Review

Machine Learning and Deep Learning Frameworks for Human–Virus Protein–Protein Interaction Prediction: Emerging Architectures, Methods, Benchmarks, and Challenges

Subhadeep Basu , Dipanwita Adhikary , Kuntal Ghosh , Swarup Chattopadhyay , Shramana Deb , Ritwick Mondal , Jayanta Roy , Anjan Chowdhury * , Julián Benito-León *

Posted Date: 29 May 2026

doi: 10.20944/preprints202605.2116.v1

Keywords: protein protein interaction (PPI); network prediction; biological databases; computational models; machine learning approaches

 doi:10.20944/preprints202605.2116.v1

Review

Machine Learning and Deep Learning Frameworks for Human–Virus Protein–Protein Interaction Prediction: Emerging Architectures, Methods, Benchmarks, and Challenges

Subhadeep Basu [1], Dipanwita Adhikary [2], Kuntal Ghosh [3], Swarup Chattopadhyay [4], Shramana Deb [5,6], Ritwick Mondal [5,6], Jayanta Roy [5,6], Anjan Chowdhury [7,*] and Julián Benito-León [8,9,10,11,*]

[1] Department of Biotechnology, Amity University, Noida, Delhi, India
[2] Department of Biotechnology and Biochemical Engineering, Indian Institute of Technology Kharagpur, India
[3] Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India
[4] School of Computer Science and Engineering, XIM University, Bhubaneswar, India
[5] Centre for Neurovascular Research, Manipal Group of Hospitals, Kolkata, India
[6] Department of Neurology, Manipal Group of Hospitals, Kolkata, India
[7] Department of Computer Science and Engineering, Indian Institute of Technology Dhanbad, India
[8] Department of Neurology, 12 de Octubre University Hospital, Madrid, Spain
[9] Instituto de Investigación Sanitaria Hospital 12 de Octubre, Madrid, Spain
[10] Centro de Investigación Biomédica en Red Sobre Enfermedades Neurodegenerativas (CIBERNED), Madrid, Spain
[11] Department of Medicine, Complutense University, Madrid, Spain
* Correspondence: anjanchowdhury@iitism.ac.in (A.C.); jbenitol67@gmail.com (J.B.-L.)

Abstract

The outbreak of coronavirus disease 2019 (COVID-19), caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has emerged as one of the most significant global health crises in recent history. Coronaviruses are a diverse group of RNA viruses classified into alpha, beta, gamma, and delta genera, with SARS-CoV-2 belonging to the beta-coronavirus family. The virus exhibits high transmissibility and causes a wide spectrum of clinical manifestations ranging from mild respiratory symptoms to severe complications such as acute respiratory distress syndrome, multi-organ failure, and death, particularly among elderly and immunocompromised individuals. Structurally, SARS-CoV-2 possesses a large single-stranded RNA genome encoding major structural proteins, including spike (S), envelope (E), membrane (M), and nucleocapsid (N) proteins, which play critical roles in host cell recognition and viral infection. Understanding the molecular mechanisms of virus–host interactions, especially protein–protein interactions (PPIs), is essential for uncovering viral pathogenesis and identifying potential therapeutic targets. Traditional experimental techniques for PPI detection, such as yeast two-hybrid and affinity purification methods, are often expensive, labor-intensive, and prone to inaccuracies. Consequently, computational approaches based on machine learning and deep learning have gained significant attention for efficient and scalable PPI prediction. These methods utilize diverse biological information, including protein sequences, structural features, genomic data, gene ontology annotations, and interaction networks, to model complex biological relationships. This survey provides a comprehensive review of computational approaches for PPI prediction, highlighting both machine learning- and deep learning-based techniques, along with their methodological advancements and performance evaluations. Furthermore, the survey discusses major biological databases and data sources commonly employed in PPI studies, offering insights into current challenges and future directions in computational PPI prediction research.

 doi:10.20944/preprints202605.2116.v1

Keywords: protein protein interaction (PPI); network prediction; biological databases; computational models; machine learning approaches

1. Background

The emergence of novel zoonotic pathogens, exemplified by the family Coronaviridae, underscores a perpetual threat to global public health, economic stability, and healthcare infrastructure [1–5]. Coronaviruses, categorized into four genera, with alpha and beta viruses predominantly infecting mammalian hosts, have historically triggered severe epidemics, including the SARS outbreak in 2003 [6,7] and the MERS outbreak in 2012 [8], culminating in the global crisis of SARS-CoV-2. These single-stranded RNA viruses possess exceptionally large genomes of approximately 27–32 kb [9] and encode four primary structural proteins—Spike (S), Envelope (E), Membrane (M), and Nucleocapsid (N)—which dictate viral pathogenesis and severe clinical manifestations ranging from acute respiratory distress to systemic multi-organ failure [5,10]. Host-cell invasion is mediated by the physical interaction between the viral spike glycoprotein's S1 receptor-binding domain (RBD) and specific host-cell surface receptors, such as DPP4 for MERS-CoV and ACE2 for SARS-CoV and SARS-CoV-2 [11–13]. Following cellular entry, the virus extensively exploits host machinery through intricate virus–host protein–protein interactions (PPIs) to facilitate replication, translation, and immune evasion. While traditional high-throughput experimental methodologies like yeast two-hybrid (Y2H) screening, affinity purification, tandem affinity purification (TAP), and protein microarrays are foundational for mapping these interactomes, they are inherently bottlenecked by high experimental costs, prolonged timelines, and elevated false-positive and false-negative rates. To address these critical limitations, computational frameworks have emerged as rapid, scalable, and high-throughput alternatives that leverage diverse biological data modalities, including primary sequences, 3D structural configurations, gene ontology annotations, and genomic networks. These in silico approaches broadly fall into two categories: traditional machine learning paradigms that rely on manual feature engineering and modern deep learning models, which encompass both sequential, non-graph-based architectures and geometry-aware graph neural networks. Given the rapid evolution of viral genomes and the overwhelming number of proposed computational frameworks, there is a critical need for a systematic evaluation of these predictive pipelines. This review addresses this imperative by providing a comprehensive, rigorous analysis of emerging machine learning and deep learning frameworks explicitly engineered for human–virus PPI prediction. The scope of this paper encompasses a detailed taxonomy of biological data sources, an in-depth comparative analysis of state-of-the-art algorithmic architectures, and established performance benchmarks. Furthermore, we evaluate a sequence-based machine learning model to demonstrate practical implementation, ultimately delineating open engineering challenges and future research trajectories in computational virology.

2. Literature Search Strategy

This review was conducted using a structured, semi-systematic literature framework to comprehensively synthesize contemporary computational approaches to predicting human–virus protein–protein interactions (PPIs). Given the rapidly evolving, interdisciplinary intersection of computational interactomics, structural bioinformatics, artificial intelligence, and graph representation learning, a semi-systematic review was chosen over a strictly systematic design to enable a broader conceptual integration of emerging algorithmic paradigms and methodological innovations across heterogeneous sources of literature.

A comprehensive electronic literature search was conducted across multiple scientific databases: PubMed/MEDLINE, Scopus, Web of Science Core Collection, IEEE Xplore Digital Library, and Google Scholar. The search targeted studies published between January 2005 and March 2025, thereby encompassing both foundational machine learning-based PPI prediction frameworks and

 doi:10.20944/preprints202605.2116.v1

contemporary transformer-based or graph neural network architectures developed during the modern deep learning era. The final database search was completed in November 2025, and only peer-reviewed articles published in English were considered for inclusion.

PubMed/MEDLINE served as the principal biomedical retrieval platform due to its integration with the National Library of Medicine (NLM) Medical Subject Headings (MeSH) indexing system. To maximize retrieval precision and semantic coverage, the search strategy combined controlled MeSH vocabulary terms with free-text keyword expansion, Boolean retrieval operators, phrase searching, truncation, and semantic synonym mapping. MeSH descriptors were selected according to their direct relevance to protein interaction biology, host–pathogen systems, structural bioinformatics, artificial intelligence, graph representation learning, and computational biology.

The principal MeSH descriptors incorporated into the search framework included "Protein Interaction Mapping," "Protein Binding," "Host-Pathogen Interactions," "Viruses," "Machine Learning," "Deep Learning," "Artificial Intelligence," "Algorithms," "Neural Networks, Computer," "Sequence Analysis, Protein," "Protein Structure, Tertiary," "Computational Biology," "Data Mining," "Graph Theory," and "Systems Biology." These controlled vocabulary descriptors were supplemented with free-text terms including "protein–protein interaction," "human–virus interactome," "host–virus interaction," "graph neural network," "geometry-aware learning," "protein language model," "transformer architecture," "deep representation learning," "graph attention network," and "structure-aware PPI prediction" to improve sensitivity for recently emerging computational methodologies that may not yet be fully indexed within conventional MeSH hierarchies. Boolean retrieval logic was implemented using a concept-block search architecture in which semantically related terms were linked with the OR operator, whereas biologically distinct conceptual domains were linked with the AND operator.

2.1. Eligibility Criteria and Study Selection

The screening pipeline strictly enforced clear boundaries to ensure both methodological relevance and data transparency. Studies were considered eligible based on the following explicit criteria: The study must describe computational methodologies for predicting protein–protein interactions in human–virus or host–pathogen systems. The predictive framework must incorporate machine learning, deep learning, graph representation learning, probabilistic modeling, transformer architectures, or hybrid computational systems. The manuscript must report benchmark transparency by detailing the explicit dataset construction methodology, negative sampling strategies, external validation procedures, and evaluation metrics. The reported performance evaluation must prioritize robust metrics such as precision–recall area under the curve (PR-AUC), Matthews correlation coefficient (MCC), F1-score, or independent test performance to guard against the artificial inflation of accuracy that is typical of highly imbalanced PPI datasets.

Conversely, records were systematically excluded from the evidence synthesis based on the following rules: Studies exclusively focused on wet-laboratory experimental interaction detection without any computational modeling component were omitted. Editorials, commentary pieces, non-peer-reviewed reports, and duplicate publications were excluded. Conference abstracts lacking methodological transparency, architectural detail, or an accessible full text were removed from the selection pool. Retrieved records underwent sequential title screening, abstract evaluation, and full-text assessment. Duplicate records identified across multiple databases were removed using a combination of automated and manual deduplication procedures. During full-text assessment, studies were critically evaluated according to computational novelty, reproducibility, biological applicability, and architectural significance. Additional backward and forward citation tracking was conducted for highly influential publications to identify relevant studies not captured during the primary database retrieval process.

 doi:10.20944/preprints202605.2116.v1

2.2. Data Extraction and Evidence Synthesis

For each eligible study, structured data extraction was performed to systematically collect information regarding biological dataset source, organism or viral system, feature representation modality, computational architecture, graph construction strategy, benchmark design, evaluation metrics, validation protocol, and predictive performance. Particular attention was paid to studies employing graph neural networks, protein language models, transformer-based architectures, multimodal feature integration, explainable artificial intelligence frameworks, and structure-aware learning systems, given their growing importance in modern computational interactomics.

The extracted studies were subsequently categorized into major computational paradigms, including feature-based machine learning approaches, network embedding frameworks, non-geometry-aware deep learning architectures, geometry-aware graph neural networks, temporal and dynamic PPIN models, heterogeneous graph learning systems, probabilistic computational models, and structure-aware residue-level learning frameworks. Comparative synthesis focused on identifying methodological trends, architectural strengths and limitations, benchmark heterogeneity, interpretability challenges, scalability constraints, and emerging research directions in AI-driven protein interaction prediction.

Because of substantial heterogeneity across datasets, benchmark construction procedures, negative sampling strategies, validation protocols, and evaluation metrics, quantitative meta-analysis was not performed. Consequently, the evidence synthesis presented in this review is qualitative and conceptual in nature, with comparative analyses intended to provide methodological interpretation and architectural insight rather than strict cross-study performance ranking. Studies lacking sufficient methodological transparency in dataset construction, negative sampling procedures, or validation strategies were interpreted with caution during comparative synthesis to minimize overestimation of reported predictive performance.

3. Biological Databases and Data Modalities for PPI Prediction (Figure 1, Table 1)

The advent of high-throughput experimental technologies has generated an exponential increase in interactome data since 2005, transforming computational protein–protein interaction (PPI) prediction from early genome-reliant mapping into a big-data discipline. While this data deluge poses distinct computational scalability challenges, it simultaneously provides a rich substrate for deep learning architectures. Modern framework paradigms increasingly favor integrative strategies that synthesize diverse data modalities to decode complex interactome networks. These multi-modal pipelines leverage distinct protein descriptors across five primary data categories: primary sequences, higher-order structures, Gene Ontology (GO) annotations, genomic contexts, and topological network features.

 doi:10.20944/preprints202605.2116.v1

Table 1. Curated Biological Databases and Resources Commonly Utilized in Computational Human–Virus Protein–Protein Interaction Prediction.

Resource Category	Database / Resource	Resource Type	Primary Application in Human–Virus PPI Prediction	Data Modality	Organism Scope	Curation Status	Representative Use in Computational Frameworks	Official URL
Protein Sequence Resources	UniProtKB /Swiss-Prot	Curated protein knowledgebase	Protein annotation, functional characterisation, sequence retrieval	Protein sequence, functional annotation	Multi-species	Expert-curated	Sequence embedding, feature engineering, transformer-based protein modelling	https://www.uniprot.org/
	TrEMBL	Computationally annotated protein repository	Large-scale sequence acquisition for deep learning pipelines	Protein sequence	Multi-species	Automatically annotated	Pretraining and large-scale sequence representation learning	https://www.uniprot.org/
	PIR	Protein information repository	Protein family and functional classification	Protein sequence and annotation	Multi-species	Curated	Evolutionary feature extraction and comparative sequence analysis	https://proteininformationresource.org/
Structural Biology Databases	PDB	Structural repository	Three-dimensional structural modelling of interacting proteins	3D protein structures	Multi-species	Curated experimental structures	Structure-aware and geometry-aware PPI prediction	https://www.rcsb.org/
	SCOP	Structural classification database	Protein fold and domain classification	Structural hierarchy	Multi-species	Curated	Structural similarity learning and fold-aware modelling	http://scop.mrc-lmb.cam.ac.uk/
	CATH	Protein domain architecture resource	Hierarchical structural classification	Structural topology and domains	Multi-species	Curated	Domain-aware graph representation learning	http://www.cathdb.info/
Gene Ontology and Functional Annotation Resources	Gene Ontology (GO)	Controlled biological ontology	Functional feature extraction and semantic similarity analysis	Biological process, molecular function, cellular component annotations	Multi-species	Curated	GO-based feature engineering and functional interaction modelling	http://geneontology.org/
	QuickGO	GO annotation browser	Rapid GO annotation retrieval	Functional annotations	Multi-species	Curated	Functional enrichment and annotation mapping	https://www.ebi.ac.uk/QuickGO/
	DAVID	Functional annotation platform	Functional clustering and pathway enrichment	Gene and protein annotations	Multi-species	Curated	Biological interpretation and pathway analysis	https://david.ncifcrf.gov/
Protein–Protein Interaction Databases	STRING	Integrated interaction network database	Known and predicted interaction retrieval	Physical and functional PPIs	Multi-species	Integrated curated + predicted	Network construction and graph-based learning	https://string-db.org/
	BioGRID	Interaction repository	Experimentally validated molecular interactions	Physical and genetic interactions	Multi-species	Curated	Benchmark dataset generation and validation	https://thebiogrid.org/
	IntAct	Molecular interaction database	Experimentally curated molecular interactions	Protein and molecular interactions	Multi-species	Expert-curated	Gold-standard interaction benchmarking	https://www.ebi.ac.uk/intact/home
	DIP	Database of interacting proteins	Experimentally verified PPIs	Protein interactions	Multi-species	Curated	Classical benchmark dataset construction	https://dip.doe-mbi.ucla.edu/dip/Main.cgi

 doi:10.20944/preprints202605.2116.v1

Abbreviations: CATH, Class, Architecture, Topology, Homologous Superfamily; CC, Cellular Component (Gene Ontology domain); DAVID, Database for Annotation, Visualization and Integrated Discovery; DIP, Database of Interacting Proteins; GO, Gene Ontology; HOS, Higher-Order Structure; HPRD, Human Protein Reference Database; MF, Molecular Function (Gene Ontology domain); MINT, Molecular Interaction Search Tool; MIPS, Munich Information Center for Protein Sequences; PDB, Protein Data Bank; PIR, Protein Information Resource; PPIN, Protein–Protein Interaction Network; SCOP, Structural Classification of Proteins; STRING, Search Tool for the Retrieval of Interacting Genes/Proteins; TrEMBL, Translated EMBL Nucleotide Sequence Data Bank; UniProt, Universal Protein Resource; URL, Uniform Resource Locator.

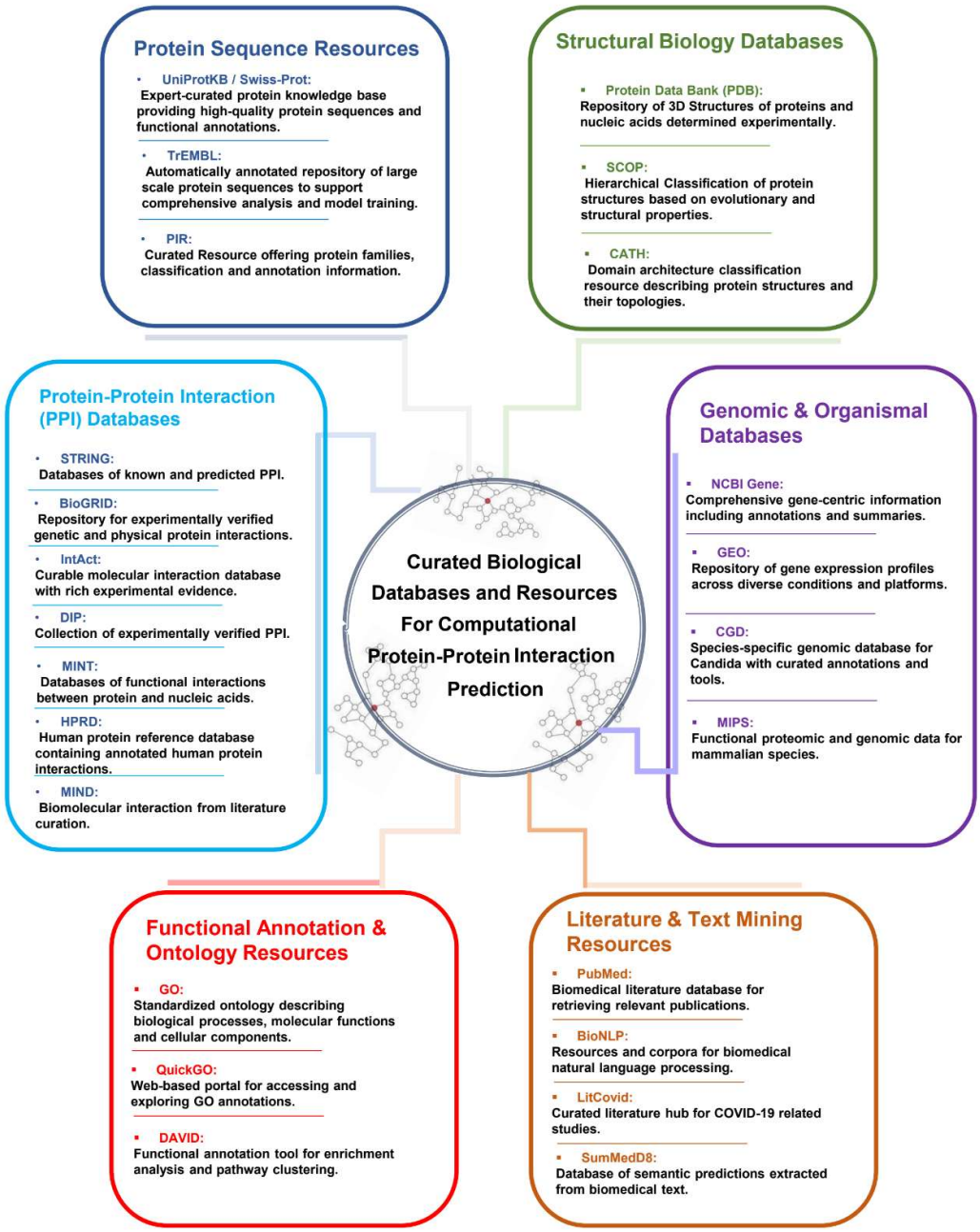

Figure 1. Biological databases and multimodal data resources used in computational protein–protein interaction (PPI) prediction. The illustration summarizes the major categories of biological databases integrated into modern PPI prediction pipelines, including sequence repositories, structural databases, interaction resources, genomic/transcriptomic datasets,

 doi:10.20944/preprints202605.2116.v1

functional annotation systems, and literature/text-mining platforms. The figure further highlights the data harmonization and integration workflow used to generate biologically relevant, high-confidence PPI networks for downstream computational and systems biology applications. Abbreviations: PPI, Protein–Protein Interaction; GO, Gene Ontology; PDB, Protein Data Bank; DAVID; Database for Annotation Visualization and Integrated Discovery, Gene Expression Omnibus; HPIDB, Host–Pathogen Interaction Database; BioGRID, Biological General Repository for Interaction Datasets; MIND, Membrane-Protein Interaction Network Database; MINT, Molecular Interaction Database; IntAct, Interaction Action; HPRD, Human Protein Reference Database; DIP; Database of Interacting Proteins; CGD, Candida Genome Database; MIPS, Munich Information Center for Protein Sequences; CATH; Class Architecture Topology Homologous superfamily; SCOP; Structural Classification of Proteins; PIR; Protein Information Resource.

3.1. Sequence-Based Information

The primary structure of a protein, represented linearly by its amino acid sequence, serves as the baseline data modality for computational PPI prediction due to its massive volume and relative ease of acquisition. Because the primary sequence intrinsically dictates higher-order folding configurations, it serves as a robust standalone proxy for predictive modeling when structural or functional annotations are missing. To capture these interactions mathematically, computational pipelines extract sequence-derived descriptors that quantify pairwise similarities. These feature vectors typically encompass fundamental physicochemical properties—including hydrophobicity and amino acid composition—as well as evolutionary conservation profiles derived from multiple sequence alignments, such as position-specific scoring matrices (PSSMs) [14–16]. Such evolutionary conservation landscapes reveal critical co-evolutionary patterns essential for predicting stable binding interfaces [17]. Standardized repositories hosting these primary sequence datasets include the Protein Information Resource (PIR), UniProt [18], and SWISS-PROT [19].

3.2. Higher-Order Structural Data

Protein higher-order structures (HOS) span hierarchical levels of organization, starting from local secondary structural motifs (α-helices and β-sheets) dictated by backbone hydrogen bonding, extending to three-dimensional tertiary conformations of single chains, and culminating in quaternary multimeric assemblies (e.g., dimers or higher-order oligomers). Guided by Anfinsen's thermodynamic hypothesis, which states that a protein's native three-dimensional structure is determined by its primary amino acid sequence [20], many computational architectures continue to prioritize sequence features. This choice is further reinforced by the stark data asymmetry in public repositories: experimentally resolved 3D structures are far scarcer than the abundance of raw sequence data, fundamentally limiting the scalability of purely structure-based models [21,22]. Nonetheless, when available, structural data provides precise geometric insights into spatial binding interfaces and thermodynamic binding affinities. Curated repositories mapping this structural landscape include the Structural Classification of Proteins (SCOP) [23], the Protein Data Bank (PDB) [24], and CATH [25].

3.3. Gene Ontology (GO) Annotations

The Gene Ontology (GO) consortium offers a highly structured, standardized framework to encapsulate functional, biological, and spatial knowledge of gene products across diverse organisms [26,27]. GO annotations are systematically partitioned into three distinct domains: molecular function (MF), which catalogues biochemical activities such as catalytic or binding mechanisms; biological process (BP), which charts broader multi-step pathways such as signal transduction cascades; and cellular component (CC), which determines precise macromolecular or subcellular localization. In the context of PPI prediction, proteins operating within the same pathway or localized within the same cellular compartment exhibit a significantly higher statistical probability of physical interaction. Computational models capture this functional proximity by computing semantic similarity metrics across GO directed acyclic graphs, markedly improving overall classification accuracy [28,29]. These curated annotations are dynamically maintained and accessed via the GO database [27] and QuickGO [30].

 doi:10.20944/preprints202605.2116.v1

3.4. Genomics-Based Features

Genomic features offer a deep evolutionary perspective on protein co-regulation and physical assembly, rooted firmly in the central dogma of molecular biology where DNA is transcribed into mRNA and subsequently translated into functional polypeptide chains [31]. Driven by advances in whole-genome sequencing, modern computational workflows exploit cross-species genomic conservation to infer physical interactions between translated proteins. Three primary genomics-based features serve as indicators of functional or physical linkage: gene fusion events, gene neighborhood conservation, and phylogenetic profiling [32,33]. Gene fusion occurs when independent genes in a reference organism are expressed as a single, contiguous open reading frame in another species, strongly implying cooperative function. Similarly, conserved gene neighborhoods (such as operons) denote spatial genomic proximity linked to co-expression, while phylogenetic profiles trace the correlated presence or absence of genes across evolutionary lineages to reveal shared functional networks. These evolutionary markers can be retrieved from specialized genomic repositories such as the Candida Genome Database [34] and the Munich Information Centre for Protein Sequences (MIPS) [35].

3.5. Network-Based Topologies (PPIN)

At the systems level, individual physical interactions are synthesized into comprehensive Protein–Protein Interaction Networks (PPINs), which provide a holistic framework for mapping complex cellular pathways, annotating uncharacterized proteins, and identifying critical disease hubs or therapeutic targets [36]. Formally, a PPIN is mathematically modeled as a graph,

$$(G = (V, E)),$$

where the vertex set V represents individual proteins and the edge set E denotes verified physical or functional interactions. This abstraction allows graph-based deep learning and machine learning algorithms to exploit both local topological features (such as node degree and centrality) and global network modularity. Comprehensive interactome repositories archiving these topological networks include STRING [37], BioGRID [38], BIND [39], DIP [40], MINT [41], HPRD [42], and IntAct [43].

As a conceptual synthesis of this entire multimodal landscape, Figure 1 delineates the overarching predictive workflow, charting the seamless trajectory from primary database curation and multi-feature extraction to the execution of an integrative computational model and subsequent downstream PPIN topological analysis for target and mechanism discovery.

4. Machine Learning (ML) Based Approaches for PPI Prediction (Table 2, Figure 2)

Driven by the deluge of high-throughput biological data, machine learning (ML) paradigms have revolutionized protein–protein interaction (PPI) prediction, offering robust computational toolsets for molecular biology and accelerated drug discovery. By leveraging large-scale datasets and advanced algorithms, ML models can accurately predict interactions and decipher complex relational networks. In this review, we categorize ML-based approaches into two primary paradigms: feature-based methods and network embedding-based frameworks.

 doi:10.20944/preprints202605.2116.v1

Table 2. Representative Machine Learning-Based Frameworks for Protein–Protein Interaction Prediction and Human–Virus PPI Modelling.

Study	Interaction Type	Organism/ System	Model Framework	Feature Representation	Negative Sampling	Split Protocol	Validation Strategy	Primary Metrics	Reported Performance	Methodological Remarks
Shen et al. (2007) [14]	General PPIN	General PPIN	SVM	Conjoint triad encoding	Random negatives	Random pair split	5-fold CV	Accuracy	83.90%	Early sequence-based PPI prediction framework
Guo et al. (2008) [49]	General PPIN	S. cerevisiae	SVM	Autocovariance descriptor	Random negatives	Random pair split	Cross-validation	Accuracy	7.40%	Introduced AC-based sequence representation
Yang et al. (2010) [57]	General PPIN	S. cerevisiae	SVM/KNN	Local Descriptor	Random negatives	Random pair split	Cross-validation	Accuracy	86.15%	Local Descriptor + KNN/SVM on S. cerevisiae dataset
Dyer et al. (2011) [50]	Human-virus PPI	Human-virus interactions	Linear SVM	Domain profile + k-mer	Not reported	Not reported	Cross-validation	Recall	65.50%	Integrated host and viral descriptors
Cui et al. (2012) [51]	Human-virus PPI	Human-virus PPIs	RBF-SVM	Conjoint triad encoding	Not reported	Not reported	Cross-validation	Accuracy	87.50%	Applied CT encoding to host-virus interactions
Wang et al. (2012) [62]	General PPIN	Not reported	NMTF	Sparse matrix completion	Not reported	Not reported	Not reported	Precision	38.19%	Nonnegative Matrix Tri-Factorisation (NMTF)-based sparse matrix completion method
You et al. (2013) [58]	General PPIN	General PPIN	Extreme Learning Machine	Multiple sequence descriptors	Random negatives	Random pair split	Cross-validation	Accuracy	87.50%	Ensemble descriptor integration
Dey et al. (2014) [46]	General PPIN	Not reported	SVM	AAC, pseudo AAC, conjoint triads	Not reported	Not reported	Not reported	Accuracy	72.33%	SVM using amino acid composition, pseudo AAC, and conjoint triads
Emamiomeh et al. (2014) [59]	General PPIN	Not reported	Ensemble (SVM, RF, MLP, NB)	Not reported	Not reported	Not reported	Not reported	Accuracy	83.00%	Ensemble stacking of SVM, RF, MLP, and NB
Huang et al. (2015) [111]	General PPIN	Not reported	WSRC	Not reported	Not reported	Not reported	Not reported	Accuracy	96.28%	WSRC classifier
You, Chan & Hu (2015) [112]	General PPIN	Not reported	Random Forest	Multi-scale local features	Not reported	Not reported	Not reported	Accuracy	94.72%	Used a multi-scale local feature representation scheme and the random forest
Xu et al. (2017) [113]	Interaction ranking	General PPIN	EssRank	Random-walk ranking	Not applicable	Not applicable	Confidence evaluation	ROC-AUC	44.00%	Ranking-based interaction prioritization
Wang et al. (2017) [114]	General PPIN	Not reported	Not reported	Legendre moments (PSSM)	Not reported	Not reported	Not reported	Accuracy	96.28%	Used Legendre moments descriptor to extract discriminatory information embedded in PSSM
Wang et al. (2017) [115]	General PPIN	Not reported	PCVM	Zernike moments	Not reported	Not reported	Not reported	Accuracy	94.48%	Used a probabilistic

 doi:10.20944/preprints202605.2116.v1

										classification vector machine model combined with a Zernike moments descriptor
Alguwaizani et al. (2018) [52]	General PPIN	Human PPIN	SVM	AAC + repeat patterns	Random negatives	Independent test set	Independent testing	ROC-AUC	94.00%	Improved sequence discrimination
Song et al. (2018) [116]	General PPIN	Not reported	Ensemble classifier	Random projection	Not reported	Not reported	Not reported	Accuracy	95.64%	An ensemble classifier with random projection using sequence and evolutionary information
Lian et al. (2019) [47]	Human-pathogen PPI	Human-Yersinia pestis	RF + noisy-OR ensemble	Ensemble feature integration	Not reported	Not reported	Cross-validation	ROC-AUC	95.00%	Pathogen-specific interaction prediction
Yang et al. (2020) [55]	General PPIN	General PPIN	Random Forest	Doc2Vec embeddings	Not reported	Random pair split	Cross-validation	ROC-AUC	87.00%	Embedding-based sequence learning
Pei et al. (2021) [60]	General PPIN	Not reported	SymNMF	Matrix factorisation	Not reported	Not reported	Not reported	ROC-AUC	94.00%	SymNMF matrix factorisation model for PPIN link prediction
Debnath et al. (2022) [48]	General PPIN	Not reported	Linear SVM	Primary sequence info	Not reported	Not reported	Not reported	Accuracy	63.00%	Linear SVM using primary protein sequence information
Pan et al. (2022) [56]	Human PPIN	Human PPIN	Rotation Forest	DHT descriptor	Random negatives	Independent test set	Independent testing	ROC-AUC	98.00%	Robust handcrafted descriptor engineering
Goldsmith et al. (2023) [63]	General PPIN	Not reported	Random walk	Not reported	Not reported	Not reported	Not reported	ROC-AUC	92.00%	Continuous-time classical and quantum random walk methods
Ma et al. (2024) [61]	General PPIN	Not reported	VKBNMF	Matrix decomposition	Not reported	Not reported	Not reported	ROC-AUC	89.00%	Kernel Bayesian nonlinear matrix factorisation-based (VKBNMF) Bayesian logistic matrix decomposition model (Avg F1 = 85%)

Abbreviations: AAC, Amino Acid Composition; AC, Autocovariance; AUC, Area Under the Receiver Operating Characteristic Curve; Avg, Average; BiGRU, Bidirectional Gated Recurrent Unit; CKSAAP, Composition of K-Spaced Amino Acid Pairs; CNN, Convolutional Neural Network; CT, Conjoint Triad; DHT, Discrete Hartley Transform; Doc2Vec, Document to Vector; ESM, Evolutionary Scale Modeling; EssRank, Essentiality Ranking; F1, F1-Score (Harmonic mean of precision and recall); HPV, Human Papillomavirus; HCV, Hepatitis C Virus; HIV, Human Immunodeficiency Virus; KNN, k-Nearest Neighbors; LSTM, Long Short-Term Memory; ML, Machine Learning; MLP, Multilayer Perceptron; NA, Not Available / Not Reported in the original study; NB, Naive Bayes; NMTF, Non-negative Matrix Tri-Factorization; PCA, Principal Component Analysis; PCVM, Probabilistic Classification Vector Machine; PPI, Protein–Protein Interaction; PseTC, Pseudo-Transition Composition; PSSM, Position-Specific Scoring Matrix; PTM, Post-Translational Modification; RBF, Radial Basis Function (Kernel); RF, Random Forest; RCNN, Recurrent Convolutional Neural Network; RoF, Rotation Forest; RNN, Recurrent Neural Network; SSAE, Stacked Sparse Autoencoder; SVM, Support Vector Machine; symLMF, Symmetric Logistic Matrix Factorization; VKBNMF, Variational Kernel Bayesian Non-negative Matrix Factorization; Word2Vec, Word to Vector; WSRC, Weighted Sparse Representation-Based Classification.

 doi:10.20944/preprints202605.2116.v1

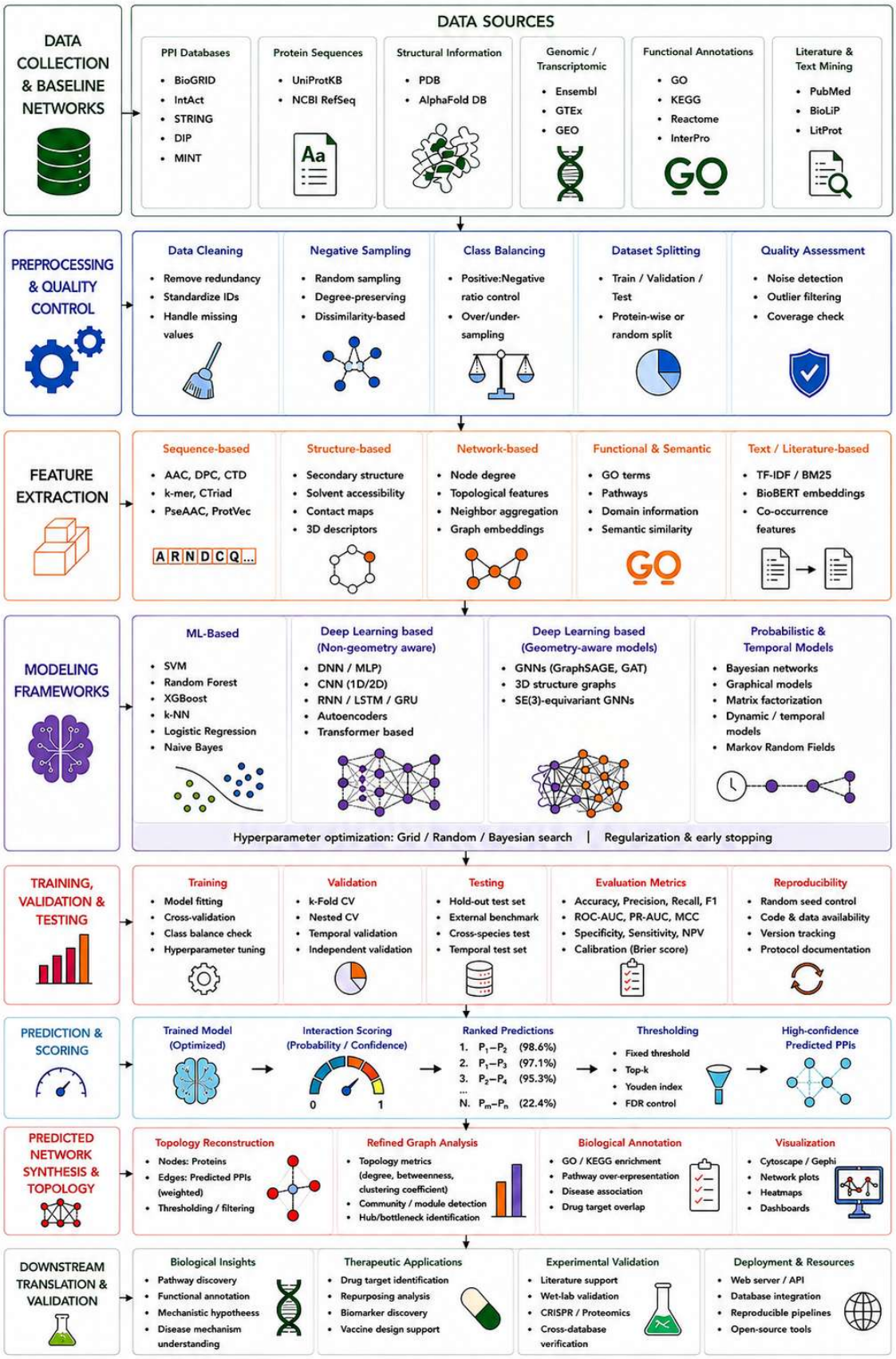

Figure 2. Analytical framework for efficient protein–protein interaction (PPI) network prediction integrating machine learning (ML), deep learning (DL), and Bayesian approaches. The schematic illustrates a unified computational workflow

 doi:10.20944/preprints202605.2116.v1

for human–virus PPI prediction, beginning with the acquisition of multi-source biological data from protein sequence repositories, structural databases, PPI databases, genomic/transcriptomic resources, functional annotation systems, and literature-derived datasets. Following preprocessing, feature extraction modules generate sequence-based descriptors, structural representations, network topology features, Gene Ontology (GO)-based semantic annotations, and text-derived embeddings for downstream predictive modelling. The ML-based framework employs engineered feature vectors and classifiers such as SVMs, RFs, Gradient Boosting, k-NNs, Logistic Regression, and Naïve Bayes. The DL-based framework integrates tensor, sequence, and graph-based representations through MLPs, CNNs, recurrent neural networks (RNNs/LSTMs/GRUs), graph neural networks (GNNs), graph attention networks (GATs), autoencoders, and attention-based architectures. Bayesian modules model interaction uncertainty using Bayesian logistic regression, Bayesian networks, Gaussian processes (GPs), Bayesian matrix factorization, and variational probabilistic learning approaches. Models are subsequently trained and validated using multiple evaluation metrics and cross-validation strategies before generating interaction probability scores. High-confidence predictions are assembled into weighted PPI networks for downstream network construction, topological analysis, hub and module identification, biological interpretation, therapeutic target prioritization, and systems-level interactome analysis. The framework highlights the integration of multimodal biological knowledge with advanced artificial intelligence paradigms for scalable and interpretable PPI network prediction. Abbreviations: PPI, Protein–Protein Interaction; PPIN, Protein–Protein Interaction Network; ML, Machine Learning; DL, Deep Learning; SVM, Support Vector Machine; RF, Random Forest; k-NN, k-Nearest Neighbors; MLP, Multilayer Perceptron; CNN, Convolutional Neural Network; RNN, Recurrent Neural Network; LSTM, Long Short-Term Memory; GRU, Gated Recurrent Unit; GNN, Graph Neural Network; GCN, Graph Convolutional Network; GAT, Graph Attention Network; GO, Gene Ontology; GP, Gaussian Process.

4.1. Feature-Based Approach

Feature-based ML represents the traditional machine learning paradigm, in which hand-crafted biological, physicochemical, or structural descriptors are mathematically extracted from protein data prior to training a downstream classifier or regressor. We outline the key classical feature-based classification techniques below.

4.1.1. SVM-Based PPIN Classification

A Support Vector Machine (SVM) [44] is a supervised learning algorithm used for classification and regression that finds an optimal separating hyperplane with a maximum margin. It efficiently handles non-linear data by using the kernel trick to map inputs into a higher-dimensional feature space. SVM was used to predict human cellular targets of SARS-CoV-2 by combining primary sequence information, amino acid composition, pseudo-amino acid composition, and conjoint triad properties. Similarly, linear SVM classifiers have been optimized strictly using sequence-derived physicochemical property scales extracted from primary structures [45–48]. Early foundational work [14] demonstrated that an SVM trained with a conjoint triad feature description and an S-kernel function could accurately replicate primary interaction topologies across complex crossover networks.

To capture localized sequence patterns, an autocovariance (AC) encoding methodology combined with an SVM was proposed [49] wherein seven critical physicochemical indices—hydrophobicity, hydrophilicity, polarity, polarizability, side-chain volume, solvent-accessible surface area, and net side-chain charge—are numerically scaled across a 30-residue sliding window, achieving an 87.36% classification accuracy on *Saccharomyces cerevisiae*. In host-pathogen contexts, Dyer et al. [50] integrated domain profiles, sequence k-mer composition, and human protein properties within a linear SVM to map human–HIV interactions. Subsequent models utilized Conjoint Triad (CT) encodings paired with radial basis function (RBF) kernels to yield superior predictive accuracies across human–papillomavirus (HPV) and hepatitis C virus (HCV) datasets [51]. This sequence-only predictive capacity was further validated by Alguwaizani et al. [52], who leveraged basic amino acid compositions and repeat patterns within an SVM architecture to outperform contemporary state-of-the-art benchmarks [53,54].

4.1.2. RF-Based PPIN Classification

Random Forest (RF) and its ensemble variants demonstrate high generalization capabilities across cross-species datasets. In recent studies, modern natural language processing (NLP)-driven embeddings such as

 doi:10.20944/preprints202605.2116.v1

doc2vec have been coupled with RF classifiers to achieve improved human–virus PPI prediction performance compared with legacy encoding schemes [55]. Similarly, combining Position-Specific Scoring Matrices (PSSMs) with a 400-dimensional discrete Hartley transform (DHT) descriptor within a Rotation Forest (RoF) classifier yielded superior predictive accuracy compared with standalone RF, SVM, k-NN, and AdaBoost models across human, yeast, and *Oryza sativa* proteomics [56]. To explicitly model bacterial pathogenesis, integrated predictors for human–*Yersinia pestis* PPIs have combined topological network properties (NetTP, NetSS) with multi-scale sequence encodings (CKSAAP, PseTC, AC) via an RF-driven noisy-OR ensemble, achieving a robust area under the receiver operating characteristic curve (AUC) of 0.922 [47].

4.1.3. KNN-Based PPIN Classification

The k-Nearest Neighbors (k-NN) algorithm represents a simpler, non-parametric approach that remains valuable when paired with robust feature extraction techniques. Implementing a unique local descriptor encoding scheme [57] paired with a k-NN classifier achieved an 86.15% accuracy on *S. cerevisiae* data [49] while maintaining competitive performance on independent *E. coli* systems, demonstrating its utility as a supplementary predictive framework.

4.1.4. Ensemble-Based PPIN Classification

To maximize predictive stability and eliminate single-model biases, contemporary investigators increasingly turn to hybrid ensemble classification frameworks. These systems frequently combine heterogeneous feature extraction strategies (including CT, AC, MAC, and localized descriptors) with Principal Component Analysis (PCA) for dimensionality reduction, feeding the condensed representations into ensemble Extreme Learning Machines to yield a highly generalized 87% accuracy across *S. cerevisiae* and *E. coli* [58]. Alternatively, stacking strategies that consolidate SVM, RF, Multi-Layer Perceptrons (MLP), and Naive Bayes (NB) models—utilizing an MLP as a meta-learner [59]—have been engineered to map intricate human–HCV interactomes by testing primary sequence, evolutionary conservation, post-translational modifications (PTMs), and native network properties, significantly outperforming single-classifier benchmarks [51].

4.2. Network Embedding-Based Approach

Let a biological network be represented as a graph $G = (V, E)$, where V represents the set of nodes (proteins), and E signify the set of edges (interactions). The primary objective of network embedding is to learn a mapping function:

$$f: V \rightarrow R^d, \text{ where } d << |V|,$$

This optimization ensures that the structural topologies, local neighborhood contexts, and semantic relationships among nodes in the native graph are preserved within a low-dimensional, dense vector space. In this compressed embedding space, nodes that lie in close topological proximity within the original graph exhibit highly similar vector representations. By mapping high-dimensional discrete graphs into continuous vectors, network embedding enables the seamless application of downstream machine learning algorithms for tasks such as node classification, cluster detection, and link prediction.

4.2.1. Matrix Factorisation-Based

Matrix factorization techniques decompose the interactome adjacency matrix into low-rank latent components to infer missing links. For instance, the symLMF framework [60] optimizes a logistic low-rank matrix factorization model specifically designed for link prediction within undirected protein–protein interaction networks (PPINs). To address host-pathogen systems, the VKBNMF model [61] implements a kernel Bayesian logistic matrix decomposition model equipped with automatic rank determination. This variational Bayesian formulation enables highly accurate predictions of human–virus protein interactions and exhibits robust cross-benchmark generalization. Furthermore, a sparse matrix completion framework utilizing Non-negative Matrix Tri-Factorization (NMTF) [62] has been engineered to predict novel

 doi:10.20944/preprints202605.2116.v1

interactions using positive-only data, successfully integrating heterogeneous biological data streams to outperform contemporary baselines on yeast networks.

4.2.2. Random Walk-Based

Random walk paradigms exploit localized and global stochastic routing across graph topologies to discover implicit node affinities. Recent implementations deploy continuous-time classical and quantum random-walk methodologies to systematically predict missing links in incomplete or noisy PPI networks [63]. Similarly, the Essentiality Ranking (EssRank) method [64] integrates interaction confidence scores, direct and indirect topological relationships, and random walk trajectories. By capturing multi-hop neighborhood features, this method accurately identifies essential proteins, significantly outperforming traditional network centrality measures and localized topology-based algorithms on yeast PPI benchmarks.

5. Deep Learning (DL) Approaches for PPI Prediction (Table 3, Figure 2)

Deep learning has attracted substantial attention across diverse scientific domains due to its ability to perform unsupervised feature learning, leading to prominent efforts to address protein–protein interaction (PPI) prediction. Computational methodologies in this domain are broadly bifurcated into geometry-aware architectures—which exploit graph neural networks (GNNs), graph attention networks (GATs), and graph autoencoders to preserve structural and relational configurations—and non-geometry-aware frameworks.

Table 3. Representative Deep-Learning framework for Protein–Protein Interaction Prediction.

Study	Task Category	Architecture	Model Family	Biological Context	Validation Strategy	Primary Metrics	Reported Performance	Methodological Remarks
Wang et al. (2017) [117]	General PPIN	SSAE+LM +PCVM	Deep feature learning	General PPIN	Cross-validation	Accuracy, MCC	Accuracy: 96.6%; MCC: 93.4%	Stacked sparse autoencoder framework
Du et al. (2017) [16]	General PPIN	DeepPPI	Deep neural network	General PPIN	Cross-validation	ROC-AUC, MCC	AUC: 97.4%; MCC: 85.1%	End-to-end deep representation learning
Ahmed et al. (2018) [68]	General PPIN	MFFN	Feedforward neural network	General PPIN	Cross-validation	ROC-AUC	93.0%	Multi-descriptor neural architecture
Yao et al. (2019) [65]	General PPIN	DeepFE-PPI	Non-geometry-aware DL	Human + S. cerevisiae	5-fold CV	Accuracy, MCC	Accuracy: 98.7%; MCC: 97.4%	Sequence embedding framework
Wang et al. (2019) [70]	RNA–protein interaction	RPIFSE	Feature-selection ensemble	RPI datasets	Cross-validation	Accuracy	89.7–99.0%	Methodologically related RPI framework
Alakus et al. (2019) [74]	General PPIN	LSTM-Prot2Vec	Recurrent neural network	General PPIN	Cross-validation	Accuracy	92.0%	Context-aware embedding model
Xie et al. (2020) [69]	PPI site prediction	CNN-based predictor	Convolutional neural network	Binding-site datasets	Cross-validation	ROC-AUC	91.2%	Site-level interaction prediction
Yang et al. (2020) [82]	Structure-aware PPIN	S-VGAE	Variational graph autoencoder	Signed interaction networks	Cross-validation	ROC-AUC, MCC	AUC: 99.9%; MCC: 98.3%	Geometry-aware graph learning
Tsukiyama et al. (2021) [76]	Human–virus PPI	LSTM-PHV	Sequence DL	Host–virus datasets	Independent testing	ROC-AUC	97.3%	Host–virus interaction prediction
Sledzieski et al. (2021) [80]	Cross-species PPI	D-SCRIPT	Transfer learning	Cross-species datasets	External validation	Recall	96.0%	Transferable cross-species prediction
Baranwal et al. (2022) [96]	Structure-aware PPIN	Struct2Graph	Graph Attention Network	Structural protein datasets	Independent testing	ROC-AUC, MCC	AUC: 99.9%; MCC: 98.8%	Structure-aware graph interaction modeling
Zheng et al. (2023) [66]	Plant PPIN	DeepAraPPI	Deep learning framework	Arabidopsis datasets	Task-specific validation	ROC-AUC	82.5–96.5%	Plant-specific interaction prediction
Kang et al. (2023) [98]	Human PPIN	AFTGAN	Transformer + GAT	SHS27k dataset	Random partition validation	Micro-F1	86.7%	Attention-based graph transformer

 doi:10.20944/preprints202605.2116.v1

Chen et al. (2025) [99]	Dynamic PPIN	DCMF-PPI	Collaborative matrix factorization	SHS27k dataset	Random partition validation	ROC-AUC, MCC	AUC: 99.9%; MCC: 98.4%	Dynamic interaction modeling

Abbreviations: AFTGAN, Attention-based Feature Transformer Graph Attention Network; AUC, Area Under the Receiver Operating Characteristic Curve; CNN, Convolutional Neural Network; DCMF-PPI, Deep Collaborative Matrix Factorization for Protein–Protein Interaction; D-SCRIPT, Deep Sequence-Contact Representations Identifying Partner Targets; DeepAraPPI, Deep Learning for Arabidopsis Protein-Protein Interaction; DeepFE-PPI, Deep Feature Embedding for Protein-Protein Interaction Prediction; GRU, Gated Recurrent Unit; LM, Language Model; LSTM, Long Short-Term Memory; MCC, Matthews Correlation Coefficient; MFFN, Multi-Layer Feed Forward Network; NA, Not Available / Not Reported in the original study; PCVM, Probabilistic Classification Vector Machine; PIPR, Protein–Protein Interaction Prediction Based on Siamese Residual RCNN; RPI, RNA–Protein Interaction; RPIFSE, RNA–Protein Interaction Feature Selection Ensemble; S-VGAE, Signed Variational Graph Autoencoder; SSAE, Stacked Sparse Autoencoder; VM, Vector Machine.

5.1. Non-Geometry-Aware DL Approaches for PPIs Prediction

Non-geometry-aware deep learning approaches treat proteins as independent entities, represented strictly by sequence or feature vectors, and seek to model the underlying interactive functions rather than explicit graph topologies. Within this paradigm, sequence-based feature extraction has been revolutionized by advanced language models (LMs) such as the LSTM-based SeqVec and the BERT-based ProtBertoperate, which process the primary structure of a protein to generate dense, residue-level feature vectors that capture complex biophysical rules. When evaluated against standard benchmarks for human and Saccharomyces cerevisiae, these language models demonstrate outstanding predictive performance, outperforming contemporary legacy baselines.

Similarly, the Res2vec residue representation methodology [65] maps residue-to-residue transitions directly from raw sequences, providing highly optimized inputs for downstream deep learning classification frameworks without requiring any 3D structural parameters. The resulting end-to-end framework, DeepFE-PPI [65], achieved predictive accuracies of 94.78% on the *S. cerevisiae* dataset and 98.71% on human interactomes, while maintaining an average accuracy of 100% across five independent species datasets encompassing *H. sapiens*, *E. coli*, *M. musculus*, *H. pylori*, and *C. elegans*.

5.1.1. Multilayer Perceptron (MLP) Based Models

Historically, the earliest and most straightforward feedforward implementations utilized Multilayer Perceptrons (MLPs), which extract and concatenate static biological descriptors—including amino acid composition (AAC), dipeptide frequencies, physicochemical properties, domain interactions, Position-Specific Scoring Matrices (PSSMs), and Gene Ontology (GO)-based similarity metrics—to learn non-linear interaction boundaries. For instance, DeepAraPPI [66] integrates domain-level embeddings derived from Domain2vec with an MLP classifier to predict Arabidopsis PPIs with 91.3% accuracy.

To address the architectural inability of standard MLPs to capture long-range sequential contexts, the SDNN-PPI framework [67] integrates a self-attention mechanism with deep neural networks; by extracting global and local descriptors via AAC, conjoint triad (CT), and autocovariance (AC) formulations, the attention layers dynamically amplify salient features prior to classification. This network demonstrated strong generalization under 5-fold cross-validation, achieving accuracies of 95.48% on *S. cerevisiae* (core subset) and 98.94% on human intraspecific data, alongside interspecific accuracies of 93.15% on human–*Bacillus anthracis* and 88.33% on human–*Yersinia pestis* datasets.

To map host-pathogen boundaries, another variant employs a Multi-Layer Feed Forward Network (MFFN) neural network [68] that handles black-box classification by fusing macroscopic interactome features (node degree, betweenness centrality, and clustering coefficients) with microscopic sequence similarities and amino acid quadruplets, outperforming traditional SVM metrics on *B. anthracis* and Human Papillomavirus (HPV) data when all multi-scale feature combinations are utilized simultaneously during training.

 doi:10.20944/preprints202605.2116.v1

5.1.2. Convolutional Neural Networks (CNNs) Based Models

Beyond feedforward networks, Convolutional Neural Networks (CNNs) are widely deployed to extract invariant, localized sequence motifs and identify structural properties along the polypeptide chain. Xie et al. [69] used a CNN topology for PPI site prediction, utilizing residue binding propensities to enrich positive training instances and achieving a remarkable area under the curve (AUC) of 0.912.

This convolutional feature extraction strategy has been expanded to heterogeneous networks within the RPIFSE framework [70] for RNA–protein interaction prediction using sequence data, which extracts deep sequence features via a CNN, handles feature perturbation to generate multiple weighted datasets, and applies an Extreme Learning Machine (ELM) classifier within a weighted voting ensemble to yield 5-fold cross-validation accuracies of 91.87%, 89.74%, 97.76%, and 98.98% across the RPI369, RPI2241, RPI488, and RPI1807 datasets, respectively [71,72]. To directly capture the bidirectional influence between paired sequences, the PIPR model [73] employs a Siamese Residual Recurrent CNN architecture that combines local residual convolutional features with global contextual information in an end-to-end, sequence-only pipeline.

5.1.3. Recurrent Neural Networks (RNNs) Based Models

Complementing convolutional frameworks, recurrent neural networks (RNNs), including Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs), are uniquely designed to resolve long-range sequential dependencies and contextual rules within amino acid sequences. For instance, a foundational LSTM pipeline developed in 2019 [74] mapped primary structures to numerical vectors using text-based protein signatures and a Prot2Vec (Protein2Vector) encoding strategy, outperforming alternative numerical mapping configurations on ROC, log loss, and accuracy. This concept was advanced by Sho Tsukiyama et al. using word2vec sequence encodings [75] inside an LSTM classifier, a strategy that directly inspired the development of LSTM-PHV [76].

Specifically engineered to map human–virus protein interactions strictly from raw sequences, LSTM-PHV handles highly skewed positive-to-negative sample distributions to deliver an AUC of 0.976 and 0.973 alongside accuracies of 0.984 and 0.985 on training and independent datasets, respectively, while outperforming existing baselines on entirely unknown or novel viral species. To address pandemic threats like COVID-19, researchers engineered a specialized mapping system [77] that normalizes and feeds distinct numerical descriptors—including Electron-Ion Interaction Potentials (EIIP), Complex Prime Number Representations (CPNR), and hydrophobicity scales—into a Deep Bidirectional Recurrent Neural Network (DeepBiRNN), achieving an average accuracy of 97.76% in predicting physical interactions between SARS-CoV-2 viral proteins and human host receptors.

5.1.4. Transfer-Based Models

The paradigm has been further advanced by transfer learning and transformer-based architectures, which leverage massive language models pre-trained on large sequence repositories. The ProBERT model [78], inspired by BERT, learns deep contextual embeddings from vast sequence data that transfer effectively to downstream prediction tasks, as demonstrated by the hybrid ProtBert-BiGRU-Attention model [79], which extracts amino acid features via ProtBERT, processes sequential context through a Bidirectional GRU, and applies an attention mechanism to isolate critical interactive features for highly accurate binary classification.

Similarly, the D-SCRIPT deep learning framework [80] tracks cross-species generalizability by mapping raw sequence strings into implicit structural contact representations to evaluate physical binding compatibility; this architecture was successfully utilized to screen the *Bos taurus* genome at a genome-wide scale to isolate functional metabolic and immune gene modules linked to rumen physiology. Furthermore, the Evolutionary Scale Modeling (ESM) framework [81] trains a deep contextual language model on 86 billion amino acids from 250 million structurally diverse, unlabeled protein sequences via self-supervised learning, automatically capturing multi-scale biological representations ranging from elemental physicochemical attributes to remote protein homologies to optimize downstream target prediction.

 doi:10.20944/preprints202605.2116.v1

5.1.5. Autoencoder-Based Models

In parallel, unsupervised architectures such as autoencoders are heavily utilized to compute compact, low-dimensional latent spaces from raw features before interaction mapping. A notable configuration combines a Stacked Sparse Autoencoder (SSAE) with Legendre Moment (LM) feature extraction, passing the compressed representations to a Probabilistic Classification Vector Machine (PCVM) to yield high 5-fold cross-validation accuracies across human (98.58%), unbalanced-human (97.71%), *H. pylori* (93.76%), and *S. cerevisiae* (96.55%) datasets, outperforming traditional standalone SVM baselines.

Similarly, DeepPPI [16] employs a deep feedforward network to extract clean, low-dimensional representations from common protein descriptors, securing a test accuracy of 92.50%. Moving toward generative variations, Variational Autoencoders (VAEs) map protein features to probabilistic continuous latent representations to generate highly informative embeddings; this paradigm is exemplified by the Signed Variational Graph Autoencoder (S-VGAE) model [82], which synthesizes sequence-derived feature vectors with topological structural elements to yield robust latent representations tailored for multi-scale PPI prediction networks.

5.2. Geometry-aware Deep Learning Approaches for PPIs Prediction

Proteins don't function alone; they interact with each other to perform a specific biological process. Information on how different proteins operate in coordination with one another to enable biological processes within the cell is primarily captured by the Protein-protein interaction (PPI) network [83]. The theory of complex networks is fundamental to many fields, including molecular and population biology, engineering, physics, sociology, and computer science [84]. The potential applications of graphical network analysis include determining a protein's or gene's function, identifying potential drug targets, designing effective strategies for treating various diseases, and enabling early diagnosis of disorders [84]. Graph network-based analysis plays a significant role in understanding the essential topological characteristics of biological networks. An unweighted, undirected graph is a typical representation of a protein-protein interaction (PPI) network, where each node represents a protein and an edge between two nodes denotes that these proteins have been found to physically interact [85]. Representing a protein-protein network as a graphical network and studying its topological characteristics helps researchers understand and extract a large amount of information from the network.

Though we have provided a mathematical representation of the PPI network in the form of a graph, $G = (V, E)$, from a deep learning perspective, we now generalize the graph representation of the PPI network as $G = (V, E, X, A)$, where $V = \{v_1, v_2, \ldots, v_n\}$ is the set of proteins (nodes) and a set of edges $E \subseteq V \times V$ represents the interaction between protein pairs (or pairs of nodes) [86]. Each node v_i is associated with a d-dimensional feature vector x_i (i.e., $x_i \in R^d$) that describes biological attributes such as sequence, structure, function, evolutionary profile, etc. The interactions are encoded by an adjacency matrix $A \subseteq R^{n \times n}$, where $A_{ij} = 1$ if proteins i and j interact; otherwise $A_{ij} = 0$. $X \in R^{n \times d}$ represents the node feature matrix, where n is the number of proteins and d is the number of associated features. This representation preserves both the topological and biological characteristics of proteins and enables graph neural network models to make better predictions and inferences. The **Figure 3** illustrates the general graphical architecture of protein–protein interaction that can be used as input to geometry-aware deep learning models for PPI prediction.

 doi:10.20944/preprints202605.2116.v1

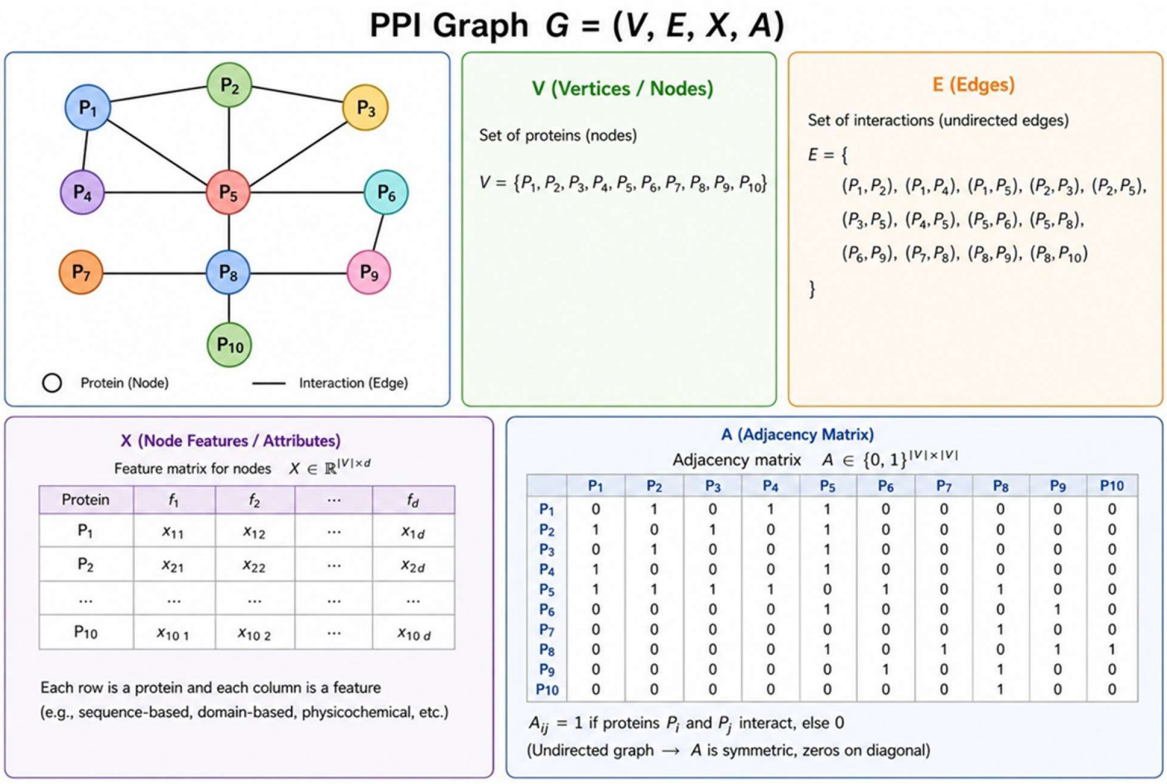

Figure 3. Mathematical and structural formulation of a Protein-Protein Interaction (PPI) graph. The network is formally defined as a tuple G = (V, E, X, A) to prepare biological interaction data for graph-based machine learning pipelines. Top Left (Visual Graph Topology): Spatial visualisation of the undirected PPI network where circles represent individual protein nodes (P_1 … P_{10}) and interconnecting lines represent physical or functional interactions (edges). Top Centre (Node Set, V): Definition of the vertex set containing all ten unique proteins present in the system, bounding the dimensionality of the graph. Top Right (Edge Set, E): The set of unordered pairs representing undirected interactions between adjacent proteins (e.g., (P_1, P_2)), establishing the network's structural connectivity. Bottom Left (Feature Matrix, X): A tabular matrix of dimensions |V| × d, mapping each protein row to a d-dimensional feature vector containing numerical attributes (e.g., amino acid sequence embeddings, structural domains, or physicochemical properties). Bottom Right (Adjacency Matrix, A): A square, binary matrix of size |V| × |V| where a value of 1 denotes a documented interaction between protein Pi and protein Pj, and 0 denotes no interaction. The matrix is symmetric with a zero diagonal due to the undirected nature of the graph.

Graph neural network (GNN) learns low-dimensional node embedding that preserves both the local and global network topology for efficient prediction of PPIs. In [87], a novel Protein-Protein Interaction (PPI) model is proposed that uses graph structure and protein primary-structure information to generate protein representations for PPI prediction. There are two main phases in this PPI prediction task: the representation phase and the prediction phase. Firstly, during the protein representation phases, protein sequences are one-hot encoded, and GCNs are used to capture graph-structural information. Following processing by the encoding and GCN modules, two matrices are obtained that contain, respectively, information about the protein sequence and the graph structure. The feature matrix and adjacent matrix are input to the model. Then, the graph structure and protein sequence information are combined to get the final representation matrix of the proteins. To make PPI predictions, this combined matrix is fed into a fully connected Deep Neural Network (DNN) to extract high-level features and make predictions. The proposed model is being validated using three datasets: human, yeast, and S.cerevisiae. The proposed method outperforms two existing sequence-based PPI models, DPPI [88] and DeepFE-PPI [65], demonstrating excellent predictive performance. Compared with the two existing PPI methods, it is found that the proposed method performed extremely well on the Yeast dataset when compared to DeepFE-PPI and DPPI. It is 2.54% and 5.78% higher than DeepFE-PPI and DPPI. The proposed method also performed extremely well on the human dataset, in contrast to DeepFE-PPI and DPPI. The proposed method is 2.69% and 4.15% higher than DeepFE-PPI and

 doi:10.20944/preprints202605.2116.v1

DPPI, respectively, on the human dataset. The proposed method also demonstrated competitive predictive performance on the S. cerevisiae dataset and outperformed the other two sequence-based PPI models with high accuracy. This proposed method uses a computationally simpler approach to protein representations and predicts with high accuracy compared with existing sequence-based complex PPI models.

Based on the underlying architectural design and learning mechanism, the geometry-aware deep learning approaches for PPI prediction can be subdivided into the following categories:

5.2.1. Spatial-Based GNNs

Spatial GNNs learn embedding by sampling and aggregating information from local neighbourhoods. This is scalable for large biological interaction networks [89,90]. The authors of [89] propose GraphSAGE, a general inductive framework that leverages node features (e.g., text attributes) to efficiently generate node embeddings for previously unseen data. They classify unseen nodes in evolving information graphs using citation and Reddit post data. The single embedding (h_v^k) of node v in layer k in GraphSAGE by aggregating the feature information from a sample neighborhood is done by the following rule:

$$h_v^{(k)} = \sigma(W^{(k)} CONCAT(h_v^{(k-1)}, AGG(h_u^{(k-1)} | u \in N_s(v))),$$

where (h_v^k) is the representation of node v at layer k, $N(v)$ denotes the set of neighboring nodes of v, AGG represents the aggregation function, W is the learnable weight matrix, and sigma (σ) is the activation function. Another, GNNs based on graph structures and message passing (MPNN) adeptly capture local patterns and global relationships in protein structures [91]. Research in [92] presents a new theory for learning link prediction heuristics, justifying learning from local subgraphs rather than entire networks, and proposes SEAL, a novel link prediction framework based on GNNs. The authors in [93] introduced DL-PPI, a novel deep learning framework for sequence-based PPI prediction. It improves feature extraction from individual protein sequences and captures inter-protein relationships using a novel Feature Relationship Network (FRN) based on Graph Neural Networks, thereby improving PPI prediction accuracy.

5.2.2. Spectral-Based GNNs

This research study [94] employs Graph Neural Networks (GNNs), such as the Graph Convolutional Network (GCN), to predict protein interactions using proteins' structural and sequence characteristics. The PDB files are used to build protein graphs, which contain 3D coordinates of atoms. The amino acid network, also known as the residue contact network, is a protein graph in which each amino acid residue is a node and the connections between them are edges. If amino acids have a pair of atoms (one from each amino acid) within the threshold distance, they are said to be connected. In [95], the authors propose a spatial graph convolution model that combines learned features across protein pairs to classify amino acid residue pairs as part of an interface. The architecture proposed here predicts protein-protein interfaces using a graph representation of the underlying protein structure. Here, GCN learns node embedding by aggregating the neighbourhood information layer-wise by using the following rule:

$$H^{(k+1)} = \sigma(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}H^{(k)}W^{(k)}),$$

where $H^0 = X$, $H^{(k)} \in R^{n \times d}$ is the node embedding matrix at layer k, $W^{(k)}$ is a trainable weight matrix, D is the degree matrix, and σ is an activation function.

5.2.3. Attention-Based GNNs

An attention-based graph neural network model assigns different weights to neighbours using an attention mechanism. The authors in [96] propose a novel multi-layer mutual graph attention network (GAT) based architecture, termed Struct2Graph, for the task of protein–protein interaction (PPI) prediction. The model takes as input coarse-grained structural representations of a given protein pair and produces the probability of interaction between the two proteins. Struct2Graph employs two graph convolutional networks (GCNs) with shared weights, along with a mutual attention mechanism, to capture and learn important geometric features that characterize the interaction patterns between the query protein pair. The research in [97] developed a GAT-based method, GAT-GO, that uses RaptorX to predict a protein's structure and Facebook's ESM-1b [81] to generate its embedding. It enhances model capacity by allowing flexible node

 doi:10.20944/preprints202605.2116.v1

feature aggregation through self-attention. To more effectively extract protein information for predicting multi-type PPIs, two attention network frameworks were integrated to construct a learning network termed AFTGAN [98]. This study assigns different weights to relationships among protein nodes, more effectively capturing protein relationships through the attention mechanism. The embedding of node or protein *i* at layer *k* is computed as a weighted combination of neighbour features: $h_i^{(k+1)} = \sigma(\sum_{j \in N(i)} \alpha_{ij} W h_j^{(k)})$. To improve stability, multiple attention heads are used for averaging in the final layer using the following formula for protein *i* at layer *(k + 1)*:

$$h_i^{(k+1)} = \sigma\left(\frac{1}{M}\sum_{i=1}^{M}\sum_{j \in N(i)} \alpha_{ij}^{M} W^m h_j^{(k)}\right)!$$

5.2.4. Graph Autoencoders

The authors in [82] consider the PPI network as an undirected graph, and propose a model based on a signed variational graph auto-encoder (S-VGAE) and a conjoint triad (CT) that effectively leverages the naturally incorporated protein sequence information and graph structure as features. The overall framework is composed of three parts. Firstly, the raw protein sequences were encoded using CT methods; secondly, vector embeddings for each protein were extracted from both sequence information and graph structure using the essential S-VGAE model. A simple three-layer feed-forward neural network is used for the classification purpose of Protein-Protein Interactions, and to train this feed-forward neural network, these S-VGAE model embeddings were used as input. The proposed model achieved high accuracy on the Human Protein Reference Database (HPRD) dataset and obtained the best results on the Caenorhabditis elegans (C. elegans), Database of Interacting Proteins (DIP) Human, Drosophila, and Escherichia coli (E. coli) datasets. The model's prediction capability is also validated against existing PPI methods, and it outperforms all existing models with excellent accuracy. Variational auto-encoder (VGAE) extends the graph auto-encoder by learning probabilistic latent embeddings of proteins in a PPI network. It models uncertainty in protein representations, which improves predictions of unknown interactions. Mathematically, instead of learning a deterministic embedding, VGAE learns the following distribution:

$$q(Z|X, A) = \prod_{i=1}^{n} N(z_i|\mu_i, diag(\sigma_i^2)),$$

where z_i is the latent embedding of protein *i*, μ_i is the mean vector, and σ_i is the standard deviation of the Gaussian Normal distribution *N*.

5.2.5. Dynamic/Temporal GNNs

Dynamic or temporal graph neural networks have been applied to PPI prediction to capture evolving interaction patterns across biological conditions. In [99], Chen et. al. proposed DCMF-PPI, which models temporal variations in protein structure and interaction networks using dynamic adjacency matrices and geometry-aware learning techniques. DCMF-PPI is a hybrid framework that combines dynamic modeling, multi-scale feature extraction, and probabilistic graph representation learning. It includes a PortT5-GAT module, where the PortT5 protein language model extracts residue-level features, incorporates temporal dependencies, and uses graph attention networks to capture context-aware structural variations in protein interactions. In temporal GNNs, protein representations evolve over time by integrating both structural and temporal dependencies. The general temporal update rule is [100]:$h_i^{(t+1)} = f(h_i^{(t)}, G_t)$, where $h_i^{(t)}$ denotes embedding of protein *i* at time step *t*, *f(.)* represents temporal graph learning function combining graph structure historical information, and $G_t = (V_t, E_t, X_t)$ denotes the graph at time step *t* where V_t is the set of proteins, E_t represents PPI interaction, and X_t denotes the node feature matrix at time step *t*. The other dynamic GNN models incorporate recurrent neural network (RNN) to capture temporal dependencies using the following formula [100,101]:

$$h_i^{(t+1)} = GRU(h_i^{(t)}, GNN(A_t, X_t)_i),$$

where A_t is the adjacency matrix at time step *t*, and $GNN(A_t, X_t)$ computes the structural embedding of nodes from a graph snapshot.

 doi:10.20944/preprints202605.2116.v1

5.2.6. Heterogeneous GNNs

Heterogeneous graph neural networks (Het-GNNs) are designed to model multi-type biological entities and relationships that naturally arise in bioinformatics data. In PPI prediction, heterogeneous graphs enable the integration of diverse biological information, such as proteins, genes, drugs, diseases, pathways, and functional annotations, within a unified framework. This kind of integration improves predictive performance by capturing complementary biological knowledge from multiple sources. Heterogeneous PPI networks comprise multiple node types and interaction types, providing a richer representation than homogeneous graphs. A heterogeneous biological interaction network represented as: $G = (V, E, X, \Gamma_v, \Gamma_e)$, where V is the set of proteins, E is the set of interactions, X is the node feature matrix, Γ_v is the set of different node types, and Γ_e is the set of different edge types. In heterogeneous GNNs [101], neighbours are grouped based on relation type. For node i, neighbour under relation r *is* defined as:

$$N_r(i) = \{j | (i, j, r) \in E\},$$

where $(i, j, r) \in E$ indicates that there exists an edge between nodes i and j with relation type r. The message aggregation for each relation is defined as:

$$m_i^{(k,r)} = \sum_{j \in N_r(i)} \frac{1}{C_{i,r}} W_r^{(k)} h_j^{(k-1)},$$

where $W_r^{(k)}$ relation-specific weight matrix, $C_{i,r}$ normalized constant, $h_j^{(k-1)}$ is the embedding of the neighbour node. The compact mathematical formulation for heterogeneous GNN is [90,101]:

$$H^{(k+1)} = \sigma\left(\sum_{r \in \Gamma_e} D_r^{-1} A_r H^{(k)} W_r^{(k)}\right)$$

where, A_r is the adjacency matrix for relation type r, D_r is the degree matrix for relation r, $W_r^{(k)}$ is a learnable parameter matrix, σ is nonlinear activation function. The authors in [90] develop Decagon, a method for predicting side effects of drug pairs by constructing a two-layer multimodal graph comprising protein–protein, drug–protein, and drug–drug interactions (side effects). Each drug–drug interaction is represented as a distinct edge type corresponding to a specific side effect. A multi-relational GCNs is then developed to predict both drug–drug interactions and their associated side effect types. Exploratory analysis shows that co-prescribed drug pairs often share common target proteins, indicating that drug–protein interaction information is valuable for modelling drug combinations.

5.2.7. Structure- or Residue-Aware GNNs

Structure-aware or residue-aware GNNs use 3D structural information of proteins to improve PPI predictions. Instead of representing proteins only at the sequence level, these approaches construct a graph in which nodes correspond to amino acid residues and edges represent spatial proximity or biochemical relationships derived from protein tertiary structures. Such structural representations enable the capture of geometric and physicochemical patterns that determine binding affinity and interaction specificity. Here, the protein structure graph can be represented as $G = (V_r, E_r, X_r)$, where V_r is the set of amino acid residues, E_r is the set of structural connections between residues, and X_r is the residue feature matrix. The structure-aware message passing incorporates the spatial relationships in the embedding space, which is defined as:

$$h_i^{(k+1)} = \sigma\left(\sum_{j \in N(i)} W^{(k)} h_j^{(k)}\right)$$

$$h_i^{(k+1)} = \sigma\left(\sum_{j \in N(i)} \phi\left(h_i^{(k)}, h_j^{(k)}, e_{ij}\right)\right)$$

where $\phi(.)$ is the geometric message function, and e_{ij} is the spatial relation between residues. Jing et al. in [102] introduce Geometric Vector Perceptrons (GVPs), which extend standard dense layers to operate on collections of Euclidean vectors and serve as a drop-in replacement for multi-layer perceptrons (MLPs) in GNN

 doi:10.20944/preprints202605.2116.v1

aggregation and feed-forward layers. GVPs process both scalar and geometric features that transform consistently under spatial rotations, enabling effective learning from protein structural data. The authors demonstrate the approach on protein structure-related tasks, including computational protein design (CPD), which aims to predict an amino acid sequence that folds into a given structure, and model quality assessment. Another work on end-to-end learning for 3D protein structure-based interface prediction is presented by Townshend et al. [103], who introduce DIPS, a dataset for interface prediction that is two orders of magnitude larger than previously available datasets. To address the limitations of hand-crafted features in handling dataset bias, they propose SASNet, the first end-to-end deep learning framework for protein interface prediction.

6. Other Methods (Table 4, Figure 2)

6.1. Bayesian and Probabilistic Models

Bayesian and probabilistic models provide a principled framework for modeling uncertainty and integrating heterogeneous biological data in protein–protein interaction (PPI) prediction. Unlike deterministic approaches, probabilistic methods explicitly represent uncertainty in both data and model parameters, making them particularly suitable for biological datasets that are often noisy, incomplete, and high-dimensional.

Table 4. Representative Bayesian, Probabilistic, and Temporal Protein–Protein Interaction Network Frameworks.

Study	Framework Category	Biological Context	Temporal Modeling	Split Protocol	Validation Strategy	Primary Metrics	Standardized Performance	Methodological Contribution	Key Limitations
Morshed et al. (2012) [106]	Bayesian co-learning	IRMA OFF dataset	No	Random split	Cross-validation	Precision, Recall	Precision: 50.0%; Recall: 75.0%	Probabilistic Gaussian-process integration	Limited scalability
Liu et al. (2019) [104]	Probabilistic feature fusion	Extended yeast PPIN	No	Random split	Cross-validation	AUROC, AUPR	AUROC: 69.0%; AUPR: 50.0%	Heterogeneous probabilistic feature integration	Limited external validation
Wang et al. (2023) [105]	Markov Random Field	Pathway activity networks	Partial	Unknown	Not reported	Activity significance	Score range: 76–394	Pathway-aware probabilistic inference	Conventional classification metrics absent
Li et al. (2024) [107]	Temporal representation learning	Dynamic PPIN	Yes	Temporal split	Temporal validation	Accuracy, Precision, Recall	Acc: 60.0%; Pre: 82.0%; Rec: 50.0%	Time-aware interaction learning	Limited recall
Li et al. (2024) [109]	Dynamic PPIN reconstruction	DIP temporal dataset	Yes	Temporal split	Reconstruction evaluation	Precision, Recall	Precision: 73.0%; Recall: 71.0%	Dynamic network reconstruction	Limited benchmark standardization
Chen et al. (2025) [99]	Dynamic matrix factorization	SHS27k dataset	Yes	Random partition	Random partition validation	F1-score	F1-score: 89.0%	Dynamic collaborative interaction modeling	Random partition may inflate performance
Megliorini et al. (2025) [108]	Integrative temporal PPIN	Multi-omics PPIN	Yes	Independent test set	Independent testing	Acc, Pre, Rec, ROC-AUC	Acc: 94.0%; AUC: 98.0%	Multi-omics temporal interaction inference	Requires large-scale omics data
He et al. (2025) [110]	Dynamic PPIN reconstruction	Babu dataset	Yes	Independent test set	Independent testing	Acc, Pre, Rec, MCC	Acc: 91.0%; MCC: 84.0%	Temporal network reconstruction	Limited biological validation

Abbreviations: AAC, Amino Acid Composition; AC, Autocovariance; AUC, Area Under the Receiver Operating Characteristic Curve; AUPR, Area Under the Precision–Recall Curve; CNN, Convolutional Neural Network; CV, Cross-Validation; DHT, Discrete Hartley Transform; DL, Deep Learning; GAT, Graph Attention Network; MCC, Matthews Correlation Coefficient; PPI, Protein–Protein Interaction; PPIN, Protein–Protein Interaction Network; RF, Random Forest;

 doi:10.20944/preprints202605.2116.v1

ROC-AUC, Receiver Operating Characteristic Area Under the Curve; RPI, RNA–Protein Interaction; SVM, Support Vector Machine; VGAE, Variational Graph Autoencoder.

In the Bayesian framework, the probability of interaction between a pair of proteins is inferred by combining prior knowledge with observed data using Bayes' theorem. Formally, given a protein pair *(i, j)* and observed features x_{ij} derived from sequence, structural, network, or functional information, the posterior probability of interaction can be expressed as:

$$P(y_{ij} = 1|x_{ij}) = \frac{P(x_{ij}|y_{ij} = 1)P(y_{ij} = 1)}{P(x_{ij})}$$

where y_{ij} denotes the interaction label. This formulation enables the integration of multiple evidence sources while accounting for their relative contributions and uncertainties.

Several studies have demonstrated the effectiveness of probabilistic data fusion strategies for PPI prediction. For instance, Liu et al. [104] proposed a fusion framework that combines diverse biological features using probabilistic modeling to improve prediction accuracy. Similarly, Wang et al. [105] introduced an active learning-based probabilistic approach that iteratively selects informative samples to enhance model performance under limited labelled data. Earlier work by Morshed et al. [106] employed Gaussian process-based probabilistic models to integrate multiple genomic and proteomic data sources, demonstrating improved robustness and generalization.

6.2. Temporal PPINs

Temporal PPINs aim to capture these time-varying interaction patterns, providing a more realistic and informative representation of cellular processes. A temporal PPIN can be modeled as a sequence of time-indexed graphs $\{G_t = (V_t, E_t)\}_{t=1}^{T}$, where V_t and E_t denote the sets of proteins and interactions at time t, respectively. Alternatively, continuous-time formulations represent interactions as timestamped events, enabling fine-grained modeling of interaction dynamics. Such representations enable the analysis of network evolution, the identification of transient interactions, and the detection of temporal functional modules.

Recent advances have focused on integrating heterogeneous temporal data sources and developing sophisticated models for dynamic PPI prediction. For instance, Li et al. [107] proposed a temporal modeling framework that captures dynamic interaction patterns using time-aware representations. Chen et al. [99] introduced a dynamic collaborative matrix factorization approach to model temporal dependencies and latent interaction structures. Similarly, Megliorini et al. [108] developed an integrative framework that combines multi-omics temporal data to enhance the reconstruction of dynamic PPINs.

In addition, the construction and reconstruction of temporal PPINs from noisy and incomplete data have been extensively studied. Li et al. [109] proposed methods for constructing temporal networks from time-series biological data, while He et al. [110] focused on reconstructing dynamic interaction networks using advanced inference techniques. These approaches highlight the importance of incorporating temporal information to better understand the evolution of biological systems. Table 4 compares several DL-based PPI prediction approaches discussed above.

Direct numerical comparisons across studies should be interpreted with caution because reported performance depends strongly on benchmark construction, negative-sampling strategy, class balance, split protocol, and evaluation metric. Accordingly, the comparative tables in this review are intended as structured summaries of the literature rather than strict cross-study rankings.

7. Discussion and Future Directions

The critical synthesis of computational frameworks for human–virus protein–protein interaction (PPI) prediction reveals a profound paradigm shift from descriptive, low-throughput experimental mapping to predictive, multi-scale in silico architecture. While traditional feature-based machine learning approaches offer exceptional computational efficiency and robust baseline performance when mapping interactions using basic sequence-derived physicochemical descriptors, they are fundamentally limited by their reliance on manual feature engineering and an inability to natively capture long-range contextual or spatial dependencies. Conversely, advanced deep learning models resolve these constraints through autonomous hierarchical feature abstraction. Within the sequential domain, large-scale pre-trained protein language

 doi:10.20944/preprints202605.2116.v1

models (LMs) have revolutionized predictive sensitivity by capturing deep evolutionary grammars and remote homologies directly from raw sequences, whereas geometry-aware graph neural networks (GNNs) preserve macro-level interactome topologies and spatial configurations.

However, a systematic evaluation of these state-of-the-art frameworks highlights major unresolved bottlenecks that hinder their clinical and translational deployment. Current predictive pipelines are heavily constrained by a data representation crisis, wherein public repositories remain noisy, biased, and fundamentally lacking in experimentally validated true-negative interactions, forcing models to rely on synthetic random negative sampling that artificially inflates accuracy metrics. Furthermore, the vast majority of high-performing deep learning models function as mathematical „black boxes," offering no clear biochemical rationale or residue-level transparency to explain *why* an interaction occurs, which severely limits their utility in targeted drug design or vaccine target selection. Crucially, these frameworks typically treat the interactome as a static, binary graph, completely abstracting away the highly volatile, context-dependent reality of viral biology, where rapid mutation rates, post-translational modifications (PTMs), and cellular localization continuously alter binding affinities.

In addressing this fragmented landscape, the primary strength of this review article lies in its comprehensive, end-to-end taxonomy that systematically bridges the gap between raw biological data modalities (sequence, structure, genomic context, GO annotations) and specific non-geometry and geometry-aware deep learning configurations. By organizing these disparate methodologies into a cohesive conceptual pipeline, this work illuminates the precise architectural trade-offs inherent in modern predictive models. Conversely, an inherent limitation of this review arises from the fast-evolving nature of the deep learning field and the highly heterogeneous evaluation metrics reported across the literature, which prevents a standardized, universal quantitative meta-analysis of algorithmic performance. Additionally, its primary focus on human-virus systems, with a particular spotlight on coronaviruses, means that certain topological anomalies unique to plant or bacterial interactomes are abstracted away.

Nevertheless, this review directly assists the scientific community by serving as a definitive methodological roadmap and decision-making manual for both computational architects and structural biologists. It eliminates the trial-and-error phase of framework design by allowing researchers to instantly align their available biological data types with the most resilient algorithmic paradigms. By establishing this clear operational baseline, this work accelerates the realization of the immediate research frontier: the development of synergistic multimodal data fusion networks capable of co-embedding primary sequences, 3D structural coordinates generated by structural prediction pipelines, and functional genomic annotations into unified, multi-scale feature spaces. Overcoming the interpretability barrier requires the systematic integration of Explainable AI (XAI) layers—such as attention-weight tracking and subgraph-relevance propagation—which will allow networks to move beyond binary interaction classification and precisely pinpoint the exact amino acid residues and spatial interfaces driving physical binding. Furthermore, to effectively counter highly mutable viral threats, computational paradigms must decouple themselves from static training distributions by engineering evolutionary domain adaptation and zero-shot learning frameworks. By capturing evolving evolutionary trajectories, these future networks will be able to proactively forecast the host interactomes of emerging zoonotic strains or newly mutated variants long before empirical datasets can be collected. Finally, integrating these intelligent pipelines with downstream virtual screening, molecular docking, and systems biology workflows will shift the field from a posture of reactive model calibration to one of proactive, structure-guided intervention, empowering global healthcare frameworks to design resilient, mutation-resistant antiviral therapeutics at unprecedented scales.

Author Contributions: S.B. contributed to 1) the conception and organization of the study and 2) the writing of the first draft of the manuscript; D.A contributed to 1) the conception and organization of the study and 2) the writing of the first draft of the manuscript; S.C., S.D. and R.M. contributed 1) the conception, organization, and execution of the research project, and 2) the review and critique of the manuscript; K.G. and J.R. contributed to 1) the conception, organization, 2) the review and critique of the manuscript research project; A.C. and J.B.L. contributed to 1) the conception, organization, 2) the review and critique of the manuscript research project. All authors have reviewed and approved the final manuscript.

Funding: This research received no external funding.

 doi:10.20944/preprints202605.2116.v1

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding authors.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. David S Hui, Esam I Azhar, Tariq A Madani, Francine Ntoumi, Richard Kock, Osman Dar, Giuseppe Ippolito, Timothy D Mchugh, Ziad A Memish, Christian Drosten, et al. The continuing 2019-ncov epidemic threat of novel coronaviruses to global health—the latest 2019 novel coronavirus outbreak in wuhan, china. *International journal of infectious diseases*, 91:264–266, 2020.
2. JONATHAN M Read, Jessica RE Bridgen, Derek AT Cummings, A Ho, and CP Jewell. Novel coronavirus 2019-ncov: early estimation of epidemiological parameters and epidemic predictions. medrxiv 2020.01. 23.20018549. *DOI*, 10(2020.01):23–20018549, 2020.
3. Victor M Corman, Olfert Landt, Marco Kaiser, Richard Molenkamp, Adam Meijer, Daniel KW Chu, Tobias Bleicker, Sebastian Bru¨nink, Julia Schneider, Marie Luisa Schmidt, et al. Detection of 2019 novel coronavirus (2019-ncov) by real-time rt-pcr. *Eurosurveillance*, 25(3):2000045, 2020.
4. Simon James Fong, Gloria Li, Nilanjan Dey, Rub'en Gonz'alez Crespo, and Enrique HerreraViedma. Composite monte carlo decision making under high uncertainty of novel coronavirus epidemic using hybridized deep learning and fuzzy rule induction. *Applied soft computing*, 93:106282, 2020.
5. Dawei Wang, Bo Hu, Chang Hu, Fangfang Zhu, Xing Liu, Jing Zhang, Binbin Wang, Hui Xiang, Zhenshun Cheng, Yong Xiong, et al. Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in wuhan, china. *jama*, 323(11):1061–1069, 2020.
6. Yanzhong Huang et al. The sars epidemic and its aftermath in china: a political perspective. *Learning from SARS: Preparing for the next disease outbreak*, pages 116–36, 2004.
7. Lee Shiu Hung. The sars epidemic in hong kong: what lessons have we learned? *Journal of the Royal Society of Medicine*, 96(8):374–378, 2003.
8. KH Kim, TE Tandi, Jae Wook Choi, JM Moon, and MS Kim. Middle east respiratory syndrome coronavirus (mers-cov) outbreak in south korea, 2015: epidemiology, characteristics and public health implications. *Journal of Hospital Infection*, 95(2):207–213, 2017.
9. Sara Hosseinzadeh Kassania, Peyman Hosseinzadeh Kassanib, Michal J Wesolowskic, Kevin A Schneidera, and Ralph Detersa. Automatic detection of coronavirus disease (covid-19) in x-ray and ct images: a machine learning based approach. *Biocybernetics and Biomedical Engineering*, 41(3):867–879, 2021.
10. Rahul Kumar, Ridhi Arora, Vipul Bansal, Vinodh J Sahayasheela, Himanshu Buckchash, Javed Imran, Narayanan Narayanan, Ganesh N Pandian, and Balasubramanian Raman. Accurate prediction of covid-19 using chest x-ray images through deep feature learning model with smote and machine learning classifiers. *MedRxiv*, pages 2020–04, 2020.
11. Yu Lung Lau and JS Malik Peiris. Pathogenesis of severe acute respiratory syndrome. *Current opinion in immunology*, 17(4):404–410, 2005.
12. Haibo Zhang, Josef M Penninger, Yimin Li, Nanshan Zhong, and Arthur S Slutsky. Angiotensin converting enzyme 2 (ace2) as a sars-cov-2 receptor: molecular mechanisms and potential therapeutic target. *Intensive care medicine*, 46:586–590, 2020.
13. Roujian Lu, Xiang Zhao, Juan Li, Peihua Niu, Bo Yang, Honglong Wu, Wenling Wang, Hao Song, Baoying Huang, Na Zhu, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *The lancet*, 395(10224):565–574, 2020.
14. Juwen Shen, Jian Zhang, Xiaomin Luo, Weiliang Zhu, Kunqian Yu, Kaixian Chen, Yixue Li, and Hualiang Jiang. Predicting protein–protein interactions based only on sequences information. *Proceedings of the National Academy of Sciences*, 104(11):4337–4341, 2007.
15. Z.-H. You, Z. Yin, H.-C. Han, D.-S. Huang, and X. Zhou. Prediction of protein–protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis. *BMC Bioinformatics*, 15(S2):S10, 2014.

 doi:10.20944/preprints202605.2116.v1

16. Xiuquan Du, Shiwei Sun, Changlin Hu, Yu Yao, Yuanting Yan, and Yanping Zhang. DeepPI: boosting prediction of protein–protein interactions with deep neural networks. *Journal of chemical information and modeling*, 57(6):1499–1510, 2017.
17. S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, 1997.
18. The UniProt Consortium. Uniprot: a worldwide hub of protein knowledge. *Nucleic Acids Research*, 47(D1):D506–D515, 2019.
19. B. Boeckmann, A. Bairoch, R. Apweiler, M.-C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, and M. Schneider. The swiss-prot protein knowledgebase and its supplement trembl in 2003. *Nucleic Acids Research*, 31(1):365–370, 2003.
20. C. B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181(4096):223–230, 1973.
21. C. Chothia. One thousand families for the molecular biologist. *Nature*, 357:543–544, 1992.
22. O. Keskin, A. Gursoy, B. Ma, and R. Nussinov. Principles of protein–protein interactions: what are the preferred ways for proteins to interact? *Chemical Reviews*, 108(4):1225–1244, 2008.
23. A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. Scop: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247(4):536–540, 1995.
24. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 2000.
25. C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton. Cath–a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1108, 1997.
26. M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.
27. The Gene Ontology Consortium. The gene ontology resource: enriching a gold mine. *Nucleic Acids Research*, 49(D1):D325–D334, 2021.
28. P. W. Lord, R. D. Stevens, A. Brass, and C. A. Goble. Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation. *Bioinformatics*, 19(10):1275–1283, 2003.
29. J. Z. Wang, Z. Du, R. Payattakool, P. S. Yu, and C. F. Chen. A new method to measure the semantic similarity of go terms. *Bioinformatics*, 23(10):1274–1281, 2007.
30. R. P. Huntley, T. Sawford, P. Mutowo-Meullenet, A. Shypitsyna, C. Bonilla, M. J. Martin, and C. O'Donovan. Quickgo: a user-friendly web-based tool for gene ontology searching. *Bioinformatics*, 31(20):3352–3353, 2015.
31. Francis Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.
32. M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Sciences*, 96(8):4285–4288, 1999.
33. R. Overbeek, M. Fonstein, M. D'Souza, G. D. Pusch, and N. Maltsev. The use of gene clusters to infer functional coupling. *Proceedings of the National Academy of Sciences*, 96(6):2896–2901, 1999.
34. M. S. Skrzypek, J. Binkley, G. Binkley, S. R. Miyasato, M. Simison, and G. Sherlock. The candida genome database: integration of genomic and functional data for candida species. *Nucleic Acids Research*, 45(D1):D592–D596, 2017.
35. H. W. Mewes, C. Amid, R. Arnold, D. Frishman, U. Gu¨ldener, G. Mannhaupt, M. Mu¨nsterk¨otter, P. Pagel, N. Strack, V. Stumpflen, J. Warfsmann, and A. Ruepp. Mips: analysis and annotation of proteins from whole genomes. *Nucleic Acids Research*, 32:D41–D44, 2004.
36. A.-L. Barab'asi, N. Gulbahce, and J. Loscalzo. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12(1):56–68, 2011.
37. D. Szklarczyk, A. L. Gable, D. Lyon, A. Junge, S. Wyder, J. Huerta-Cepas, M. Simonovic, N. T. Doncheva, J. H. Morris, P. Bork, L. J. Jensen, and C. von Mering. String v11: protein–protein association networks with increased coverage. *Nucleic Acids Research*, 47(D1):D607–D613, 2019.
38. C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers. Biogrid: a general repository for interaction datasets. *Nucleic Acids Research*, 34:D535–D539, 2006.
39. G. D. Bader, I. Donaldson, C. Wolting, B. F. Ouellette, T. Pawson, and C. W. Hogue. Bind: the biomolecular interaction network database. *Nucleic Acids Research*, 31(1):248–250, 2003.
40. I. Xenarios, D. W. Rice, L . Salw'ınski, M. K. Baron, E. M. Marcotte, and D. Eisenberg. Dip: the database of interacting proteins. *Nucleic Acids Research*, 30(1):303–305, 2002.

 doi:10.20944/preprints202605.2116.v1

41. A. Zanzoni, L. Montecchi-Palazzi, G. Quondam, P. Ausiello, and G. Cesareni. Mint: a molecular interaction database. *FEBS Letters*, 513(1):135–140, 2002.
42. S. Perry, A. I. Su, N. Elango, , et al. The human protein reference database (hprd) and human proteinpedia. *Nucleic Acids Research*, 37:D767–D772, 2009.
43. S. Kerrien, B. Aranda, L. Breuza, A. Bridge, F. Broackes-Carter, C. Chen, M. Duesbury, M. Dumousseau, M. Feuermann, U. Hinz, C. Jandrasits, R. C. Jimenez, J. Khadake, U. Mahadevan, P. Masson, I. Pedruzzi, E. Pfeiffenberger, P. Porras, A. Raghunath, B. Roechert, S. Orchard, and H. Hermjakob. The intact molecular interaction database in 2012. *Nucleic Acids Research*, 40:D841–D846, 2012.
44. Yunqian Ma and Guodong Guo. *Support vector machines applications*, volume 649. Springer, 2014.
45. P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–2504, 2003.
46. Lopamudra Dey, Sanjay Chakraborty, and Anirban Mukhopadhyay. Machine learning techniques for sequence-based prediction of viral–host interactions between sars-cov-2 and human proteins. *Biomedical journal*, 43(5):438–450, 2020.
47. Xianyi Lian, Shiping Yang, Hong Li, Chen Fu, and Ziding Zhang. Machine-learning-based predictor of human–bacteria protein–protein interactions by incorporating comprehensive hostnetwork properties. *Journal of proteome research*, 18(5):2195–2205, 2019.
48. Soumyadeep Debnath and Ayatullah Faruk Mollah. A supervised machine learning approach for sequence based protein-protein interaction (ppi) prediction. *arXiv preprint arXiv:2203.12659*, 2022.
49. Yanzhi Guo, Lezheng Yu, Zhining Wen, and Menglong Li. Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences. *Nucleic acids research*, 36(9):3025–3030, 2008.
50. Matthew D Dyer, TM Murali, and Bruno W Sobral. Supervised learning and prediction of physical interactions between human and hiv proteins. *Infection, Genetics and Evolution*, 11(5):917– 923, 2011.
51. Guangyu Cui, Chao Fang, and Kyungsook Han. Prediction of protein-protein interactions between viruses and human by an svm model. In *BMC bioinformatics*, volume 13, pages 1–10. BioMed Central, 2012.
52. Saud Alguwaizani, Byungkyu Park, Xiang Zhou, De-Shuang Huang, and Kyungsook Han. Predicting interactions between virus and host proteins using repeat patterns and composition of amino acids. *Journal of healthcare engineering*, 2018, 2018.
53. Ranjan Kumar Barman, Sudipto Saha, and Santasabuj Das. Prediction of interactions between viral and host proteins using supervised machine learning methods. *PloS one*, 9(11):e112034, 2014.
54. Fatma-Elzahraa Eid, Mahmoud ElHefnawi, and Lenwood S Heath. Denovo: virus-host sequencebased protein–protein interaction prediction. *Bioinformatics*, 32(8):1144–1150, 2016.
55. Xiaodi Yang, Shiping Yang, Qinmengge Li, Stefan Wuchty, and Ziding Zhang. Prediction of human-virus protein-protein interactions through a sequence embedding-based machine learning method. *Computational and structural biotechnology journal*, 18:153–161, 2020.
56. Jie Pan, Shiwei Wang, Changqing Yu, Liping Li, Zhuhong You, and Yanmei Sun. A novel ensemble learning-based computational method to predict protein-protein interactions from protein primary sequences. *Biology*, 11(5):775, 2022.
57. Lei Yang, Jun-Feng Xia, and Jie Gui. Prediction of protein-protein interactions from protein sequence using local descriptors. *Protein and Peptide Letters*, 17(9):1085–1090, 2010.
58. Zhu-Hong You, Ying-Ke Lei, Lin Zhu, Junfeng Xia, and Bing Wang. Prediction of proteinprotein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis. In *BMC bioinformatics*, volume 14, pages 1–11. Springer, 2013.
59. Abbasali Emamjomeh, Bahram Goliaei, Javad Zahiri, and Reza Ebrahimpour. Predicting protein–protein interactions between human and hepatitis c virus via an ensemble learning method. *Molecular Biosystems*, 10(12):3147–3154, 2014.
60. Fen Pei, Qingya Shi, Haotian Zhang, and Ivet Bahar. Predicting protein–protein interactions using symmetric logistic matrix factorization. *Journal of chemical information and modeling*, 61(4):1670–1682, 2021.
61. Yingjun Ma, Yongbiao Zhao, and Yuanyuan Ma. Kernel bayesian nonlinear matrix factorization based on variational inference for human–virus protein–protein interaction prediction. *Scientific Reports*, 14(1):5693, 2024

 doi:10.20944/preprints202605.2116.v1

62. Hua Wang, Heng Huang, Chris Ding, and Feiping Nie. Predicting protein-protein interactions from multimodal biological data sources via nonnegative matrix tri-factorization. In *Annual International Conference on Research in Computational Molecular Biology*, pages 314–325. Springer, 2012.
63. Mark Goldsmith, Harto Saarinen, Guillermo Garc'ıa-P'erez, Joonas Malmi, Matteo AC Rossi, and Sabrina Maniscalco. Link prediction with continuous-time classical and quantum walks. *Entropy*, 25(5):730, 2023.
64. Bin Xu, Jihong Guan, Yang Wang, and Zewei Wang. Essential protein detection by random walk on weighted protein-protein interaction networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 16(2):377–387, 2017
65. Yu Yao, Xiuquan Du, Yanyu Diao, and Huaixu Zhu. An integration of deep learning with feature embedding for protein–protein interaction prediction. *PeerJ*, 7:e7126, 2019.
66. Jingyan Zheng, Xiaodi Yang, Yan Huang, Shiping Yang, Stefan Wuchty, and Ziding Zhang. Deep learning-assisted prediction of protein–protein interactions in arabidopsis thaliana. *The Plant Journal*, 114(4):984–994, 2023.
67. Xue Li, Peifu Han, Gan Wang, Wenqi Chen, Shuang Wang, and Tao Song. Sdnn-ppi: self attention with deep neural network effect on protein-protein interaction prediction. *BMC genomics*, 23(1):474, 2022.
68. Ibrahim Ahmed, Peter Witbooi, and Alan Christoffels. Prediction of human-bacillus anthracis protein–protein interactions using multi-layer neural network. *Bioinformatics*, 34(24):4159–4164, 2018.
69. Zengyan Xie, Xiaoya Deng, and Kunxian Shu. Prediction of protein–protein interaction sites using convolutional neural network and improved data sets. *International journal of molecular sciences*, 21(2):467, 2020.
70. Lei Wang, Xin Yan, Meng-Lin Liu, Ke-Jian Song, Xiao-Fei Sun, and Wen-Wen Pan. Prediction of rna-protein interactions by combining deep convolutional neural network with feature selection ensemble method. *Journal of theoretical biology*, 461:230–238, 2019.
71. Xiaodi Yang, Shiping Yang, Xianyi Lian, Stefan Wuchty, and Ziding Zhang. Transfer learning via multi-scale convolutional neural layers for human–virus protein–protein interaction prediction. *Bioinformatics*, 37(24):4771–4778, 2021.
72. Jiyun Zhou, Qin Lu, Ruifeng Xu, Lin Gui, and Hongpeng Wang. Cnnsite: Prediction of dnabinding residues in proteins using convolutional neural network with sequence features. In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 78–85. IEEE, 2016.
73. Muhao Chen, Chelsea J-T Ju, Guangyu Zhou, Xuelu Chen, Tianran Zhang, Kai-Wei Chang, Carlo Zaniolo, and Wei Wang. Multifaceted protein–protein interaction prediction based on siamese residual rcnn. *Bioinformatics*, 35(14):i305–i314, 2019.
74. Talha Burak Alakus and Ibrahim Turkoglu. Prediction of protein-protein interactions with lstm deep learning model. In *2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, pages 1–5. IEEE, 2019.
75. Long Ma and Yanqing Zhang. Using word2vec to process big text data. In *2015 IEEE International Conference on Big Data (Big Data)*, pages 2895–2897. IEEE, 2015.
76. Sho Tsukiyama, Md Mehedi Hasan, Satoshi Fujii, and Hiroyuki Kurata. Lstm-phv: prediction of human-virus protein–protein interactions by lstm with word2vec. *Briefings in Bioinformatics*, 22(6):bbab228, 2021.
77. Talha Burak Alakus and Ibrahim Turkoglu. A novel protein mapping method for predicting the protein interactions in covid-19 disease by deep learning. *Interdisciplinary Sciences: Computational Life Sciences*, 13:44–60, 2021.
78. Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. Prottrans: toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):7112–7127, 2021.
79. Qian Gao, Chi Zhang, Ming Li, and Tianfei Yu. Protein–protein interaction prediction model based on protbert-bigru-attention. *Journal of Computational Biology*, 31(9):797–814, 2024.
80. Samuel Sledzieski, Rohit Singh, Lenore Cowen, and Bonnie Berger. D-script translates genome to phenome with sequence-based, structure-aware, genome-scale predictions of protein-protein interactions. *Cell Systems*, 12(10):969–982, 2021.
81. Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the national academy of sciences*, 118(15):e2016239118, 2021.
82. Fang Yang, Kunjie Fan, Dandan Song, and Huakang Lin. Graph-based prediction of proteinprotein interactions with attributed signed graph embedding. *BMC bioinformatics*, 21(1):1–16, 2020.

 doi:10.20944/preprints202605.2116.v1

83. Matteo Pellegrini, David Haynor, and Jason M Johnson. Protein interaction networks. *Expert review of proteomics*, 1(2):239–249, 2004.
84. Georgios A Pavlopoulos, Maria Secrier, Charalampos N Moschopoulos, Theodoros G Soldatos, Sophia Kossida, Jan Aerts, Reinhard Schneider, and Pantelis G Bagos. Using graph theory to analyze biological networks. *BioData mining*, 4:1–27, 2011.
85. Nataˇsa Prˇzulj and Desmond J Higham. Modelling protein–protein interaction networks via a stickiness index. *Journal of the Royal Society Interface*, 3(10):711–716, 2006.
86. Min Wu, Xiaoli Li, Chee-Keong Kwoh, and See-Kiong Ng. A core-attachment based method to detect protein complexes in ppi networks. *BMC bioinformatics*, 10(1):169, 2009.
87. Leilei Liu, Yi Ma, Xianglei Zhu, Yaodong Yang, Xiaotian Hao, Li Wang, and Jiajie Peng. Integrating sequence and network information to enhance protein-protein interaction prediction using graph convolutional networks. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1762–1768. IEEE, 2019.
88. Somaye Hashemifar, Behnam Neyshabur, Aly A Khan, and Jinbo Xu. Predicting protein–protein interactions through sequence-based deep learning. *Bioinformatics*, 34(17):i802–i810, 2018.
89. Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.
90. Marinka Zitnik, Monica Agrawal, and Jure Leskovec. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, 34(13):i457–i466, 2018.
91. Bharti Khemani, Shruti Patil, Ketan Kotecha, and Sudeep Tanwar. A review of graph neural networks: concepts, architectures, techniques, challenges, datasets, applications, and future directions. *Journal of Big Data*, 11(1):18, 2024.
92. Muhan Zhang and Yixin Chen. Link prediction based on graph neural networks. *Advances in neural information processing systems*, 31, 2018.
93. Jiahui Wu, Bo Liu, Jidong Zhang, Zhihan Wang, and Jianqiang Li. Dl-ppi: a method on prediction of sequenced protein–protein interaction based on deep learning. *BMC bioinformatics*, 24(1):473, 2023.
94. Kanchan Jha, Sriparna Saha, and Hiteshi Singh. Prediction of protein–protein interaction using graph neural networks. *Scientific Reports*, 12(1):8360, 2022.
95. Alex Fout, Jonathon Byrd, Basir Shariat, and Asa Ben-Hur. Protein interface prediction using graph convolutional networks. *Advances in neural information processing systems*, 30, 2017
96. Mayank Baranwal, Abram Magner, Jacob Saldinger, Emine S Turali-Emre, Paolo Elvati, Shivani Kozarekar, J Scott VanEpps, Nicholas A Kotov, Angela Violi, and Alfred O Hero. Struct2graph: A graph attention network for structure based predictions of protein–protein interactions. *BMC bioinformatics*, 23(1):370, 2022.
97. Boqiao Lai and Jinbo Xu. Accurate protein function prediction via graph attention networks with predicted structure information. *Briefings in Bioinformatics*, 23(1):bbab502, 2022.
98. Yanlei Kang, Arne Elofsson, Yunliang Jiang, Weihong Huang, Minzhe Yu, and Zhong Li. Aftgan: prediction of multi-type ppi based on attention free transformer and graph attention network. *Bioinformatics*, 39(2):btad052, 2023.
99. Siqi Chen, Anhong Zheng, Weichi Yu, and Chao Zhan. Dcmf-ppi: a protein-protein interaction predictor based on dynamic condition and multi-feature fusion. *BMC bioinformatics*, 26(1):1–27, 2025.
100. Youngjoo Seo, Micha¨el Defferrard, Pierre Vandergheynst, and Xavier Bresson. Structured sequence modeling with graph convolutional recurrent networks. In *International conference on neural information processing*, pages 362–373. Springer, 2018.
101. Aldo Pareja, Giacomo Domeniconi, Jie Chen, Tengfei Ma, Toyotaro Suzumura, Hiroki Kanezashi, Tim Kaler, Tao Schardl, and Charles Leiserson. Evolvegcn: Evolving graph convolutional networks for dynamic graphs. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 5363–5370, 2020.
102. Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael JL Townshend, and Ron Dror. Learning from protein structure with geometric vector perceptrons. *arXiv preprint arXiv:2009.01411*, 2020.
103. Raphael Townshend, Rishi Bedi, Patricia Suriana, and Ron Dror. End-to-end learning on 3d protein structure for interface prediction. *Advances in Neural Information Processing Systems*, 32, 2019
104. Wenting Liu and Jagath C Rajapakse. Fusing gene expressions and transitive protein-protein interactions for inference of gene regulatory networks. *BMC systems biology*, 13(Suppl 2):37, 2019.
105. Chuanyuan Wang, Shiyu Xu, Duanchen Sun, and Zhi-Ping Liu. Activeppi: quantifying protein– protein interaction network activity with markov random fields. *Bioinformatics*, 39(9):btad567, 2023.

 doi:10.20944/preprints202605.2116.v1

106. Nizamul Morshed, Madhu Chetty, and Nguyen Xuan Vinh. Fusgp: bayesian co-learning of gene regulatory networks and protein interaction networks. In *International Conference on Neural Information Processing*, pages 369–377. Springer, 2012.
107. Zeqian Li, Yijia Zhang, and Peixuan Zhou. Temporal protein complex identification based on dynamic heterogeneous protein information network representation learning. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 21(5):1154–1164, 2024.
108. Laura Megliorini et al. An integrative approach to infer dynamic protein-protein interaction networks: the case of plasmodium falciparum interactomes. 2025.
109. Peng Li, Shufang Guo, Chenghao Zhang, Mosharaf Md Parvej, and Jing Zhang. A construction method for a dynamic weighted protein network using multi-level embedding. *Applied Sciences*, 14(10):4090, 2024.
110. Yue He and Fei Zhu. Reconstruction of dynamic protein–protein interaction network via graph convolutional network. *Expert Systems with Applications*, 259:125140, 2025.
111. Y. A. Huang, Z. H. You, X. Gao, L. Wong, and L. Wang. Using weighted sparse representation model combined with discrete cosine transformation to predict protein–protein interactions from protein sequence. *BioMed Research International*, 2015:902198, 2015.
112. Z. H. You, K. C. Chan, and P. Hu. Predicting protein–protein interactions from primary protein sequences using a novel multi-scale local feature representation scheme and the random forest. *PLOS ONE*, 10(5):e0125811, 2015.
113. Bin Xu, Jihong Guan, Yang Wang, and Zewei Wang. Essential protein detection by random walk on weighted protein-protein interaction networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 16(2):377–387, 2017.
114. Y. B. Wang, Z. H. You, L. P. Li, Y. A. Huang, and H. C. Yi. Detection of interactions between proteins by using legendre moments descriptor to extract discriminatory information embedded in pssm. *Molecules*, 22(8):1366, 2017.
115. Y. Wang, Z. You, X. Li, X. Chen, T. Jiang, and J. Zhang. Pcvmzm: Using the probabilistic classification vector machines model combined with a zernike moments descriptor to predict protein–protein interactions from protein sequences. *International Journal of Molecular Sciences*, 18(5):1029–1041, 2017.
116. X. Y. Song, Z. H. Chen, X. Y. Sun, Z. H. You, L. P. Li, and Y. Zhao. An ensemble classifier with random projection for predicting protein–protein interactions using sequence and evolutionary information. *Applied Sciences*, 8(1):89–103, 2018.
117. Yan-Bin Wang, Zhu-Hong You, Xiao Li, Tong-Hai Jiang, Xing Chen, Xi Zhou, and Lei Wang. Predicting protein–protein interactions from protein sequences by a stacked sparse autoencoder deep neural network. *Molecular BioSystems*, 13(7):1336–1344, 2017.