. Bioinformatics 2011;27(20):2820–7.
[230] Zhang C, Zheng W, Mortuza S, Li Y, Zhang Y. Deepmsa: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. Bioinformatics 2020;36(7):2105–12.
[231] Mulnaes D, Porta N, Clemens R, Apanasenko I, Reiners J, Gremer L, Neudecker P, Smits SH, Gohlke H. Topmodel: template-based protein structure prediction at low sequence identity using top-down consensus and deep neural networks. Journal of chemical theory and computation 2020;16(3):1953–67.
[232] Yan Y, Wen Z, Wang X, Huang S-Y. Addressing recent docking challenges: A hybrid strategy to integrate template-based and free protein-protein docking, Proteins: Structure. Function, and Bioinformatics 2017;85(3):497–512.
[233] Dapkūnas J, Venclovas Č. Template-based modeling of protein complexes using the ppi3d web server. Protein Structure Prediction, Springer 2020:139–55.
[234] Lee H, Baek M, Lee GR, Park S, Seok C. Template-based modeling and ab initio refinement of protein oligomer structures using galaxy in capri round 30, Proteins: Structure. Function, and Bioinformatics 2017;85(3):399–407.
[235] Kundrotas PJ, Lensink MF, Alexov E. Homology-based modeling of 3d structures of protein–protein complexes using alignments of modified sequence profiles. International journal of biological macromolecules 2008;43(2):198–208.
[236] Ogmen U, Keskin O, Aytuna AS, Nussinov R, Gursoy A. Prism: protein interactions by structural matching. Nucleic acids research 2005;33(suppl_2):W331–6.
[237] You Z-H, Chan KC, 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.
[238] Huang Y-A, You Z-H, Gao X, Wong L, Wang L. Using weighted sparse representation model combined with discrete cosine transformation to predict protein-protein interactions from protein sequence. BioMed research international 2015.
[239] You Z-H, Yu J-Z, Zhu L, Li S, Wen Z-K. A mapreduce based parallel svm for large-scale predicting protein–protein interactions. Neurocomputing 2014;145:37–43.
[240] Northey TC, Barešić A, Martin AC. Intpred: a structure-based predictor of protein–protein interaction sites. Bioinformatics 2018;34(2):223–9.
[241] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, arXiv preprint arXiv:1609.02907.
[242] Madi K, Paquet E, Kheddouci H. New graph distance for deformable 3d objects recognition based on triangle-stars decomposition. Pattern Recognition 2019;90:297–307.
[243] Chen J, Zheng S, Zhao H, Yang Y. Structure-aware protein solubility prediction from sequence through graph convolutional network and predicted contact map. Journal of cheminformatics 2021;13(1):1–10.
[244] Rao J, Zhou X, Lu Y, Zhao H, Yang Y. Imputing single-cell rna-seq data by combining graph convolution and autoencoder neural networks. Iscience 2021;24(5):102393.
[245] Y. Song, S. Zheng, Z. Niu, Z.-H. Fu, Y. Lu, Y. Yang, Communicative representation learning on attributed molecular graphs., in: IJCAI, Vol. 2020, 2020, pp. 2831–2838.
[246] Liu L, Zhu X, Ma Y, Piao H, Yang Y, Hao X, Fu Y, Wang L, Peng J. Combining sequence and network information to enhance protein–protein interaction prediction. BMC bioinformatics 2020;21(16):1–13.
[247] A. Fout, J. Byrd, B. Shariat, A. Ben-Hur, Protein interface prediction using graph convolutional networks, Advances in neural information processing systems 30.
[248] M. Baranwal, A. Magner, J. Saldinger, E.S. Turali-Emre, S. Kozarekar, P. Elvati, J. S. VanEpps, N.A. Kotov, A. Violi, A.O. Hero, Struct2graph: A graph attention network for structure based predictions of protein-protein interactions, bioRxiv.
[249] Kufareva I, Budagyan L, Raush E, Totrov M, Abagyan R. Pier: protein interface recognition for structural proteomics, Proteins: Structure. Function, and Bioinformatics 2007;67(2):400–17.
[250] Liang S, Zhang C, Liu S, Zhou Y. Protein binding site prediction using an empirical scoring function. Nucleic acids research 2006;34(13):3698–707.
[251] Qin S, Zhou H-X. meta-ppisp: a meta web server for protein-protein interaction site prediction. Bioinformatics 2007;23(24):3386–7.
[252] Porollo A, Meller J. Prediction-based fingerprints of protein–protein interactions, Proteins: Structure. Function, and Bioinformatics 2007;66(3):630–45.
[253] 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(12):e28766.
[254] 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. Proceedings of the National Academy of Sciences 2011;108(49):E1293–301.
[255] Balakrishnan S, Kamisetty H, Carbonell JG, Lee S-I, Langmead CJ. Learning generative models for protein fold families, Proteins: Structure. Function, and Bioinformatics 2011;79(4):1061–78.
[256] Vreven T, Moal IH, Vangone A, Pierce BG, Kastritis PL, Torchala M, Chaleil R, Jiménez-García B, Bates PA, Fernandez-Recio J, et al. Updates to the integrated protein–protein interaction benchmarks: docking benchmark version 5 and affinity benchmark version 2. Journal of molecular biology 2015;427(19):3031–41.
[257] Krissinel E, Henrick K. Inference of macromolecular assemblies from crystalline state. Journal of molecular biology 2007;372(3):774–97.
[258] Wang G, Dunbrack Jr RL. Pisces: a protein sequence culling server. Bioinformatics 2003;19(12):1589–91.

[259] Zhu H, Domingues FS, Sommer I, Lengauer T. Noxclass: prediction of protein-protein interaction types. BMC bioinformatics 2006;7(1):1–15.
[260] Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. Cath–a hierarchic classification of protein domain structures. Structure 1997;5 (8):1093–109.
[261] Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, et al. String v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic acids research 2019;47(D1):D607–13.
[262] Trabuco LG, Betts MJ, Russell RB. Negative protein–protein interaction datasets derived from large-scale two-hybrid experiments. Methods 2012;58(4):343–8.
[263] A. Bateman, Uniprot: a universal hub of protein knowledge, in: Protein Science, Vol. 28, WILEY 111 RIVER ST, HOBOKEN 07030–5774, NJ USA, 2019, pp. 32–32.
[264] 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. Journal of proteome research 2010;9(10):4992–5001.
[265] Cocco S, Feinauer C, Figliuzzi M, Monasson R, Weigt M. Inverse statistical physics of protein sequences: a key issues review. Reports on Progress in Physics 2018;81(3):032601.
[266] Lehmann M, Kostrewa D, Wyss M, Brugger R, D'Arcy A, Pasamontes L, van Loon AP. From dna sequence to improved functionality: using protein sequence comparisons to rapidly design a thermostable consensus phytase. Protein engineering 2000;13(1):49–57.
[267] Socolich M, Lockless SW, Russ WP, Lee H, Gardner KH, Ranganathan R. Evolutionary information for specifying a protein fold. Nature 2005;437 (7058):512–8.
[268] Porebski BT, Buckle AM. Consensus protein design. Protein Engineering, Design and Selection 2016;29(7):245–51.
[269] Sievers F, Higgins DG. Clustal omega for making accurate alignments of many protein sequences. Protein Science 2018;27(1):135–45.
[270] Hugenholtz P, Tyson GW. Metagenomics. Nature 2008;455(7212):481–3.
[271] Cumberworth A, Lamour G, Babu MM, Gsponer J. Promiscuity as a functional trait: intrinsically disordered regions as central players of interactomes. Biochemical Journal 2013;454(3):361–9.
[272] Bock JR, Gough DA. Predicting protein–protein interactions from primary structure. Bioinformatics 2001;17(5):455–60.
[273] Chou K-C, Maggiora GM. Domain structural class prediction. Protein Engineering 1998;11(7):523–38.
[274] Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Marchler GH, Song JS, et al. Cdd/sparcle: the conserved domain database in 2020. Nucleic acids research 2020;48(D1):D265–8.
[275] X. Li, L. Yang, X. Zhang, X. Jiao, Prediction of protein-protein interactions based on domain, Computational and mathematical methods in medicine 2019.
[276] Wojcik J, Schächter V. Protein-protein interaction map inference using interacting domain profile pairs. Bioinformatics 2001;17(suppl_1):S296–305.
[277] Wan KK, Park J, Suh JK. Large scale statistical prediction of protein-protein interaction by potentially interacting domain (pid) pair. Genome Informatics 2002;13:42–50.
[278] Kamada M, Sakuma Y, Hayashida M, Akutsu T. Prediction of protein-protein interaction strength using domain features with supervised regression. The Scientific World Journal 2014.
[279] Singhal M, Resat H. A domain-based approach to predict protein-protein interactions. Bmc Bioinformatics 2007;8(1):1–19.
[280] Lee S-A, Chan C-H, Tsai C-H, Lai J-M, Wang F-S, Kao C-Y, Huang C-YF. Ortholog-based protein-protein interaction prediction and its application to inter-species interactions. BMC bioinformatics 2008;9(12):1–9.
[281] Koonin EV. Orthologs, paralogs, and evolutionary genomics. Annu. Rev. Genet. 2005;39:309–38.
[282] Winter C, Henschel A, Kim WK, Schroeder M. Scoppi: a structural classification of protein–protein interfaces. Nucleic acids research 2006;34 (suppl_1):D310–4.
[283] Hosur R, Xu J, Bienkowska J, Berger B. iwrap: an interface threading approach with application to prediction of cancer-related protein–protein interactions. Journal of molecular biology 2011;405(5):1295–310.
[284] Valente GT, Acencio ML, Martins C, Lemke N. The development of a universal in silico predictor of protein-protein interactions. PloS one 2013;8(5):e65587.
[285] You Z-H, Li J, Gao X, He Z, Zhu L, Lei Y-K, Ji Z. Detecting protein-protein interactions with a novel matrix-based protein sequence representation and support vector machines. BioMed research international 2015.
[286] Shen J, Zhang J, Luo X, Zhu W, Yu K, Chen K, Li Y, Jiang H. Predicting protein–protein interactions based only on sequences information. Proceedings of the National Academy of Sciences 2007;104(11):4337–41.
[287] Zhang S-W, Hao L-Y, Zhang T-H. Prediction of protein–protein interaction with pairwise kernel support vector machine. International journal of molecular sciences 2014;15:3220–33.
[288] 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 research 2008;36(9):3025–30.
[289] Ma W, Cao Y, Bao W, Yang B, Chen Y. Act-svm: Prediction of protein-protein interactions based on support vector basis model. Scientific Programming 2020.
[290] T. Bepler, B. Berger, Learning protein sequence embeddings using information from structure, arXiv preprint arXiv:1902.08661.
[291] Ma W, Bao W, Cao Y, Yang B, Chen Y. Prediction of protein-protein interaction based on deep learning feature representation and random forest. International Conference on Intelligent Computing, Springer 2021:654–62.
[292] Du X, Sun S, Hu C, Yao Y, Yan Y, Zhang Y. Deepppi: boosting prediction of protein–protein interactions with deep neural networks. Journal of chemical information and modeling 2017;57(6):1499–510.
[293] Wold S, Jonsson J, Sjörström M, Sandberg M, Rännar S. Dna and peptide sequences and chemical processes multivariately modelled by principal component analysis and partial least-squares projections to latent structures. Analytica Chimica Acta 1993;277(2):239–53.
[294] Davies MN, Secker A, Freitas AA, Clark E, Timmis J, Flower DR. Optimizing amino acid groupings for gpcr classification. Bioinformatics 2008;24 (18):1980–6.
[295] Tong JC, Tammi MT. Prediction of protein allergenicity using local description of amino acid sequence. Frontiers in Bioscience 2008;13(16):6072–8.
[296] Su X-R, You Z-H, Chen Z-H, Yi H-C, Guo Z-H. Protein-protein interaction prediction by integrating sequence information and heterogeneous network representation. In: International Conference on Intelligent Computing. Springer; 2021. p. 617–26.
[297] Manekar SC, Sathe SR. A benchmark study of k-mer counting methods for high-throughput sequencing. GigaScience 2018;7(12):giy125.
[298] Hashemifar S, Neyshabur B, Khan AA, Xu J. Predicting protein–protein interactions through sequence-based deep learning. Bioinformatics 2018;34: i802–10.
[299] Chen M, Ju CJ-T, Zhou G, Chen X, Zhang T, Chang K-W, Zaniolo C, Wang W. Multifaceted protein–protein interaction prediction based on siamese residual rcnn. Bioinformatics 2019;35(14):i305–14.
[300] 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):1–16.
[301] Xu W, Gao Y, Wang Y, Guan J. Protein–protein interaction prediction based on ordinal regression and recurrent convolutional neural networks. BMC bioinformatics 2021;22(6):1–21.
[302] Hu X, Feng C, Zhou Y, Harrison A, Chen M. Deeptrio: a ternary prediction system for protein–protein interaction using mask multiple parallel convolutional neural networks. Bioinformatics 2022;38(3):694–702.
[303] Xie Y, Zhang T. A fault diagnosis approach using svm with data dimension reduction by pca and lda method. In: 2015 Chinese Automation Congress (CAC). IEEE; 2015. p. 869–74.
[304] Saha I, Zubek J, Klingström T, Forsberg S, Wikander J, Kierczak M, Maulik U, Plewczynski D. Ensemble learning prediction of protein–protein interactions using proteins functional annotations. Molecular BioSystems 2014;10 (4):820–30.
[305] X. Wang, J. Xu, W. Shi, J. Liu, Ogru: An optimized gated recurrent unit neural network, in: Journal of Physics: Conference Series, Vol. 1325, IOP Publishing, 2019, p. 012089.
[306] T.N. Kipf, M. Welling, Variational graph auto-encoders, arXiv preprint arXiv:1611.07308.
[307] Cervantes J, Garcia-Lamont F, Rodríguez-Mazahua L, Lopez A. A comprehensive survey on support vector machine classification: Applications, challenges and trends. Neurocomputing 2020;408:189–215.
[308] Yang L, Xia J-F, Gui J. Prediction of protein-protein interactions from protein sequence using local descriptors. Protein and Peptide Letters 2010;17 (9):1085–90.
[309] Wong L, You Z-H, Li S, Huang Y-A, Liu G. Detection of protein-protein interactions from amino acid sequences using a rotation forest model with a novel pr-lpq descriptor. In: International Conference on Intelligent Computing. Springer; 2015. p. 713–20.
[310] Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. p. 785–94.
[311] Chang C-C, Lin C-J. Libsvm: a library for support vector machines. ACM transactions on intelligent systems and technology (TIST) 2011;2 (3):1–27.
[312] Yao Y, Du X, Diao Y, Zhu H. An integration of deep learning with feature embedding for protein–protein interaction prediction. PeerJ 2019;7:e7126.
[313] Deane CM, Salwinski Ł, Xenarios I, Eisenberg D. Protein interactions: two methods for assessment of the reliability of high throughput observations. Molecular & Cellular Proteomics 2002;1(5):349–56.
[314] Martin S, Roe D, Faulon J-L. Predicting protein–protein interactions using signature products. Bioinformatics 2005;21(2):218–26.
[315] Zhou YZ, Gao Y, Zheng YY. Prediction of protein-protein interactions using local description of amino acid sequence. In: Advances in computer science and education applications. Springer; 2011. p. 254–62.
[316] Hamp T, Rost B. Evolutionary profiles improve protein–protein interaction prediction from sequence. Bioinformatics 2015;31(12):1945–50.
[317] 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.
[318] Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. The database of interacting proteins: 2004 update. Nucleic acids research 2004;32 (suppl_1):D449–51.
[319] D. Szklarczyk, J.H. Morris, H. Cook, M. Kuhn, S. Wyder, M. Simonovic, A. Santos, N.T. Doncheva, A. Roth, P. Bork, et al., The string database in 2017:

quality-controlled protein–protein association networks, made broadly accessible, Nucleic acids research (2016) gkw937.
[320] Moal IH, Fernández-Recio J. Skempi: a structural kinetic and energetic database of mutant protein interactions and its use in empirical models. Bioinformatics 2012;28(20):2600–7.
[321] Kong M, Zhang Y, Xu D, Chen W, Dehmer M. Fctp-wsrc: protein–protein interactions prediction via weighted sparse representation based classification. Frontiers in genetics 2020;11:18.
[322] Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 2006;22(13):1658–9.
[323] Fu L, Niu B, Zhu Z, Wu S, Li W. Cd-hit: accelerated for clustering the next-generation sequencing data. Bioinformatics 2012;28(23):3150–2.
[324] F. Richoux, C. Servantie, C. Borès, S. Téletchéa, Comparing two deep learning sequence-based models for protein-protein interaction prediction, arXiv preprint arXiv:1901.06268.
[325] Li W, Jaroszewski L, Godzik A. Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics 2001;17(3):282–3.
[326] Oughtred R, Stark C, Breitkreutz B-J, Rust J, Boucher L, Chang C, Kolas N, O'Donnell L, Leung G, McAdam R, et al. The biogrid interaction database: 2019 update. Nucleic acids research 2019;47(D1):D529–41.
[327] U. Consortium. Uniprot: a worldwide hub of protein knowledge. Nucleic acids research 2019;47(D1):D506–15.
[328] Popova M, Isayev O, Tropsha A. Deep reinforcement learning for de novo drug design, Science advances, 4, 2018. p. eaap7885..
[329] Segler MH, Kogej T, Tyrchan C, Waller MP. Generating focused molecule libraries for drug discovery with recurrent neural networks. ACS central science 2018;4(1):120–31.
[330] Gupta A, Müller AT, Huisman BJ, Fuchs JA, Schneider P, Schneider G. Generative recurrent networks for de novo drug design. Molecular informatics 2018;37(1–2):1700111.
[331] D. Duvenaud, D. Maclaurin, J. Aguilera-Iparraguirre, R. Gómez-Bombarelli, T. Hirzel, A. Aspuru-Guzik, R.P. Adams, Convolutional networks on graphs for learning molecular fingerprints, arXiv preprint arXiv:1509.09292.
[332] Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, Leswing K, Pande V. Moleculenet: a benchmark for molecular machine learning. Chemical science 2018;9(2):513–30.
[333] Gómez-Bombarelli R, Wei JN, Duvenaud D, Hernández-Lobato JM, Sánchez-Lengeling B, Sheberla D, Aguilera-Iparraguirre J, Hirzel TD, Adams RP, Aspuru-Guzik A. Automatic chemical design using a data-driven continuous representation of molecules. ACS central science 2018;4(2):268–76.
[334] Kusner MJ, Paige B, Hernández-Lobato JM. Grammar variational autoencoder. International Conference on Machine Learning, PMLR 2017:1945–54.
[335] N. Killoran, L.J. Lee, A. Delong, D. Duvenaud, B.J. Frey, Generating and designing dna with deep generative models, arXiv preprint arXiv:1712.06148.
[336] Kamisetty H, Ovchinnikov S, Baker D. Assessing the utility of coevolution-based residue–residue contact predictions in a sequence-and structure-rich era. Proceedings of the National Academy of Sciences 2013;110(39):15674–9.
[337] Hopf TA, Schärfe CP, Rodrigues JP, Green AG, Kohlbacher O, Sander C, Bonvin AM, Marks DS. Sequence co-evolution gives 3d contacts and structures of protein complexes. Elife 2014;3:e03430.
[338] Keefe AD, Szostak JW. Functional proteins from a random-sequence library. Nature 2001;410(6829):715–8.
[339] Fisher MA, McKinley KL, Bradley LH, Viola SR, Hecht MH. De novo designed proteins from a library of artificial sequences function in escherichia coli and enable cell growth. PloS one 2011;6(1):e15364.
[340] Murphy GS, Greisman JB, Hecht MH. De novo proteins with life-sustaining functions are structurally dynamic. Journal of molecular biology 2016;428 (2):399–411.
[341] Wan C, Jones DT. Protein function prediction is improved by creating synthetic feature samples with generative adversarial networks. Nature Machine Intelligence 2020;2(9):540–50.
[342] Hawkins-Hooker A, Depardieu F, Baur S, Couairon G, Chen A, Bikard D. Generating functional protein variants with variational autoencoders. PLoS Computational Biology 2021;17(2):e1008736.
[343] Olivecrona M, Blaschke T, Engkvist O, Chen H. Molecular de-novo design through deep reinforcement learning. Journal of cheminformatics 2017;9 (1):1–14.
[344] Yang KK, Wu Z, Arnold FH. Machine-learning-guided directed evolution for protein engineering. Nature methods 2019;16(8):687–94.
[345] O'Connell J, Li Z, Hanson J, Heffernan R, Lyons J, Paliwal K, Dehzangi A, Yang Y, Zhou Y. Spin2: Predicting sequence profiles from protein structures using deep neural networks, Proteins: Structure. Function, and Bioinformatics 2018;86(6):629–33.
[346] Chen S, Sun Z, Lin L, Liu Z, Liu X, Chong Y, Lu Y, Zhao H, Yang Y. To improve protein sequence profile prediction through image captioning on pairwise residue distance map. Journal of chemical information and modeling 2019;60 (1):391–9.
[347] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, arXiv preprint arXiv:1511.06434.
[348] Riesselman AJ, Ingraham JB, Marks DS. Deep generative models of genetic variation capture the effects of mutations. Nature methods 2018;15 (10):816–22.
[349] Tubiana J, Cocco S, Monasson R. Learning protein constitutive motifs from sequence data. Elife 2019;8:e39397.
[350] Riesselman A, Shin J-E, Kollasch A, McMahon C, Simon E, Sander C, Manglik A, Kruse A, Marks D. Accelerating protein design using autoregressive generative models. BioRxiv 2019:757252.
[351] S. Sinai, E. Kelsic, G.M. Church, M.A. Nowak, Variational auto-encoding of protein sequences, arXiv preprint arXiv:1712.03346.
[352] Repecka D, Jauniskis V, Karpus L, Rembeza E, Rokaitis I, Zrimec J, Poviloniene S, Laurynenas A, Viknander S, Abuajwa W, et al. Expanding functional protein sequence spaces using generative adversarial networks. Nature Machine Intelligence 2021;3(4):324–33.
[353] S. Bai, J.Z. Kolter, V. Koltun, An empirical evaluation of generic convolutional and recurrent networks for sequence modeling, arXiv preprint arXiv:1803.01271.
[354] H. Zhang, I. Goodfellow, D. Metaxas, A. Odena, Self-attention generative adversarial networks, in: International conference on machine learning, PMLR, 2019, pp. 7354–7363.
[355] W. Ding, K. Nakai, H. Gong, Protein design via deep learning, Briefings in Bioinformatics.
[356] Greener JG, Moffat L, Jones DT. Design of metalloproteins and novel protein folds using variational autoencoders. Scientific reports 2018;8(1):1–12.
[357] W. Boomsma, J. Frellsen, Spherical convolutions and their application in molecular modelling., in: NIPS, Vol. 2, 2017, p. 6.
[358] M. Weiler, M. Geiger, M. Welling, W. Boomsma, T. Cohen, 3d steerable cnns: Learning rotationally equivariant features in volumetric data, arXiv preprint arXiv:1807.02547.
[359] Wang S, Sun S, Li Z, Zhang R, Xu J. Accurate de novo prediction of protein contact map by ultra-deep learning model. PLoS Computational Biology 2017;13(1):e1005324.
[360] Spencer M, Eickholt J, Cheng J. A deep learning network approach to ab initio protein secondary structure prediction. IEEE/ACM Transactions on Computational Biology and Bioinformatics 2014;12(1):103–12.
[361] Wang S, Peng J, Ma J, Xu J. Protein secondary structure prediction using deep convolutional neural fields. Scientific reports 2016;6(1):1–11.
[362] P. Das, K. Wadhawan, O. Chang, T. Sercu, C.D. Santos, M. Riemer, V. Chenthamarakshan, I. Padhi, A. Mojsilovic, Pepcvae: Semi-supervised targeted design of antimicrobial peptide sequences, arXiv preprint arXiv:1810.07743.
[363] D. Repecka, V. Jauniskis, L. Karpus, E. Rembeza, J. Zrimec, S. Poviloniene, I. Rokaitis, A. Laurynenas, W. Abuajwa, O. Savolainen, et al., Expanding functional protein sequence space using generative adversarial networks, bioRxiv (2019) 789719.
[364] A. Madani, B. McCann, N. Naik, N.S. Keskar, N. Anand, R.R. Eguchi, P.-S. Huang, R. Socher, Progen: Language modeling for protein generation, arXiv preprint arXiv:2004.03497.
[365] Strokach A, Becerra D, Corbi-Verge C, Perez-Riba A, Kim PM. Fast and flexible protein design using deep graph neural networks. Cell Systems 2020;11 (4):402–11.
[366] Muller AT, Hiss JA, Schneider G. Recurrent neural network model for constructive peptide design. Journal of chemical information and modeling 2018;58(2):472–9.
[367] R. Lim, Methods for accelerating machine learning in high performance computing, University of Oregon–Area-2019-01.
[368] Pan J, Li L-P, Yu C-Q, You Z-H, Guan Y-J, Ren Z-H. Sequence-based prediction of plant protein-protein interactions by combining discrete sine transformation with rotation forest. Evolutionary Bioinformatics 2021;17. 11769343211050067.
[369] Jia J, Li X, Qiu W, Xiao X, Chou K-C. ippi-pseaac (cgr): Identify protein-protein interactions by incorporating chaos game representation into pseaac. Journal of theoretical biology 2019;460:195–203.
[370] Chang C-K, Lin S-M, Satange R, Lin S-C, Sun S-C, Wu H-Y, Kehn-Hall K, Hou M-H. Targeting protein-protein interaction interfaces in covid-19 drug discovery, Computational and Structural. Biotechnology Journal 2021;19:2246–55.
[371] S. Ferrari, F. Pellati, M. Costi, Disruption of protein-protein interfaces (2013).
[372] Rual J-F, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, et al. Towards a proteome-scale map of the human protein–protein interaction network. Nature 2005;437(7062):1173–8.
[373] Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, et al. A human protein-protein interaction network: a resource for annotating the proteome. Cell 2005;122 (6):957–68.
[374] F. Browne, H. Zheng, H. Wang, F. Azuaje, From experimental approaches to computational techniques: a review on the prediction of protein-protein interactions., Advances in Artificial Intelligence (16877470).