

Structural and functional analysis of protein

Neetu Singh Yadav^{1,2}, Pawan Kumar³ and Indra Singh⁴

¹Department of Biochemical Engineering & Biotechnology, Indian Institute of Technology Delhi, New Delhi, India, ²Department of Chemical & Biomolecular Engineering, Clemson University, Clemson, SC, United States, ³Bioinformatics Center, National Institute of Immunology, New Delhi, India, ⁴School of Biotechnology, Banaras Hindu University, Varanasi, India

12.1 Protein preliminaries

Protein as a biomolecule is considered as a basic functional unit of the cellular function and its physical interactions with other molecules regulate the structural and functional aspects of cell biology. Once decoded from the mRNA sequence, the fate of the protein is determined and translocated at the desired location to perform its function. A world-famous experiment carried out by Anfinsen and coworkers proposed that a precious combination of amino acid sequence carries protein folding imprint required to fold the chain of amino acid into the specific spatial shape of a protein under specific cellular conditions (Sela, White, & Anfinsen, 1957). Once determined, protein structural information can provide mechanistic insights into molecular and cellular processes. Thus it is possible to relate the protein structural information to form a hypothesis concerning its function and further associate it with its amino acid sequence.

12.2 Growth of the protein structural database

As mentioned above, protein is a major functional and structural unit; however, due to the different level experimental or technical limitation, a large chunk of known protein sequences is deposited in the different databases with uncharacterized function. Among the known protein sequences, further challenges at the expression level or crystallization level, and the only subset of proteins are solved for its 3D structure by different techniques, such as crystallography (X-ray), NMR spectroscopy, and 3D electron cryo-microscopy (3DEM), and other structure-determination techniques. From the start of the Protein Data Bank in 1971 with just seven solved structures, now it has become the primary database with open-access web-resource for structural biology (Crystallography: Protein Data Bank, 1971). Currently the Protein Data Bank (PDB) database has more than 170,000 structural entries representing different biomolecules, such as protein, DNA, RNA, and also their complexes. As shown in Fig. 12.1, approximately 90% of 3D structures are solved using the crystallography technique while NMR spectroscopy and electron microscopy based structures account for ~9% and ~1%.

The importance of PDB repository data can be understood from the fact that more than 200 different life-sciences oriented data resources depend on this primary database (Burley et al., 2018). To maintain and manage the standard required at different levels structural entry process, such as deposition, bio-curation, validation of the deposited data, and further distribution of the structural data across the different PDB partners, worldwide PDB (wwPDB) was set up in 2003 (Berman, Henrick, & Nakamura, 2003) with the task to ensure a high-quality data availability with no limitation on usage (Berman, Kleywegt, Nakamura, & Markley, 2013). After the establishment of wwPDB, a unified system (OneDep) (Young et al., 2017) was put forward in 2014 to handle evolving requirements in coming decades as more and more large molecular assembly, large polymer, and small ligand chemical structures will come up with different level challenges. From time-to-time various decisions are taken since 2008, structure factor data are essential to submit with all X-ray crystal structures, and since 2010, for NMR-based structures, chemical shift information is necessary to supply. In the case of the Electron Microscopy Data Bank entries, an electron density map is mandatory to submit since 2016 (Smart et al., 2018).

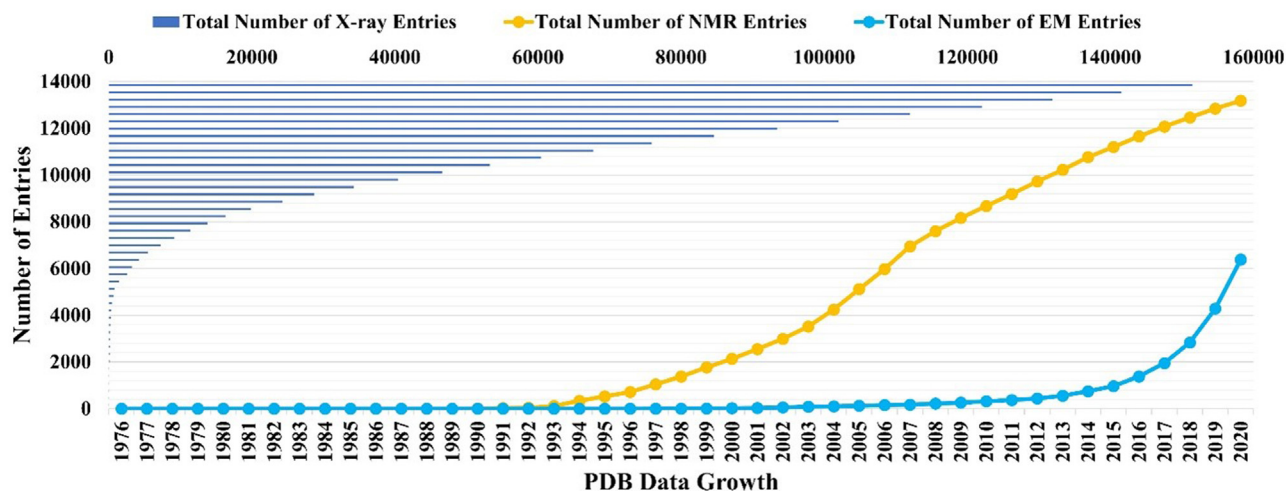


FIGURE 12.1 Growth of the 3D protein structural data in the Protein Data Bank database.

Since the last decade tremendous improvement in crystallography techniques has enabled the deposition of large amounts of crystal structure data in the global achievement of 3D biomolecular structure. High-throughput X-ray technique, such as Pan-Dataset Density Analysis, has tremendously increased the processing capability of the X-ray diffraction data with remarkably clear-cut detail (Pearce et al., 2017). The impact of this method can be understood from the fact that more than 160 crystal structures were submitted in the PDB database for different SARS-CoV-2 proteins to fight against the current Covid-19 pandemic. Submitted apo and fragment bound complex structures are utilized to understand the molecular interaction details necessary for high affinity novel inhibitors design for these target proteins.

12.3 Structural topology and fold classification scheme

Assessing the structural similarity is one of the approaches to establish the protein structural, functional, and evolutionary relationship with other related proteins, and also this similarity assessment allows one to understand the function and annotation of the uncharacterized protein structure (Deng, Zhong, Liu, Luo, & Liu, 2019; Konopka, Nebel, & Kotulska, 2012; Petrey et al., 2015; Zhang et al., 2012). Different schemes are proposed to relate the two protein structures (Khan, Panday, & Ghosh, 2018), and among them, the Class, Architecture, Topology, and Homology Protein Structure Database (CATH) and the Structural Classification of Proteins (SCOP) are the most approved approaches, which classify the protein structure at different hierarchy levels. SCOP (<http://scop.mrc-lmb.cam.ac.uk/>) that stands for Structural Classification of Protein classifies the protein structures in different hierarchy levels with the domain as a fundamental unit in experimentally determined structure (Fox, Brenner, & Chandonia, 2014). The SCOP that was first released in 1994 with all PDB entries (3019) of that time (Murzin, Brenner, Hubbard, & Chothia, 1995) embedded the notion of evolution by grouping the protein sequence not only based on similarity but also by characteristic evolutionary features. The SCOP comprises the following level species (distinct/original protein sequence); protein (similar set of sequences having similar function); family (proteins with similar sequences with typically distinct functions); and at last superfamily (bridging the evolutionary conserved structural features). Superfamilies are further classified as structural fold and classes (Fig. 12.2).

As with SCOP, CATH (Sillitoe et al., 2019) (<https://www.cathdb.info/>) is also an online resource that classifies the deposited protein structures in the four-level hierarchy, such as homologous superfamily, topology, architecture, and class based on the increasing structural and sequence similarity assessment using the sensitive tools, such as SSAP (Orengo & Taylor, 1996), HMMER3 (Mistry, Finn, Eddy, Bateman, & Punta, 2013), and PRC (Brandt & Heringa, 2009). At the first level, CATH tries to capture the domain superfamily information by analyzing over 95 million protein sequences (Lewis et al., 2018). Homologous superfamily assignments can recognize the evolutionary relationship among the proteins with the same structural core and different functions.

Apart from the structural classification approach, protein topological features are utilized to navigate the folding and scaffold profiles (Lewis et al., 2018; Sadowski & Taylor, 2010). Protein topology-based analysis also enabled us to analyze the protein modularity nature and associate this with protein function and structural diversity (Mills, Beuning, & Ondrechen, 2015; Slabinski et al., 2007). Topology-driven online resources, such as ProLego (Khan et al., 2018), were

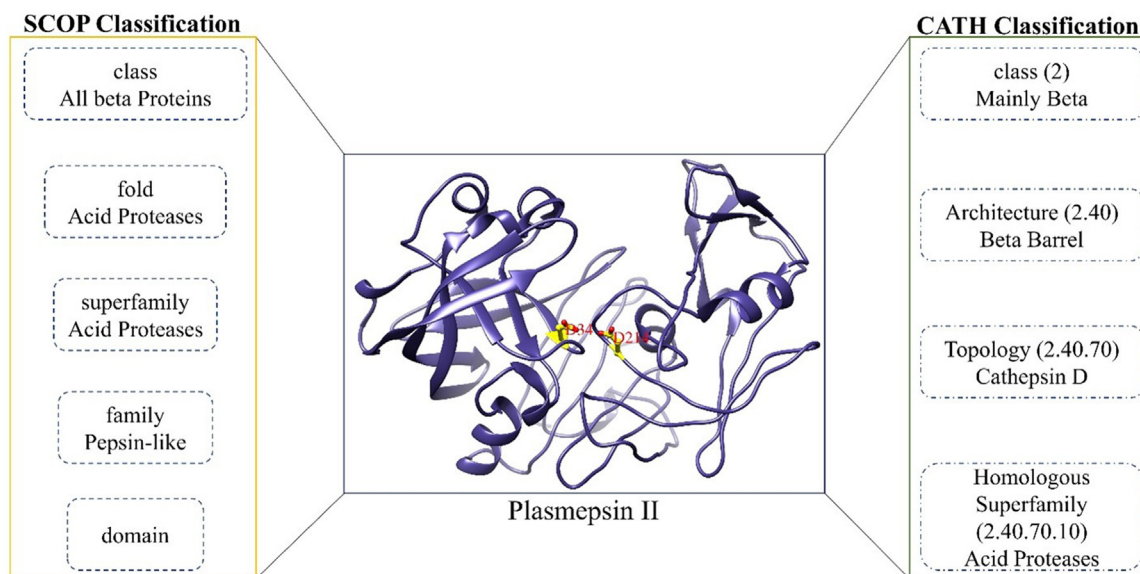


FIGURE 12.2 Structural classification (Structural Classification of Proteins and Class, Architecture, Topology, and Homology Protein Structure Database) for the acid protease of *Plasmodium falciparum* (Protein Data Bank ID: 2BJU).

developed to understand the topology and modularity nature in the available protein structural database. Given the protein sequence of choice, the program first built the secondary structure (SS) contact list, which is further utilized for adjacency matrix and corresponding 1D topology string generation encompassing contact and relative orientation of SS.

12.4 D-Structure quality assessment protocol

Once the 3D structure of proteins determined by any modern-day approach (Cavalli, Salvatella, Dobson, & Vendruscolo, 2007; Slabinski et al., 2007; Yip, Fischer, Paknia, Chari, & Stark, 2020), all deposited structures in the PDB undergo a validation process to address the quality benchmark of the deposited 3D coordinates (Smart et al., 2018). OneDep was established by the wwPDB partners to handle deposition, bio-curation, and validation tasks. The validation task includes the assessment of the experimental data along with the stereochemical and geometric fit of the atomic coordinates against experimental data (Shao et al., 2017). In 2014 the validation pipeline was integrated with the OneDep platform for the assessment of crystallography structures and in 2016 the validation pipeline for NMR and 3DEM structural entries was also started. Before the actual deposition of the structural data to the OneDep system, the depositor can perform the structural quality check using the wwPDB validation server (<https://validate.wwpdb.org>). A comprehensive structure validation report against each entry is available for users to check before actual scientific uses (<https://www.wwpdb.org/validation/validation-reports>). Along with a full validation report, a validation metric is also available summarizing the structural and geometric quality factors against all deposited structures. Fig. 12.3 shows the validation metric (slider image) for three different PDB IDs of higher (4DI8), intermediate (2HYU), and lower quality (2GUW) along with residues in purple color and sphere style showing poor fit electron density.

12.5 Protein 3D structure prediction

Predicting how proteins fold in nature in vivo is considered as the holy grail of proteomics and theoretical chemistry even today. With the advancement in the genomic sequence techniques, the number of known protein sequences has increased exponentially, while the rate of protein structure determination trails behind due to associated technical difficulties with structural biological experiments. By the end of 2020 the number of sequences deposited to the UniprotKB database is ~200 million (<https://www.uniprot.org/>) (Bairoch et al., 2005), whereas the corresponding protein structures deposited in RCSB-PDB databank only about ~170,000 (<https://www.rcsb.org/stats/growth/growth-released-structures>) (Berman et al., 2000). Thus developing an efficient computer-based method generating high-resolution three-dimensional structures of protein becomes the possible approach to fill the gap.

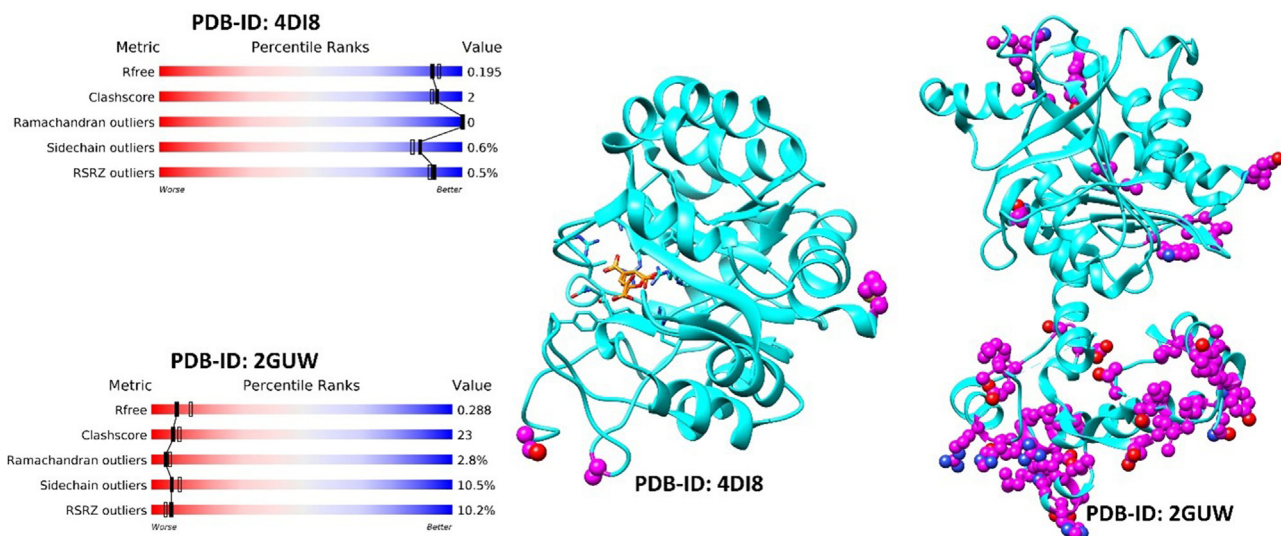


FIGURE 12.3 Structure validation metrics for two proteins (Protein Data Bank ID: 4DI8 and 2GUW) depicting the different geometric and structural factors to assess the structure accuracy. Left-side slider image displays the structure quality factors and deviation compared to all structures having similar resolution deposited in the PDB. Right-side figure shows the inaccuracy associated with the 3D coordinate of the residue compared with the available electron density map.

Depending upon whether the similar protein has been solved before experimentally, the Protein Structure Prediction (PSP) methods can be divided into two classes. If the query protein has a template of structures available in RCSB-PDB the task is relatively easier. High-resolution models of such sequences often build on the framework of available structures. This procedure is called template-based modeling (TMB) (Karplus, Barrett, & Hughey, 1998; Šali & Blundell, 1993; Söding, 2005). Although the TMB techniques can successfully generate a model, still it cannot answer the question of how and why a protein adopts a particular fold in nature. On the other hand, if structural homology (or protein templates) does not exist, and/or cannot identify, models have to build from scratch. The modeling of such a structure is known as *ab initio* modeling (Klepeis, Wei, Hecht, & Floudas, 2005; Liwo, Khalili, & Scheraga, 2005; Taylor et al., 2008), *de novo* modeling (Bradley, Malmström, et al., 2005; Bradley, Misura, & Baker, 2005), free modeling (FM) (Jauch, Yeo, Kolatkar, & Clarke, 2007; Kinch, Li, Monastyrskyy, Kryshtafovych, & Grishin, 2016), and physics-based modeling (Ołdziej et al., 2005). This offers a solution to the PSP problem and provides the physicochemical fundamentals of how proteins fold in biological solutions. Generally, in *ab initio* modeling, a conformational search is conducted under the guidance of a predesigned energy function. This results in a myriad of structural decoys (possible conformations) subsequently the final models are selected from them. The success of *ab initio* modeling confides on three factors; (1) description of accurate energy function which enables to define the native structure of the protein having the minimum energy (thermodynamically more stable) as compared to the other decoys; (2) an efficient search method helps in quickly identify the low energy states through conformational search; and (3) an algorithm that can select near-native models from the available decoy conformation. Despite the significant effort, and advancement of computational techniques, the *ab initio* prediction of 3D structures of protein from sequences without using the template is limited.

12.5.1 Energy functions

Energy function and the search procedure are intricately coupled to each other, so much so the decoupling procedure often results in the loss of power of validity of the modeling procedure. The energy function, depending upon the usage of the statistics of the existing 3D structures in the PDB, can be classified into two groups: (1) physics-based energy function and (2) knowledge-based energy function.

12.5.1.1 Physics-based energy function and its applications

In the physics-based *ab initio* modeling technique, the interaction between atoms is parametrized by employing quantum mechanics and Coulombic potential, along with the fundamental parameters (e.g., electron charge and Planck constant) (Hagler, Huler, & Lifson, 1974; Weiner et al., 1984). Although the promising role offered by the quantum

mechanics structure modeling; however, due to the extensive computational resource requirement, serious attempts are not reported to predict the structure using quantum mechanics. In the absence of quantum mechanical treatments, the starting point for the structure prediction is shifted toward the use of force fields. In the force-field, unlike quantum mechanics, atoms are considered as particles, interacting via a well-defined potential energy function. The rules governing interatomic interactions are derived by comparing FF (force field) data collected from the experimental and quantum calculation (Hagler et al., 1974; Weiner et al., 1984). The well-established examples of classical physics-based FFs are AMBER (Cornell et al., 1995; Duan & Kollman, 1998; Weiner et al., 1984), CHARMM (Brooks et al., 2009; Neria, Fischer, & Karplus, 1996), OPLS (Jorgensen, Chandrasekhar, Madura, Impey, & Klein, 1983), and GROMOS (van Gunsteren et al., 1996). These classical force fields contain terms to define the internal coordinate, such as a bond, angles, and torsion angles along nonbonded interactions. The significant difference among the different FF lies in the representation of atoms types and the corresponding interaction coefficients.

12.5.1.1.1 Coupling molecular dynamics simulations with physics-based potentials

To study protein folding problems, the aforementioned classical FFs are usually integrated with molecular dynamics simulations (MDS). Protein folding pathway funnel prediction via MDS provides detailed insights into the 3D-structure folding and folding funnel type. Duan and Kollman performed the very first simulation of villin headpiece subdomain (a 36-residue long protein) in an explicit water medium for around six months on high-performance computers (HPCs) of that time (Duan & Kollman, 1998), where their best model was showing 4.5 Å root mean square deviation (RMSD) from the crystal state. Now with the availability of Folding@Home, a worldwide distributed HPC, the folding process of villin was further analyzed by Pande et al. and achieved 1.7 Å RMSD in 300 μs of simulation length. Since then a significant success has been seen in the *ab initio* folding simulations using MD (Chowdhury, Lee, Xiong, & Duan, 2003; Ensign, Kasson, & Pande, 2007; Freddolino & Schulten, 2009), although all have required large computational facilities. Nonetheless, the advancement *ab initio* folding has also revealed several physics-based force fields hampering their general applicability (Best & Hummer, 2009; Best, Buchete, & Hummer, 2008; Freddolino, Liu, Gruebele, & Schulten, 2008).

Since the majority of given FFs are characterized by many local energy minima, which can easily trap the protein folding in a simulation, thus offer difficulties in identifying global minima. The flurry of FF development tried to mitigate these known limitations by generating new parameter sets applicable to fold a wide variety of protein structures consistently (Lindorff-Larsen et al., 2010; Mittal & Best, 2010; Piana, Lindorff-Larsen, & Shaw, 2011), leaving simulation time as the main barrier for *ab initio* MD folding simulations. Later on, with the progress and upgradation in computer hardware, this barrier began to crumble. The Anton supercomputer from the Shaw group has enabled the complete reversible folding pathway of selected functional protein up to ~100-amino-acid long in explicit solvent media (Lindorff-Larsen, Piana, Dror, & Shaw, 2011; Piana, Lindorff-Larsen, & Shaw, 2012, 2013; Piana, Klepeis, & Shaw, 2014). Following a different route, the GPU-enabled MD simulation packages have accelerated the *ab initio* folding simulations to reach up ~1 μs/day of simulation length for small proteins (10–100 residues) in implicit solvent (Nguyen, Maier, Huang, Perrone, & Simmerling, 2014).

12.5.1.1.2 Structure refinement

Another area, where physics-based MD simulations have contributed most is the structure refinement. Such simulation aims to build structures closer to the native state starting from low-resolution protein models, by refining the peptide backbone packing and local side-chain interactions. When the starting model is near to the native state and the intended conformational changes are small, the simulation time would be much less as compared to the *ab initio* protein natural state simulation. Among the early reports of structure-based refinement is of GCN4 leucine zipper coiled-coil dimer (33 residues long). The low-resolution coiled-coil dimer was assembled with Monte Carlo (MC) simulation a priori to the MD refinement. Further by employing the helical dihedral-angle restraints, in CHARMM Vieth, Koliński, Brooks, and Skolnick (1994) have successfully generated the refined structure of GCN4, in the TIP3P water model, having backbone RMSD < 1 Å. Subsequently, by using AMBER 5.0, Simons, Kooperberg, Huang, and Baker (1997) attempted to refine ~360 low-resolution models of 12 small proteins (<75 amino acids) generated by ROSETTA. However, no systematic structure improvement was reported by the authors (Lee, Tsai, Baker, & Kollman, 2001). Next, Fan and Mark using GROMACS 3.5 (Lindahl, Hess, & Van Der Spoel, 2001) tried to refine 60 ROSETTA models of 11 small proteins (<85 residues) in an explicit model and they reported that 11/60 models have shown improvement by 10% in RMSD, while 18/60 resulted in worse RMSD after refinement. Similar results were obtained by Chen and Brooks (2007) in CHARMM22 (MacKerell et al., 1998). They have shown that during the refinement of five CASP comparative

modeling targets, four out of five have received an RMSD value of 1 Å. In the reported work, they implied an implicit model, such as the generalized born approximation, to enhance the calculation speed. In addition, the spatial restraints determined from the starting models were also employed to further suggest the refinement in the structure (Chen & Brooks, 2007).

On the other hand, Zhang and coworkers have proposed the use of analogous fragments generated using the already available protein structures to model the physics-based FF for further enhancement. In the reported approach, the starting model was fragmented into many segments of 2–4 SS elements. Next, these fragments were aligned via the PDB structure library to identify analogous fragments (Zhang & Skolnick, 2004a, 2005b). Furthermore, a distant map constructed from similar structural elements is applied as a restraint to redesign the molecular dynamics of the energy funnel. This approach was validated on 181 benchmarking and 26 CASP proteins, and it was reported that models having TM score more than 0.5 can be pulled near to the original state with a relatively higher GDT-HA score.

Interestingly, Summa and Levitt (2007) by exploring various molecular mechanics potentials [AMBER99, OPLS-AA (Kaminski, Friesner, Tirado-Rives, & Jorgensen, 2001), GROMOS-96 (van Gunsteren et al., 1996), and ENCAD (Levitt, Hirshberg, Sharon, & Daggett, 1995)] for energy minimizing 75 proteins in vacuo have made a noteworthy observation. They reported that the knowledge-based atomic contact potentials have exceeded molecular mechanics-based contact potentials by driving all the proteins closer to their native states. In the case of molecular mechanics potentials, except for AMBER99, all others have moved the decoys further away from their respective native states. One of the possible explanations for the failure of MM potentials could be attributed to the vacuum simulation without salvation. Still, observation clearly demonstrated that the conjunction of knowledge-based potentials along with the physics-based force field is a better method for the refinement of protein structures.

The innate computational cost associated with the physics-based potential methods associated with MD simulation on the folding timescale surpasses the success of physics-based methods. However, adding the quick search approaches, for example, MC simulations and genetic algorithms (GA) integrated with classical physics-based function also reported encouraging results in this regard. A solely thermodynamic method, UNited RESidue developed by the Scheraga and coworkers (Liwo, Lee, Ripoll, Pillardy, & Scheraga, 1999; Liwo et al., 2005; Ołdziej et al., 2005), is the most faithful method and applied for different CASP targets. UNRES can also use the coarse-grained function with an optimization approach for enhanced protein conformational space search (Ołdziej et al., 2005). Such an approach was initially employed specifically for the prediction of the protein 3D model. UNRES is a highly reduced model where each protein residue is described by only two interacting regions: (1) united side chains and (2) united peptide group. This representation helps in the reduction of the number of atoms by 10. Thus enable to model polypeptide chains of >100 residue length, which in turn elevates the speed of UNRES simulation by ~1000–4000-fold, as compared to the all-atom simulation. The low energy models of UNRES can be described as all-atom functions by ECEPP/3 (Nemethy et al., 1992). It has successfully predicted a model for 95-residue long alpha-helical protein with 4.2 Å RMSD between it and other models. This has established clearly that ab initio-based methods can sometimes improve the 3D-protein models where the template-based methods do not reliably work.

ASTRO-FOLD, a hierarchical algorithm, is an additional instance of physics-based multi-stage modeling method (Klepeis & Floudas, 2003; Klepeis et al., 2005). The 3D structure of a protein from its amino acid chain is predicted by the combination of global and optimization framework. The important features of ASTO-FOLD can be understood as: (1) SS estimation by implementing a novel optimization algorithm. In this method, the secondary structure elements were predicted by the free energy function of aligning peptides and all other possible interactions between two hydrophobic amino acids. The free energy calculation terms consist of entropy, polarization, ionization, and binding site contributions of each protein fragments, (2) use of mixed integer linear program (MILP) for the beta-sheet topology prediction, (3) residue-to-residue interaction estimation by employing a distance-dependent FF and MILP construction, and (5) non-linear optimization for the dihedral clustering. This method was reported to the prediction of 102 amino acid peptides in a double-blind manner (Klepeis & Floudas, 2003; Klepeis et al., 2005).

12.5.1.2 Knowledge-based energy function and its application

The term knowledge-based potentials indicate a set of energy sets calculated by applying the statistical and regularities of the 3D structures available in the PDB. Skolnick categorizes the potentials into two types (Skolnick, 2006). The first covers generic and sequence-independent terms which includes H-bonding and protein backbone stiffness (Zhang, Kolinski, & Skolnick, 2003). The second contains protein-amino acid-dependent factors, for example, pairwise contact term (Skolnick, Godzik, Jaroszewski, & Kolinski, 1997), distance-associated contact term (Lu & Skolnick, 2001; Samudrala & Moult, 1998; Shen & Sali, 2006), and SS propensities (Zhang & Skolnick, 2005a, 2005b; Zhang et al.,

2003). Naturally, a variety of protein sequences, depending on their local and global environment, prefers either helical and/or extended beta secondary structures. Albeit the majority of knowledge-based FFs include SS propensities, yet no available force field can reproduce these subtlety properties successfully. Here, we illustrated the examples of methods that were designed based on the knowledge-based energy function.

12.5.1.2.1 ROSETTA

ROSETTA, an ab initio method for PSP, was pioneered by Bowie and Eisenberg. This method uses the assembly-based approach to arrange the small fragments extracted from the PDB to generate the model 3D-structures (Bowie & Eisenberg, 1994). This query sequence-based fragment alignment is guided by the similarity score, the SS prediction information, Ramachandran map probabilities distribution. Although the authors had tested the length of the different fragments, they eventually adopted the fixed 9-mer long fragment as the core of the building process while 3-mers fragments were used as refinement probes.

So far ROSETTA has shown success in FM targets in the CASP experiments. In recent ROSETTA models, authors have adopted a two-stage protocol in which reduced models are developed having conformations defined by C β backbone atoms. Furthermore, these reduced forms are refined by all-atom energy functions. A major success attributed to this protocol is the blind structure prediction of target T0281 (70 residues) from CASP6 at Rosetta@home computing network. Inspired by the success of ROSETTA, several groups started developing their own energy functions to mitigate the limited availability of the ROSETTA's energy function. The popular derivatives of ROSETTA are Simfold (Fujitsuka, Chikenji, & Takada, 2006) and Profesy (Lee, Kim, Joo, Kim, & Lee, 2004), which tried to incorporate the more detailed structural and functional parameters; however, minimal success has been achieved by these programs (Fujitsuka et al., 2006; Lee et al., 2004).

12.5.1.2.2 Iterative Threading ASSEMBLY Refinement (I-TASSER)

This program was developed by Zhang and Skolnick (2004a, 2004b) for PSP and structure-based function annotation. I-TASSER program developed 3D protein structures by employing a knowledge-based approach. In this approach, the first step is the threading of a protein query sequence, to determine the closest template structure from the PDB database. Subsequently, fragments covering more than five continuous residues are further considered to reassemble complete models, while the fragments that do not align are modeled via ab initio modeling (Zhang et al., 2003). Here, the fragment assembly is performed using MC simulations (Zhang, Kihara, & Skolnick, 2002). The energy terms applied in TASSER consist of predicted secondary structure propensities, backbone h-bonds pattern, short- and long-range structural interactions, and computed hydrophobic energy. The precise distribution of the weights for each term is further optimized with the help of decoy set analysis (Zhang et al., 2003).

Another successful derivative is I-TASSER (Roy, Kucukural, & Zhang, 2010; Yang et al., 2015). I-TASSER tries to refine the TASSER cluster by simulating in an iterative fragment assembly manner. Given a protein sequence, I-TASSER first builds the three-dimensional models of the protein by performing multiple threading alignments to identify evolutionary relatives using PSI-BLAST (Altschul et al., 1997). Then, the secondary structure is predicted by PSIPRED (Jones, 1999). Second, the continuous fragments obtained through threading alignments are assembled by structural conformations of the perfectly aligned sections (Fig. 12.4). To improve the performance, I-TASSER starts with a reduced search model. This approach tries to minimize the total number of conformations required for screening. Replica-Exchange Monte Carlo (REMC) simulation is used for the fragment assembly. The fragment assembly simulations are carried out by using the knowledge-based FF and have statistical expressions obtained from PDB for hydrogen bond, hydrophobicity, C α , and side-chain correlations.

Models from the selected centroid are further utilized for final model identification and further refinement by fragment assembly simulations. I-TASSER uses the spatial restraints derived from the first round of TASSER models, to filter out any clashes from first-round models (Fig. 12.5). Next, the TM-align program selects the template structures (Zhang & Skolnick, 2005a, 2005b) from PDB. From CASP6 to CASP11 experiments, I-TASSER maintains itself as one of the premier methods for PSP (Battey et al., 2007; Cozzetto et al., 2009; Mariani, Kiefer, Schmidt, Haas, & Schwede, 2011; Montelione, 2012). Further improvements and developed versions can be discussed in the literature (Wu & Zhang, 2008a, 2008b, 2010).

12.5.1.2.3 QUARK

Quark is one of the ab initio PSP and protein folding methods. QUARK models are created by employing a large fragment size of up to 20 residues to assemble the models utilizing both knowledge and physics-based energy expression

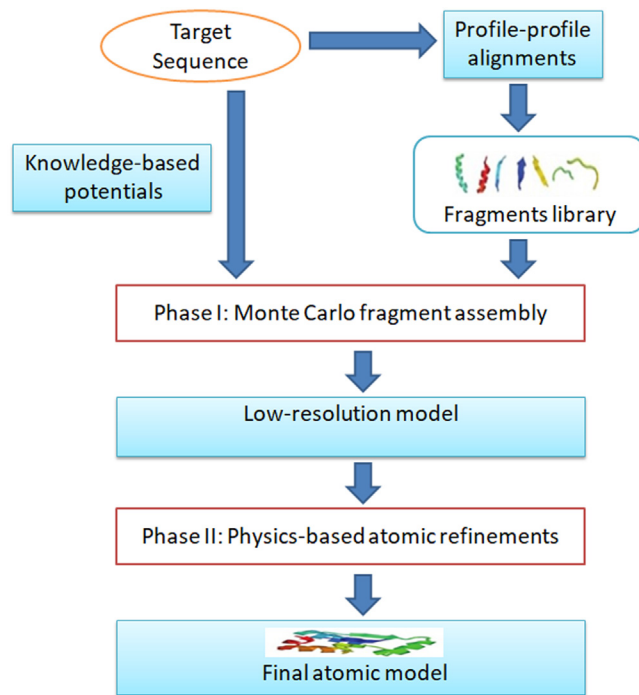


FIGURE 12.4 ROSETTA flowchart protocol. Fragments are first created from unrelated protein structures in the Protein Data Bank, which are used to assemble full-length models by simulated annealing simulations guided by a knowledge-based force field. In the second phase, selected models are refined at the atomic level using a physics-based potential.

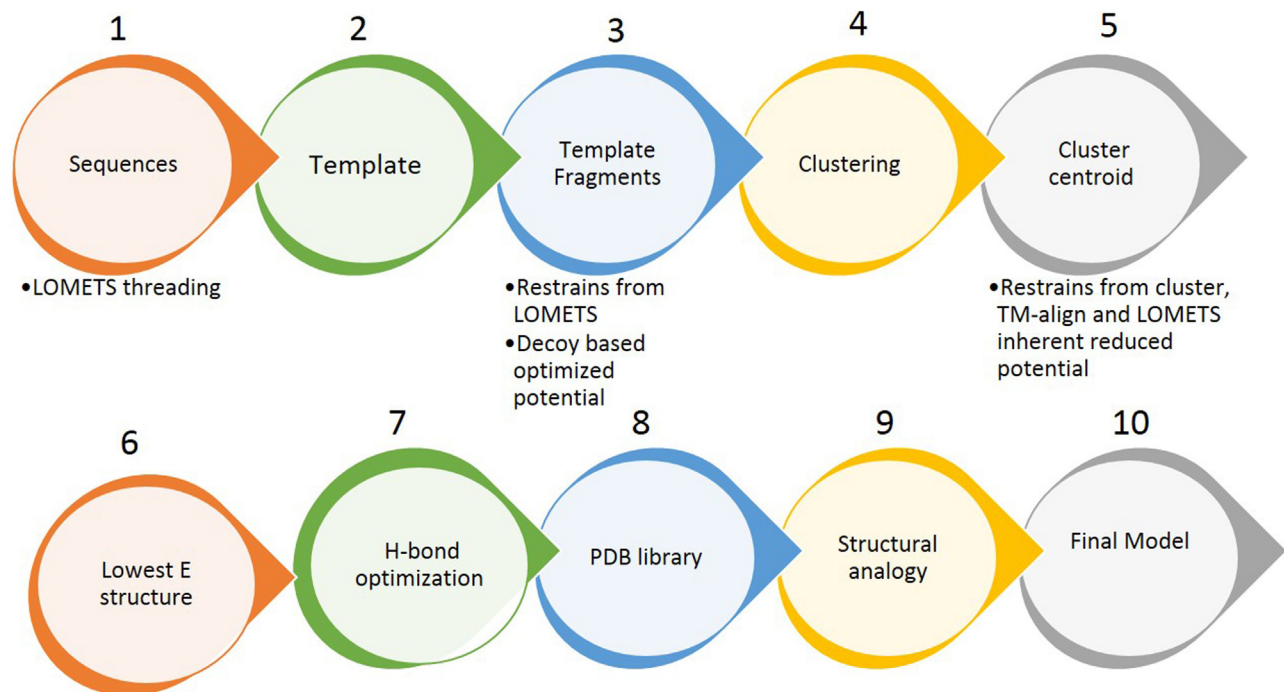


FIGURE 12.5 The general framework for the I-TASSER protocol in protein structure and function prediction. The *blue* and *red* colors in protein structures represent the N and C termini, respectively.

(Xu & Zhang, 2012). The focus during QUARK development was set to elaborate on the design of both the force field and search engine. QUARK enhances the force field development and search engine performance by representing the protein residue by full backbone atoms, and center of mass for the side chain.

For a given protein query sequence, the QUARK approach first predicts the structural features by employing the neural network (NN) model. Next, the global structural fold is identified by using REMC simulation by gathering small

fragments. In the ROSETTA, and I-TASSER methods, such fragments are generated by gapless threading alignment from template libraries. However, the fragment length of QUARK lies in between 1 and 20 residues, while the fragment length of ROSETTA and I-TASSER is 3–9-mer.

The REMC simulations and physics- and knowledge-based potentials contain a term for hydrogen bonds, van der Waals interactions, solvation, Coulomb, and backbone-torsion angles, along with internal coordinate information. Since its development, QUARK has also maintained itself among the best ab initio PSP method (Kinch et al., 2011, 2016; Tai, Bai, Taylor, & Lee, 2014). It was ranked number 1 server in FM in CASP9 and 10. QUARK accepts protein sequences <200 amino acids long; in addition, QUARK users cannot submit multiple jobs at once. Since no global template is required in QUARK modeling, hence, the QUARK server is suitable for proteins lacking homologous templates in the PDB library.

12.6 Machine learning in PSP

Machine learning (ML)-based PSP model generation now has very enriched literature information. Under the umbrella of ML-based modeling, many mathematical frameworks were developed in which NN, SVMs have been widely and successfully used (Heffernan et al., 2015; Shen & Bax, 2013). Recently the applicability of ML has been shifted to predict/model the 2D features, such as interresidue distance methods and residue–residue contact maps. Recognizing the contact maps is similar to 2D image recognition. The deep learning (DL) approach has shown tremendous success in the classification and interpretation of such maps (LeCun, Bengio, & Hinton, 2015). Fascinated by DL success, protein modelers have started to practice DL techniques to understand the order and patterns in the biological sequences and structural information available in different databases. Convolution neural network (CNN) has particularly demonstrated tremendous success in image analysis tasks (Jones & Kandathil, 2018; Krizhevsky, Sutskever, & Hinton, 2017; Liu, Palmedo, Ye, Berger, & Peng, 2018). Moreover, ML-based approaches are more crucial in template FM as shown by the top-performing CASP13 PSP methods.

12.6.1 Feature engineering and representation

ML features are independent input variables to a predictive model. Selecting and/or applying a suitable feature is a fundamental step in implementing ML algorithms. Features have multiple usages like in classification problems it allows the model to differentiate between one category of data and another (Table 12.1). In the case of regression, it helps in fitting a suitable function to the given input. a suitable set of features is called feature engineering, where a category of features serving one object is called as a feature vector. In N-dimensions the space corresponding to feature vectors is generally referred to as feature space. In the case of proteins, the feature space is composed of amino acid sequences, protein–protein interactions (PPI) (Kerepesi, Daróczy, Sturm, Vellai, & Benczúr, 2018; Lee, Tu, Deng, Sun, & Chen, 2006), and physicochemical properties (Gligorijević, Barot, & Bonneau, 2018; Wang, Zhang, Jia, Ren, & Yu, 2017). Motifs are specific amino acid patterns, associated with a specific biological function. Hence, the existence of a motif

TABLE 12.1 List of features applied for the functional classification of proteins.

Feature	Advantage	Disadvantage
Physicochemical properties	Simple	Fail to provide details about the protein
Protein/peptide sequence	Provides detailed information	Requires a conversion into numeric data for machine learning application
Protein–protein interaction network	Adjacent proteins are highly probable of sharing function	Robustness is experimental data dependent
Biomedical text	Provides detailed information	Results are sensitive toward the information of selected terms
Immunohistochemistry images	Provides a rich source of information and features	Needs extensive computational resources and data, particularly useful in the case of subcellular localization
Representation learning	No need for manual feature engineering and selection	Needs extensive computational resources and larger datasets

in a classification problem can be considered as a binary feature. Furthermore, several studies have used N-terminal targeting sequences as a feature (Höglund, Dönnies, Blum, Adolph, & Kohlbacher, 2006; Shatkay et al., 2007). Moreover, auto covariance, Moran autocorrelation, conjoint triad, and local description have been used as successful features in mining the interaction information of sequence (You, Lei, & Zhu, 2013).

Albeit ML requires numerical features to boost an appropriate model; however, the derivation of numerical features is simple for protein sequence, PPI networks, and physiochemical properties. Given a conversion to numerical formats, the use of text-based features is also possible. Furthermore, advancement in the natural language processing system has resulted in the exponential usage of biomedical literature-based features, for protein function prediction (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013; Verspoor, 2014).

12.6.2 Feature selection

Like other places, the usage of ML in biology also faces the problem of dimensionality. This indicates that for the available data, the feature space is larger than the sample space, thus resulting in the performance degradation. One way to mitigate this issue is by applying “dimensionality reduction,” which is to filter the feature space. In principle, one can utilize each combinatorial subset of feature and identify one; however, this method is suitable only for small-scaled data. Considering the issue, Molina et al., by applying three rules, has proposed the use of feature selection algorithms (FSAs) for the characterization of search problems in the hypothesis space. **The first rule concerns devising a common strategy to explore hypotheses. The second rule pertains to the selection of successor candidates, and the third is an evaluation of the successor candidate’s fitness.** This will give the flexibility to analyze hypotheses for further search processes (Molina, Belanche, & Nebot, 2002). Interested readers can find specific applications of these techniques in different biological problem analysis in the literature (Saeyns, Inza, & Larrañaga, 2007; Wang, Wang, & Chang, 2016). Moreover, FSAs are broadly grouped into three main classes: (1) wrapper method, (2) filter method, and (3) embedded method. Wrapper algorithm evaluates candidate feature subsets by applying a similar predictive model (SVM or Random Forest) on the selected feature after building a final classification model.

The filter method evaluates the feature subsets by applying a proxy method in place of the error rate given by the algorithm. This is computationally cheap and provides the fitness of the given feature set (Al-Shahib, Breitling, & Gilbert, 2005; Tang et al., 2018). Furthermore, the embedded method selects the features as part of the model building process. The two most common examples of an embedded method are LASSO and random forest. In the LASSO method along with building a linear model, regression coefficients are penalized with the L1 penalty. This penalty reduces many coefficients to zero. Features having coefficient values greater than zero are selected while the rest are discarded. In the Random Forest method, features are first ranked and then selected by applying the best-first forward search application (Lou et al., 2014).

12.6.3 ML algorithms

In biology, ML algorithms can be widely applied for protein function prediction, where each input sequence needs to map to a discrete function/output. The models predicting the function of protein are generally trained by implementing supervised, unsupervised, or semisupervised learning approach. Where supervised learning methods have the labeled training datasets. Unsupervised learning is related to finding, clusters, and structure in the data. Semisupervised learning lies in between supervised and unsupervised learning methods, where the training sets usually consist of a combination of off-labeled data and unlabeled data. Among the several ML algorithms, the simplest is the logistic regression. You, Huang, and Zhu (2018) have used logistic regression in the prediction of cellular components, molecular function, and biological process derived from MEDLINE biomedical literature databases. Furthermore, a kernel logistic regression model based upon the PPI networks was developed (Lee et al., 2006; Ni et al., 2009). Naive Bayes classification is based upon Bayes theorem, based upon the assumption that given a class variable all features are independent of each other. It has been used to predict the PPI network (You et al., 2013) and gene functions (Fabris & Freitas, 2014).

As stated, earlier SVM is the supervised learning method, generally successful in higher-dimensional space. It maximizes the separation between points belonging to different classes in some N-dimensional space and therefore determines maximum-margin hyperplanes. In case, when the data is not linearly separable in the associated feature space, for easier separation a mapping is needed to higher-dimensional space. This functionality is usually achieved by applying kernel function, for example, Gaussian or polynomial function. Furthermore, the use of SVM in protein function prediction can be found in references (Badal, Kundrotas, & Vakser, 2018; Cai, Han, Ji, Chen, & Chen, 2003; Cai, Wang, Sun, & Chen, 2003). The k-nearest neighbors (kNN) is one of the simplest algorithms used in linear regression

and classification. It is a nonparametric method that classifies a given data based upon the similarity measure, that is, the closest number of 15 points (k) is associated feature space (Hu et al., 2011; Makrodimitris, van Ham, & Reinders, 2019; Wong & Shatkay, 2013).

Ensemble methods require plenty of base models to generate a suitable model. This method can be divided into two categories; sequential and parallel. In sequential methods, boosting is applied for building an ensemble, where each model is trained on the same data set, but weights are adjusted on each data point according to the error of the last prediction. The AdaBoost (Freund & Schapire, 1997) and Gradient Boosting (Friedman, 2001) falls under this category. Furthermore, the XGBoost algorithm (Chen & Guestrin, 2016), a scalable tree boosting system, is used for the classification of aging and nonaging human proteins (Kerepesi et al., 2018). A decision tree is another simplest ML model, where each leaf node designates the output or decision of the model after traversing a specific route/path through the tree's branches (Blockeel, Schietgat, Struyf, Džeroski, & Clare, 2006; Vens, Struyf, Schietgat, Džeroski, & Blockeel, 2008; Yang & Yang, 2006).

A NN model consists of a sequence of interconnected layers known as neurons and/or perception. In a classification problem, the output from the network is designed by one-hot encoding, where the categorical variables are designed by a binary vector. Initially, the vector is the string of zero which is assigned to one at the index of categorical value. The output of the given neuron/perceptron is computed using an activation-function. The weights generated by the previous neuronal layers serve as input to these activation functions. General activation functions are tanh, sigmoid, and ReLU (rectified linear unit). The model training aims to provide suitable weight for a given input such that a correct output can be obtained. One way to deliver a better output the typical NN has multiple intermediate hidden layers. The neural response is another algorithm that mimics the function of the brain. This algorithm simulates the neuronal action/response of the visual cortex. The neural response, build upon the similarity of different protein sequences, defines a distance metric, which is further used to predict the protein function (Li et al., 2016). Driven by the fact that one protein can perform multiple functions, which may be further categorized into subfunction, authors have performed hierarchical multilabel classification using a local multilayer perceptron (Cerri, Barros, de Carvalho, & Jin, 2016). Since a larger number of output labels hinder the performance of ML algorithms, hence, an ensemble of hundreds of NNs, each having 100 output, trained to speculate the protein function (Nievola, Paraiso, & Freitas, 2015). Later on, a hierarchical NN, having the inherent hierarchical nature, was also developed and tested (Clark & Radivojac, 2011).

DL is another approach in the ML; it mimics the working of the human brain in making decisions, processing data, detecting objects, and speech recognition. It is an example of unsupervised learning from unstructured and unlabeled data. In DL, unlike typical ML procedure predicting continuous or discrete-valued output, it is mainly focused on learning data representation, also called feature learning. Instead of traditional feature searching and engineering, this allows to automatically discover the required features. Like NNs, DL architecture is associated with several hidden layers. With the advancement in GPU, the training of DL models is becoming feasible; however, the only caveat in using DL is the availability of large amounts of data. A big dataset is required for the estimation of large parameters of DL (LeCun et al., 2015).

Fa et al. predicted the human protein functions by using Multi-Task Deep Neural Network models (MT-DNN) (Fa, Cozzetto, Wan, & Jones, 2018). MT-DNN can not only efficiently leverage the cross talk in bigger datasets but also maximizes the regularization effect. The studies have shown that the performance of MT-DNN is higher than alternative ML methods, such as FFPred and BLAST (Fa et al., 2018). A deep semantic text representation (DeepText2GO) was used for improving large-scale protein function prediction for biomedical literature (You et al., 2018). Furthermore, multifunctional enzyme function prediction with multilabeled DL is studied in reference (Zou, Tian, Gao, & Li, 2019).

12.6.4 ML models' implementation and evaluation

Owing to the increased demand for ML, several platforms have been designed across different disciplines. Scikit-Learn is the most commonly used python-based framework (Pedregosa et al., 2011). Although several packages in R and MATLAB have also been developed simultaneously yet, from 2017 by the IEEE symposium, Scikit-Learn has rated as the most commonly used package (Cass, 2018). Considering the computationally exhaustive character of DL, several GPU, cluster-based libraries, and frameworks, such as TensorFlow (Abadi et al., 2016), Keras (Chollet, 2017), and PyTorch (Adam et al., 2017), have been developed. With the increasing amount of data availability, the GPU-enabled DL algorithms have gained more attention toward the prediction of protein function (Kulmanov, Khan, & Hoehndorf, 2018; Miranda & Hu, 2018; Nauman, Rehman, Politano, & Benso, 2019).

ML models by nature are generally adapted to achieve an acceptable performance by utilizing appropriate hyperparameters. These hyperparameters are usually set before the actual training process as opposed to the parameters

TABLE 12.2 Summary of generally used metrics applied for protein function prediction.

Metric	Advantage	Disadvantage
Accuracy	Distinguish the correctly labeled samples from all the samples	Provide misleading information particularly in case of imbalance
Precision	Can differentiate how many samples labeled as a class of interest (COI) truly belongs to COI?	Does not take a false negative into account
Recall	Can correctly predict samples belonging to COI among all the samples	Does not take false positive into account
F1 score	Useful in cases of class imbalance	Less initiative as compared to other metrics
AUROC	The score is independent of the threshold set,	Provides misleading information particularly in case of imbalance
F_{\max}	Considers predictions across the full spectrum from high to low sensitivity	Penalizes specific predictions

learned during the model training. This selection allows the model to generalize and perform better on unseen data. Optimizers, such as Adam (Kingma & Ba, 2014), RMSprop (Tieleman & Hinton, 2012), or stochastic gradient descent, are examples of few general hyperparameters. Particularly, the activation function or the number of neurons in each layer in the NN is hyperparameters. The performance of NNs can be enhanced by training the model with many layers. The widely accepted approach is to search for the space of hyperparameters.

Protein function prediction is considered as a classification problem, where the parameter set used to evaluate the performance of ML models generally included accuracy, sensitivity, specificity (recall), precision, and F1 score. F1 score handles class imbalance better than accuracy. On the other hand, ROC (receiver operating characteristic) provides better visualization and model performance. ROC curves plot the rate of true positive (TP) as the function of FP (false positive) and show the relationship between sensitivity and specificity for every possible cut-off. The larger area under ROC indicates better performance, as it indicates that a higher TP rate is obtained for a similar FP rate. A similar curve termed as “precision-recall” (PR) can also be obtained from the Area Under PR curve (AUPR). A summary of commonly used metrics is listed in Table 12.2.

Given, the variety of domains, certain algorithms and models have proven to learn input-output mappings more effectively than others. However, it is important to train various ML models and compare their performance. Recently, a study has compared the performance of logistic regression, Naive Bayes, SVM, decision tree and NN to evaluate the suitability of using dissimilarity representations. Where SVM was giving better results in terms of F1 score and AUC parameters (De Santis, Martino, Rizzi, & Mascioli, 2018). SVM and kNN algorithms were trained on sequence motifs for enzyme classification (Ben-Hur & Brutlag, 2006). In the reference You et al. (2013), a comparison was made between SVM and an extreme machine-learning algorithm to predict the PPI network from protein sequences. Similarly, in another study sequence-based prediction of DNA-binding protein was predicted by hybrid feature selection using Gaussian Naive Bayes trained together with a decision tree, random forest, logistic regression, kNN, and SVM with both polynomial and RBF kernels (Lou et al., 2014).

12.7 Conclusion

Albeit algorithms for protein structure or function, the prediction has grown yet many challenges remain. Despite these provocations, we believe that protein structure prediction and design methods will continue to develop. We anticipate that the advancement in artificial intelligence may provide us the solution to the fifty-year old grand challenge of biology. It would be far exciting to speculate the progress of ML and pattern recognition in the field of protein design. Recently, an AI-based method AlphaFold2 demonstrates the stunning potential of computer aided structure prediction. Just as 50 years back Anfinsen laid out a challenging problem to the scientific community. The progress in bioinformatics methods and techniques announced today give use a bright hope and confidence that the computational methods will become one of humanity’s valuable asset in expending the frontiers of scientific knowledge.

Conflict of interest

The authors declare that they have no conflict of interest.

References

- Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... Isard, M. (2016). *Tensorflow: A system for large-scale machine learning*. . 12th USENIX symposium on operating systems design and implementation (OSDI) (Vol. 16, pp. 265–283). USENIX.
- Adam, P., Sam, G., Soumith, C., Gregory, C., Edward, Y., Zachary, D., ... Adam, L. (2017). Automatic differentiation in PyTorch. In *Proceedings of neural information processing systems*.
- Al-Shahib, A., Breitling, R., & Gilbert, D. (2005). FrankSum: New feature selection method for protein function prediction. *International Journal of Neural Systems*, 15(04), 259–275.
- Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, 25(17), 3389–3402.
- Badal, V. D., Kundrotas, P. J., & Vakser, I. A. (2018). Natural language processing in text mining for structural modeling of protein complexes. *BMC Bioinformatics*, 19(1), 84.
- Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., ... Magrane, M. (2005). The universal protein resource (UniProt). *Nucleic Acids Research*, 33 (Suppl. 1), D154–D159.
- Battey, J. N., Kopp, J., Bordoli, L., Read, R. J., Clarke, N. D., & Schwede, T. (2007). Automated server predictions in CASP7. *Proteins: Structure, Function, and Bioinformatics*, 69(S8), 68–82.
- Ben-Hur, A., & Brutlag, D. (2006). *Sequence motifs: Highly predictive features of protein function. Feature extraction* (pp. 625–645). Springer.
- Berman, H., Henrick, K., & Nakamura, H. (2003). Announcing the worldwide Protein Data Bank. *Nature Structural & Molecular Biology*, 10(12), 980. Available from <https://doi.org/10.1038/nsb1203-980>.
- Berman, H. M., Kleywegt, G. J., Nakamura, H., & Markley, J. L. (2013). The future of the Protein Data Bank. *Biopolymers*, 99(3), 218–222. Available from <https://doi.org/10.1002/bip.22132>.
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., ... Bourne, P. E. (2000). The protein data bank. *Nucleic Acids Research*, 28(1), 235–242.
- Best, R. B., Buchete, N.-V., & Hummer, G. (2008). Are current molecular dynamics force fields too helical? *Biophysical Journal*, 95(1), L07–L09.
- Best, R. B., & Hummer, G. (2009). Optimized molecular dynamics force fields applied to the helix – coil transition of polypeptides. *The Journal of Physical Chemistry B*, 113(26), 9004–9015.
- Blockeel, H., Schietgat, L., Struyf, J., Džeroski, S., & Clare, A. (2006). Decision trees for hierarchical multilabel classification: A case study in functional genomics. In *European conference on principles of data mining and knowledge discovery* (pp. 18–29). Springer.
- Bowie, J. U., & Eisenberg, D. (1994). An evolutionary approach to folding small alpha-helical proteins that uses sequence information and an empirical guiding fitness function. *Proceedings of the National Academy of Sciences of the United States of America*, 91(10), 4436–4440.
- Bradley, P., Malmström, L., Qian, B., Schonbrun, J., Chivian, D., Kim, D. E., ... Baker, D. (2005). Free modeling with Rosetta in CASP6. *Proteins: Structure, Function, and Bioinformatics*, 61(S7), 128–134.
- Bradley, P., Misura, K. M., & Baker, D. (2005). Toward high-resolution de novo structure prediction for small proteins. *Science (New York, N.Y.)*, 309(5742), 1868–1871.
- Brandt, B. W., & Heringa, J. (2009). webPRC: The Profile Comparer for alignment-based searching of public domain databases. *Nucleic Acids Research*, 37(Web Server issue), W48–W52. Available from <https://doi.org/10.1093/nar/gkp279>.
- Brooks, B. R., Brooks, C. L., 3rd, Mackerell, A. D., Jr, Nilsson, L., Petrella, R. J., Roux, B., ... Venable, R. M., ... Yang, W., York, D. M., & Karplus, M. (2009). CHARMM: The biomolecular simulation program. *Journal of Computational Chemistry*, 30(10), 1545–1614. Available from <https://doi.org/10.1002/jcc.21287>.
- Burley, S. K., Berman, H. M., Christie, C., Duarte, J. M., Feng, Z., Westbrook, J., ... Zardecki, C. (2018). RCSB Protein Data Bank: Sustaining a living digital data resource that enables breakthroughs in scientific research and biomedical education. *Protein Science*, 27(1), 316–330. Available from <https://doi.org/10.1002/pro.3331>.
- Cai, C., Han, L., Ji, Z. L., Chen, X., & Chen, Y. Z. (2003). SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Research*, 31(13), 3692–3697.
- Cai, C., Wang, W., Sun, L., & Chen, Y. (2003). Protein function classification via support vector machine approach. *Mathematical Biosciences*, 185(2), 111–122.
- Cass, S. (2018). The 2017 top programming languages. *IEEE Spectrum*, 31.
- Cavalli, A., Salvatella, X., Dobson, C. M., & Vendruscolo, M. (2007). Protein structure determination from NMR chemical shifts. *Proceedings of the National Academy of Sciences of the United States of America*, 104(23), 9615–9620. Available from <https://doi.org/10.1073/pnas.0610313104>.
- Cerri, R., Barros, R. C., de Carvalho, A. C., & Jin, Y. (2016). Reduction strategies for hierarchical multi-label classification in protein function prediction. *BMC Bioinformatics*, 17(1), 373.
- Chen, J., & Brooks, C. L., III (2007). Can molecular dynamics simulations provide high-resolution refinement of protein structure? *Proteins: Structure, Function, and Bioinformatics*, 67(4), 922–930.

- Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794).
- Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1251–1258).
- Chowdhury, S., Lee, M. C., Xiong, G., & Duan, Y. (2003). Ab initio folding simulation of the Trp-cage mini-protein approaches NMR resolution. *Journal of Molecular Biology*, 327(3), 711–717.
- Clark, W. T., & Radivojac, P. (2011). Analysis of protein function and its prediction from amino acid sequence. *Proteins: Structure, Function, and Bioinformatics*, 79(7), 2086–2096.
- Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., . . . Kollman, P. A. (1995). A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society*, 117(19), 5179–5197.
- Cozzetto, D., Kryshchuk, A., Fidelis, K., Moulton, J., Rost, B., & Tramontano, A. (2009). Evaluation of template-based models in CASP8 with standard measures. *Proteins: Structure, Function, and Bioinformatics*, 77(S9), 18–28.
- Crystallography: Protein Data Bank. (1971). *Nature: New Biology*, 233(42), 223. Available from <https://doi.org/10.1038/newbio233223b0>.
- De Santis, E., Martino, A., Rizzi, A., & Mascioli, F. M. F. (2018). Dissimilarity space representations and automatic feature selection for protein function prediction. In *2018 international joint conference on neural networks (IJCNN)* (pp. 1–8). IEEE.
- Deng, L., Zhong, G., Liu, C., Luo, J., & Liu, H. (2019). MADOKA: An ultra-fast approach for large-scale protein structure similarity searching. *BMC Bioinformatics*, 20(Suppl. 19), 662. Available from <https://doi.org/10.1186/s12859-019-3235-1>.
- Duan, Y., & Kollman, P. A. (1998). Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science (New York, N.Y.)*, 282(5389), 740–744.
- Ensign, D. L., Kasson, P. M., & Pande, V. S. (2007). Heterogeneity even at the speed limit of folding: Large-scale molecular dynamics study of a fast-folding variant of the villin headpiece. *Journal of Molecular Biology*, 374(3), 806–816.
- Fa, R., Cozzetto, D., Wan, C., & Jones, D. T. (2018). Predicting human protein function with multi-task deep neural networks. *PLoS One*, 13(6), e0198216.
- Fabris, F., & Freitas, A. A. (2014). *An efficient algorithm for hierarchical classification of protein and gene functions*. In *25th international workshop on database and expert systems applications* (pp. 64–68). IEEE.
- Fox, N. K., Brenner, S. E., & Chandonia, J.-M. (2014). SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Research*, 42(D1), D304–D309.
- Freddolino, P. L., Liu, F., Gruebele, M., & Schulten, K. (2008). Ten-microsecond molecular dynamics simulation of a fast-folding WW domain. *Biophysical Journal*, 94(10), L75–L77.
- Freddolino, P. L., & Schulten, K. (2009). Common structural transitions in explicit-solvent simulations of villin headpiece folding. *Biophysical Journal*, 97(8), 2338–2347.
- Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1), 119–139.
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 1189–1232.
- Fujitsuka, Y., Chikenji, G., & Takada, S. (2006). SimFold energy function for de novo protein structure prediction: Consensus with Rosetta. *Proteins: Structure, Function, and Bioinformatics*, 62(2), 381–398.
- Gligorijević, V., Barot, M., & Bonneau, R. (2018). deepNF: Deep network fusion for protein function prediction. *Bioinformatics (Oxford, England)*, 34(22), 3873–3881.
- Hagler, A., Huler, E., & Lifson, S. (1974). Energy functions for peptides and proteins I. Derivation of a consistent force field including the hydrogen bond from amide crystals. *Journal of the American Chemical Society*, 96(17), 5319–5327.
- Heffernan, R., Paliwal, K., Lyons, J., Dehzangi, A., Sharma, A., Wang, J., . . . Zhou, Y. (2015). Improving prediction of secondary structure, local backbone angles and solvent accessible surface area of proteins by iterative deep learning. *Scientific Reports*, 5(1), 1–11.
- Höglund, A., Dönnies, P., Blum, T., Adolph, H.-W., & Kohlbacher, O. (2006). MultiLoc: Prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition. *Bioinformatics (Oxford, England)*, 22(10), 1158–1165.
- Hu, L., Huang, T., Shi, X., Lu, W.-C., Cai, Y.-D., & Chou, K.-C. (2011). Predicting functions of proteins in mouse based on weighted protein-protein interaction network and protein hybrid properties. *PLoS One*, 6(1), e14556.
- Jauch, R., Yeo, H. C., Kolatkar, P. R., & Clarke, N. D. (2007). Assessment of CASP7 structure predictions for template free targets. *Proteins: Structure, Function, and Bioinformatics*, 69(S8), 57–67.
- Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology*, 292(2), 195–202.
- Jones, D. T., & Kandathil, S. M. (2018). High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features. *Bioinformatics (Oxford, England)*, 34(19), 3308–3315.
- Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., & Klein, M. L. (1983). Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79(2), 926–935. Available from <https://doi.org/10.1063/1.445869>.
- Kaminski, G. A., Friesner, R. A., Tirado-Rives, J., & Jorgensen, W. L. (2001). Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *The Journal of Physical Chemistry B*, 105(28), 6474–6487.
- Karplus, K., Barrett, C., & Hughey, R. (1998). Hidden Markov models for detecting remote protein homologies. *Bioinformatics (Oxford, England)*, 14(10), 846–856.

- Kerepesi, C., Daróczy, B., Sturm, Á., Vellai, T., & Benczúr, A. (2018). Prediction and characterization of human ageing-related proteins by using machine learning. *Scientific Reports*, *8*(1), 4094.
- Khan, T., Panday, S. K., & Ghosh, I. (2018). ProLego: Tool for extracting and visualizing topological modules in protein structures. *BMC Bioinformatics*, *19*(1), 167. Available from <https://doi.org/10.1186/s12859-018-2171-9>.
- Kinch, L., Yong Shi, S., Cong, Q., Cheng, H., Liao, Y., & Grishin, N. V. (2011). CASP9 assessment of free modeling target predictions. *Proteins: Structure, Function, and Bioinformatics*, *79*(S10), 59–73.
- Kinch, L. N., Li, W., Monastyrskyy, B., Kryshtafovych, A., & Grishin, N. V. (2016). Evaluation of free modeling targets in CASP11 and ROLL. *Proteins: Structure, Function, and Bioinformatics*, *84*, 51–66.
- Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.
- Klepeis, J. L., & Floudas, C. A. (2003). ASTRO-FOLD: A combinatorial and global optimization framework for ab initio prediction of three-dimensional structures of proteins from the amino acid sequence. *Biophysical Journal*, *85*(4), 2119–2146.
- Klepeis, J. L., Wei, Y., Hecht, M. H., & Floudas, C. A. (2005). Ab initio prediction of the three-dimensional structure of a de novo designed protein: A double-blind case study. *Proteins: Structure, Function, and Bioinformatics*, *58*(3), 560–570.
- Konopka, B. M., Nebel, J. C., & Kotulska, M. (2012). Quality assessment of protein model-structures based on structural and functional similarities. *BMC Bioinformatics*, *13*, 242. Available from <https://doi.org/10.1186/1471-2105-13-242>.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90.
- Kulmanov, M., Khan, M. A., & Hoehndorf, R. (2018). DeepGO: Predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics (Oxford, England)*, *34*(4), 660–668.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.
- Lee, H., Tu, Z., Deng, M., Sun, F., & Chen, T. (2006). Diffusion kernel-based logistic regression models for protein function prediction. *OmicS: A Journal of Integrative Biology*, *10*(1), 40–55.
- Lee, J., Kim, S. Y., Joo, K., Kim, I., & Lee, J. (2004). Prediction of protein tertiary structure using PROFESY, a novel method based on fragment assembly and conformational space annealing. *Proteins: Structure, Function, and Bioinformatics*, *56*(4), 704–714.
- Lee, M. R., Tsai, J., Baker, D., & Kollman, P. A. (2001). Molecular dynamics in the endgame of protein structure prediction. *Journal of Molecular Biology*, *313*(2), 417–430.
- Levitt, M., Hirshberg, M., Sharon, R., & Daggett, V. (1995). Potential energy function and parameters for simulations of the molecular dynamics of proteins and nucleic acids in solution. *Computer Physics Communications*, *91*(1), 215–231.
- Lewis, T. E., Sillitoe, I., Dawson, N., Lam, S. D., Clarke, T., Lee, D., . . . Lees, J. (2018). Gene3D: Extensive prediction of globular domains in proteins. *Nucleic Acids Research*, *46*(D1), D435–D439. Available from <https://doi.org/10.1093/nar/gkx1069>.
- Li, Y. H., Xu, J. Y., Tao, L., Li, X. F., Li, S., Zeng, X., . . . Zhang, C. (2016). SVM-Prot 2016: A web-server for machine learning prediction of protein functional families from sequence irrespective of similarity. *PLoS One*, *11*(8), e0155290.
- Lindahl, E., Hess, B., & Van Der Spoel, D. (2001). GROMACS 3.0: A package for molecular simulation and trajectory analysis. *Molecular Modeling Annual*, *7*(8), 306–317.
- Lindorff-Larsen, K., Piana, S., Dror, R. O., & Shaw, D. E. (2011). How fast-folding proteins fold. *Science (New York, N.Y.)*, *334*(6055), 517–520.
- Lindorff-Larsen, K., Piana, S., Palmo, K., Maragakis, P., Klepeis, J. L., Dror, R. O., & Shaw, D. E. (2010). Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Structure, Function, and Bioinformatics*, *78*(8), 1950–1958.
- Liu, Y., Palmedo, P., Ye, Q., Berger, B., & Peng, J. (2018). Enhancing evolutionary couplings with deep convolutional neural networks. *Cell Systems*, *6*(1), 65–74, e63.
- Liwo, A., Khalili, M., & Scheraga, H. A. (2005). Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(7), 2362–2367.
- Liwo, A., Lee, J., Ripoll, D. R., Pillardy, J., & Scheraga, H. A. (1999). Protein structure prediction by global optimization of a potential energy function. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(10), 5482–5485.
- Lou, W., Wang, X., Chen, F., Chen, Y., Jiang, B., & Zhang, H. (2014). Sequence based prediction of DNA-binding proteins based on hybrid feature selection using random forest and Gaussian naive Bayes. *PLoS One*, *9*(1), e86703.
- Lu, H., & Skolnick, J. (2001). A distance-dependent atomic knowledge-based potential for improved protein structure selection. *Proteins: Structure, Function, and Bioinformatics*, *44*(3), 223–232.
- MacKerell, A. D., Jr, Bashford, D., Bellott, M., Dunbrack, R. L., Jr, Evanseck, J. D., Field, M. J., . . . Ha, S. (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. *The Journal of Physical Chemistry B*, *102*(18), 3586–3616.
- Makrodimitris, S., van Ham, R. C., & Reinders, M. J. (2019). Improving protein function prediction using protein sequence and GO-term similarities. *Bioinformatics (Oxford, England)*, *35*(7), 1116–1124.
- Mariani, V., Kiefer, F., Schmidt, T., Haas, J., & Schwede, T. (2011). Assessment of template based protein structure predictions in CASP9. *Proteins: Structure, Function, and Bioinformatics*, *79*(S10), 37–58.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). *Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems* (pp. 3111–3119). The MIT Press.
- Mills, C. L., Beuning, P. J., & Ondrechen, M. J. (2015). Biochemical functional predictions for protein structures of unknown or uncertain function. *Computational and Structural Biotechnology Journal*, *13*, 182–191. Available from <https://doi.org/10.1016/j.csbj.2015.02.003>.

- Miranda, L. J., & Hu, J.A. (2018). Deep learning approach based on stacked denoising autoencoders for protein function prediction. In *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)* (pp. 480–485). IEEE.
- Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A., & Punta, M. (2013). Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Research*, *41*(12), e121. Available from <https://doi.org/10.1093/nar/gkt263>.
- Mittal, J., & Best, R. B. (2010). Tackling force-field bias in protein folding simulations: Folding of Villin HP35 and Pin WW domains in explicit water. *Biophysical Journal*, *99*(3), L26–L28.
- Molina, L. C., Belanche, L., & Nebot, À. (2002). Feature selection algorithms: A survey and experimental evaluation. In *2002 IEEE International Conference on Data Mining. Proceedings* (pp. 306–313). IEEE.
- Montelione, G. (2012). Template based modeling assessment in CASP10. In *10th community wide experiment on the critical assessment of techniques for protein structure prediction* (pp. 9–12), Gaeta, Italy.
- Murzin, A. G., Brenner, S. E., Hubbard, T., & Chothia, C. (1995). SCOP: A structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, *247*(4), 536–540. Available from [https://doi.org/10.1016/S0022-2836\(05\)80134-2](https://doi.org/10.1016/S0022-2836(05)80134-2).
- Nauman, M., Rehman, H. U., Politano, G., & Benso, A. (2019). Beyond homology transfer: Deep learning for automated annotation of proteins. *Journal of Grid Computing*, *17*(2), 225–237.
- Nemethy, G., Gibson, K. D., Palmer, K. A., Yoon, C. N., Paterlini, G., Zagari, A., . . . Scheraga, H. A. (1992). Energy parameters in polypeptides. 10. Improved geometrical parameters and nonbonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides. *The Journal of Physical Chemistry*, *96*(15), 6472–6484.
- Neria, E., Fischer, S., & Karplus, M. (1996). Simulation of activation free energies in molecular systems. *The Journal of Chemical Physics*, *105*(5), 1902–1921.
- Nguyen, H., Maier, J., Huang, H., Perrone, V., & Simmerling, C. (2014). Folding simulations for proteins with diverse topologies are accessible in days with a physics-based force field and implicit solvent. *Journal of the American Chemical Society*, *136*(40), 13959–13962.
- Ni, Q., Wang, Z. -Z., Han, Q., Li, G., Wang, X., & Wang, G. (2009). Using logistic regression method to predict protein function from protein-protein interaction data. In *2009 3rd international conference on bioinformatics and biomedical engineering* (pp. 1–4). IEEE.
- Nievolá, J. C., Paraiso, E. C., & Freitas, A. A. (2015). A hierarchical neural network for predicting protein functions. In *2015 IEEE 15th international conference on bioinformatics and bioengineering (BIBE)* (pp. 1–5). IEEE.
- Oldziej, S., Czaplowski, C., Liwo, A., Chinchio, M., Nancias, M., Vila, J., . . . Scheraga H. A. (2005). Physics-based protein-structure prediction using a hierarchical protocol based on the UNRES force field: Assessment in two blind tests. *Proceedings of the National Academy of Sciences of the United States of America*, *102* (21), 7547–7552.
- Orengo, C. A., & Taylor, W. R. (1996). SSAP: Sequential structure alignment program for protein structure comparison. *Methods in Enzymology*, *266*, 617–635. Available from [https://doi.org/10.1016/s0076-6879\(96\)66038-8](https://doi.org/10.1016/s0076-6879(96)66038-8).
- Pearce, N. M., Krojer, T., Bradley, A. R., Collins, P., Nowak, R. P., Talon, R., . . . von Delft, F. (2017). A multi-crystal method for extracting obscured crystallographic states from conventionally uninterpretable electron density. *Nature Communications*, *8*, 15123. Available from <https://doi.org/10.1038/ncomms15123>.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning research*, *12*, 2825–2830.
- Petrey, D., Chen, T. S., Deng, L., Garzon, J. I., Hwang, H., Lasso, G., . . . Honig, B. (2015). Template-based prediction of protein function. *Current Opinion in Structural Biology*, *32*, 33–38. Available from <https://doi.org/10.1016/j.sbi.2015.01.007>.
- Piana, S., Klepeis, J. L., & Shaw, D. E. (2014). Assessing the accuracy of physical models used in protein-folding simulations: Quantitative evidence from long molecular dynamics simulations. *Current Opinion in Structural Biology*, *24*, 98–105.
- Piana, S., Lindorff-Larsen, K., & Shaw, D. E. (2011). How robust are protein folding simulations with respect to force field parameterization? *Biophysical Journal*, *100*(9), L47–L49.
- Piana, S., Lindorff-Larsen, K., & Shaw, D. E. (2012). Protein folding kinetics and thermodynamics from atomistic simulation. *Proceedings of the National Academy of Sciences of the United States of America*, *109* (44), 17845–17850.
- Piana, S., Lindorff-Larsen, K., & Shaw, D. E. (2013). Atomic-level description of ubiquitin folding. *Proceedings of the National Academy of Sciences of the United States of America*, *110* (15), 5915–5920.
- Roy, A., Kucukural, A., & Zhang, Y. (2010). I-TASSER: A unified platform for automated protein structure and function prediction. *Nature Protocols*, *5*(4), 725–738.
- Sadowski, M. I., & Taylor, W. R. (2010). Protein structures, folds and fold spaces. *Journal of Physics Condensed Matter: An Institute of Physics Journal*, *22*(3), 033103. Available from <https://doi.org/10.1088/0953-8984/22/3/033103>.
- Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, *23*(19), 2507–2517.
- Šali, A., & Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, *234*(3), 779–815.
- Samudrala, R., & Moult, J. (1998). An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *Journal of Molecular Biology*, *275*(5), 895–916.
- Sela, M., White, F. H., Jr, & Anfinsen, C. B. (1957). Reductive cleavage of disulfide bridges in ribonuclease. *Science (New York, N.Y.)*, *125*(3250), 691–692.
- Shao, C., Yang, H., Westbrook, J. D., Young, J. Y., Zardecki, C., & Burley, S. K. (2017). Multivariate analyses of quality metrics for crystal structures in the PDB archive. *Structure (London, England: 1993)*, *25*(3), 458–468. Available from <https://doi.org/10.1016/j.str.2017.01.013>.

- Shatkay, H., Höglund, A., Brady, S., Blum, T., Dönnies, P., & Kohlbacher, O. (2007). SherLoc: High-accuracy prediction of protein subcellular localization by integrating text and protein sequence data. *Bioinformatics (Oxford, England)*, *23*(11), 1410–1417.
- Shen, My, & Sali, A. (2006). Statistical potential for assessment and prediction of protein structures. *Protein Science*, *15*(11), 2507–2524.
- Shen, Y., & Bax, A. (2013). Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks. *Journal of Biomolecular NMR*, *56*(3), 227–241.
- Sillitoe, I., Dawson, N., Lewis, T. E., Das, S., Lees, J. G., Ashford, P., . . . Orengo, C. A. (2019). CATH: Expanding the horizons of structure-based functional annotations for genome sequences. *Nucleic Acids Research*, *47*(D1), D280–D284. Available from <https://doi.org/10.1093/nar/gky1097>.
- Simons, K. T., Kooperberg, C., Huang, E., & Baker, D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *Journal of Molecular Biology*, *268*(1), 209–225.
- Skolnick, J. (2006). In quest of an empirical potential for protein structure prediction. *Current Opinion in Structural Biology*, *16*(2), 166–171.
- Skolnick, J., Godzik, A., Jaroszewski, L., & Kolinski, A. (1997). Derivation and testing of pair potentials for protein folding. When is the quasichemical approximation correct? *Protein Science*, *6*(3), 676–688.
- Slabinski, L., Jaroszewski, L., Rodrigues, A. P., Rychlewski, L., Wilson, I. A., Lesley, S. A., & Godzik, A. (2007). The challenge of protein structure determination—Lessons from structural genomics. *Protein Science*, *16*(11), 2472–2482.
- Smart, O. S., Horsky, V., Gore, S., Svobodova Varekova, R., Bendova, V., Kleywegt, G. J., & Velankar, S. (2018). Worldwide Protein Data Bank validation information: Usage and trends. *Acta Crystallographica Section D: Structural Biology*, *74*(Pt 3), 237–244. Available from <https://doi.org/10.1107/S2059798318003303>.
- Söding, J. (2005). Protein homology detection by HMM–HMM comparison. *Bioinformatics (Oxford, England)*, *21*(7), 951–960.
- Summa, C. M., & Levitt, M. (2007). Near-native structure refinement using in vacuo energy minimization. *Proceedings of the National Academy of Sciences of the United States of America*, *104* (9), 3177–3182.
- Tai, C. H., Bai, H., Taylor, T. J., & Lee, B. (2014). Assessment of template-free modeling in CASP10 and ROLL. *Proteins: Structure, Function, and Bioinformatics*, *82*, 57–83.
- Tang, H., Zhao, Y.-W., Zou, P., Zhang, C.-M., Chen, R., Huang, P., & Lin, H. (2018). HBPred: A tool to identify growth hormone-binding proteins. *International Journal of Biological Sciences*, *14*(8), 957.
- Taylor, W. R., Bartlett, G. J., Chelliah, V., Klose, D., Lin, K., Sheldon, T., & Jonassen, I. (2008). Prediction of protein structure from ideal forms. *Proteins: Structure, Function, and Bioinformatics*, *70*(4), 1610–1619.
- Tieleman, T., & Hinton, G. (2012). Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning*, *4*(2), 26–31.
- van Gunsteren, W. F., Billetter, S., Eising, A., Hünenberger, P., Krüger, P., Mark, A., . . . Tironi, I. (1996). *Biomolecular simulation: The GROMOS96 manual and user guide*. Vdf Hochschulverlag AG an der ETH Zürich, Zürich.
- Vens, C., Struyf, J., Schietgat, L., Džeroski, S., & Blockeel, H. (2008). Decision trees for hierarchical multi-label classification. *Machine Learning*, *73* (2), 185.
- Verspoor, K. M. (2014). *Roles for text mining in protein function prediction. Biomedical literature mining* (pp. 95–108). Springer.
- Vieth, M., Koliński, A., Brooks, C. L., III, & Skolnick, J. (1994). Prediction of the folding pathways and structure of the GCN4 leucine zipper. *Journal of Molecular Biology*, *237*.
- Wang, J., Zhang, L., Jia, L., Ren, Y., & Yu, G. (2017). Protein-protein interactions prediction using a novel local conjoint triad descriptor of amino acid sequences. *International Journal of Molecular Sciences*, *18*(11), 2373.
- Wang, L., Wang, Y., & Chang, Q. (2016). Feature selection methods for big data bioinformatics: A survey from the search perspective. *Methods (San Diego, Calif.)*, *111*, 21–31.
- Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G., . . . Weiner, P. (1984). A new force field for molecular mechanical simulation of nucleic acids and proteins. *Journal of the American Chemical Society*, *106*(3), 765–784.
- Wong, A., & Shatkay, H. (2013). Protein function prediction using text-based features extracted from the biomedical literature: The CAFA challenge. *BMC Bioinformatics*, *14*(Suppl. 3), S14.
- Wu, S., & Zhang, Y. (2008a). A comprehensive assessment of sequence-based and template-based methods for protein contact prediction. *Bioinformatics (Oxford, England)*, *24*(7), 924–931.
- Wu, S., & Zhang, Y. (2008b). MUSTER: Improving protein sequence profile–profile alignments by using multiple sources of structure information. *Proteins: Structure, Function, and Bioinformatics*, *72*(2), 547–556.
- Xu, D., & Zhang, Y. (2012). Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins: Structure, Function, and Bioinformatics*, *80*(7), 1715–1735.
- Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., & Zhang, Y. (2015). The I-TASSER Suite: Protein structure and function prediction. *Nature Methods*, *12*(1), 7–8.
- Yang, J. Y., & Yang, M. Q. (2006). Assessing protein function using a combination of supervised and unsupervised learning. In *Sixth IEEE symposium on bioinformatics and bioengineering (BIBE'06)* (pp. 35–44). IEEE.
- Yip, K. M., Fischer, N., Paknia, E., Chari, A., & Stark, H. (2020). Atomic-resolution protein structure determination by cryo-EM. *Nature*, *587*(7832), 157–161. Available from <https://doi.org/10.1038/s41586-020-2833-4>.
- You, R., Huang, X., & Zhu, S. (2018). DeepText2Go: Improving large-scale protein function prediction with deep semantic text representation. *Methods (San Diego, Calif.)*, *145*, 82–90.

- You, Z.-H., Lei, Y.-K., Zhu, L., Xia, J., & Wang, B. (2013). Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis. *BMC Bioinformatics*, *14*(Suppl. 8), S10.
- Young, J. Y., Westbrook, J. D., Feng, Z., Sala, R., Peisach, E., Oldfield, T. J., . . . Burley, S. K. (2017). OneDep: Unified wwPDB System for deposition, biocuration, and validation of macromolecular structures in the PDB archive. *Structure (London, England: 1993)*, *25*(3), 536–545, Epub 2017 Feb 9. PMID: 28190782; PMCID: PMC5360273.
- Zhang, Q. C., Petrey, D., Deng, L., Qiang, L., Shi, Y., Thu, C. A., . . . Hunter, T. (2012). Structure-based prediction of protein–protein interactions on a genome-wide scale. *Nature*, *490*(7421), 556–560.
- Zhang, Y., Kihara, D., & Skolnick, J. (2002). Local energy landscape flattening: Parallel hyperbolic Monte Carlo sampling of protein folding. *Proteins: Structure, Function, and Bioinformatics*, *48*(2), 192–201.
- Zhang, Y., Kolinski, A., & Skolnick, J. (2003). TOUCHSTONE II: A new approach to ab initio protein structure prediction. *Biophysical Journal*, *85*(2), 1145–1164.
- Zhang, Y., & Skolnick, J. (2004a). Automated structure prediction of weakly homologous proteins on a genomic scale. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(20), 7594–7599.
- Zhang, Y., & Skolnick, J. (2004b). SPICKER: A clustering approach to identify near-native protein folds. *Journal of Computational Chemistry*, *25*(6), 865–871.
- Zhang, Y., & Skolnick, J. (2005a). The protein structure prediction problem could be solved using the current PDB library. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(4), 1029–1034.
- Zhang, Y., & Skolnick, J. (2005b). TM-align: A protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, *33*(7), 2302–2309.
- Zou, Z., Tian, S., Gao, X., & Li, Y. (2019). mldeepre: Multi-functional enzyme function prediction with hierarchical multi-label deep learning. *Frontiers in Genetics*, *9*, 714.