**Minireview Article**

**Optimization in Protein structure and function prediction of silk of Tasar silkworm, *Antheraea mylitta* Drury**

**Abstract**

 Protein structure and function prediction is an arduous task in computational biology. Understanding the structure of protein facilitates in understanding its function. The Tasar silkworm, *Antheraea mylitta* Drury,is a vital species in sericulture, producing high-quality silk fibres with unique properties. Beyond its economic significance, the Tasar silkworm is a model organism for studying insect biology, development, and evolution. Despite recent advances in genomics and proteomics, which have enabled the exploration of the Tasar silkworm proteome, there still exist hitches in the atomic accuracy of its protein structure prediction in the era of deep learning. Here, we endeavour to bridge the existing knowledge gap of algorithmic modelling techniques, especially deep neural networks in protein prediction methods of this valuable lepidopteran, leading to a better understanding of its biological processes and potential applications.

 **Keywords:** Optimization, Protein structure and function, algorithmic modelling, deep neural network

**Introduction**

 Tasar cocoons are reported to be the largest among all the silk producing insects in the world [1]. Tasar silk producing silkworm *Antheraea mylitta* Drury is highly heterogenic with several ecoraces in varied topographical diverse areas exhibiting diversity in phenotypic characters and forms an integral part of forest ecosystem [2] However, it is on the verge of extinction due to its weakness in voltinism, emergence, hatching, low yield etc. [3],[4],[5]. The extinction process can be mitigated to some extent by deciphering the proteome enabling the identification of functional proteins, their interactions and regulation. The current paper discusses the optimization of machine learning and deep learning methods to be used for protein structure and function prediction of silk of Tasar silkworm *Antheraea mylitta* Drury.

 The advancements in computational biology have made molecular modelling of biological compounds a reality [6] For instance, proteins—composed of polypeptide chains-undergo intricate folding processes at multiple structural levels to become functional. After synthesis by ribosomes, these chains first fold into secondary structures such as alpha helices and beta sheets, often connected by turns [7].This is followed by tertiary [8] and quaternary folding [9], which are essential for achieving the final three-dimensional conformation necessary for proper protein function. In nature, although many varieties of Protein folding are seen, but SCOP (Structural Classification of Proteins) and CATH are two most common ones [10]. The amino acids that make up the polypeptide determine the folding, aka the interaction between the structures, determines the physio-chemical nature of the Proteins. Hence understanding the constitutive aminoacids of a protein is crucial to understand the the mechanism of function of proteins [11]

Although understanding the constituents of the proteins in the language of Amino Acids sounds easy, but physically understanding the inter-aminoacid interactions in secondary structures and further folding is tricky. Hence, for this purpose use of traditional prediction methods along with AI powered one have proven quite helpful.

**PROTEIN PREDICTION METHODS**

Various methods are been employed for protein prediction, which broadly categorized into several approaches. Homology-based methods include tools such as BLAST (Basic Local Alignment Search Tool) and PSI-BLAST (Position-Specific Iterative BLAST), along with HMMER, which utilizes Hidden Markov Models (HMM) for sequence alignment and comparison. Machine learning-based approaches involve algorithms like Support Vector Machines (SVM), Random Forests, Neural Networks, and Gradient Boosting, which can identify patterns from large datasets. Deep learning-based techniques offer advanced analysis through architectures such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) networks. Additionally, physicochemical property-based methods predict proteins based on features like hydrophobicity, charge, and secondary structure. Ab initio methods, which predict protein structures from scratch without relying on homologous templates, include tools such as Rosetta, trRosetta (transform-restrained Rosetta), I-TASSER (Iterative Threading Assembly Refinement), and QUARK. Hybrid methods combine multiple strategies from the aforementioned categories to enhance prediction accuracy. Furthermore, template-based modeling compares the target protein with known structures as templates or by aligning smaller fragments from existing proteins.

Protein structure prediction can be approached using two broad categories: **Experimental Methods** and **Computational Methods**. Experimental techniques such as X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy (Cryo-EM) are widely employed for determining the three-dimensional (3D) structure of proteins, each with its own strengths and limitations [12]. In X-ray crystallography, the protein must be purified and crystallized under optimal conditions before being exposed to an X-ray beam to obtain diffraction patterns [13] In NMR spectroscopy, it analyzes proteins in solution, especially if the protein structure is flexible[14]. Cryo-EM, involving electron beams and lenses, provides direct imaging of biomolecules, and in certain cases, electron diffraction from 2D/3D crystals is employed as well, to determine the structural information [15],[16]. As for the Computational methods, Homology Modelling, Fold Recognition (also known as threading), and AB Initio Modelling are widely used. Homology modelling is considered as one of the most straightforward and reliable techniques when the target protein shares >30% sequence identity with a known template structure [17],[18]. The process involves multiple steps such as template identification (using BLAST: [19] or FASTA: [20] sequence alignment (e.g., CLUSTAL W: [21] structural modelling, loop modelling, side chain optimization, energy minimization, and model validation ([22]sequentially. Loop regions, which often contribute to protein flexibility and functional specificity, are modelled using knowledge-based methods supported by software such as MODELLER [23], Swiss-Model [24] Insight ([25], and 3D-Jigsaw [26]. Energy functions using Monte Carlo or Molecular Dynamics Simulations can be applied to optimize loop conformations ([27], However, a low energy state does not always indicate correct folding, as some misfolded structures can also exhibit energetically favourable conformations [28] Hence, model validation should include verification for Torsion Angles, Bond Lengths, Bond Angles, and Hydrophobic Core Packing [29] In cases where sequence similarity is below 30%, Threading or Fold Recognition both are preferred [30],[31]. These techniques specialize in detecting distant homologs by matching the target sequence to known folds using structural features. 3D profiles, which can assess the compatibility of amino acids between multiple proteins with structural environments, are also preferred in some, like Tools of Profile-3D and PROSPECT [32],[33]. Ab Initio Modelling, such as the ROSETTA platform (developed by: David Baker lab; [34],[35] can construct protein models without a template, using fragment assembly techniques based on structures from the Protein Data Bank (PDB). Despite its innovative strategy, the accuracy of Ab Initio Models is comparatively lower compared to template-based methods [36]

Protein structure prediction is continuously benchmarked through community-wide experiments like the Critical Assessment of protein Structure Prediction (CASP), which classifies prediction targets into Template-Based Modelling (TBM) and Free Modelling (FM) categories [37] Furthermore, machine learning techniques (ML Techniques), including Support Vector Machines (SVM) and kernel methods, are increasingly being integrated into structural bioinformatics recently. These approaches leverage both homogeneous and heterogeneous datasets to recognize structural and functional patterns ,[.38]

 **UTILISING PROTEOMICS AND PROTEIN PREDICTION METHODS IN TASAR SILKWORM**

Proteomic techniques can also be effectively used to study protein alterations throughout embryonic development, larval stages, and pupation of the tropical tasar silkworm *Antheraea mylitta* Drury. By integrating proteomics with genomics, we can gain a deeper understanding of gene expression and protein function in various silkworm species, including both closely and distantly related species of *Antheraea mylitta*. This integrative approach may be proven as a very useful approach towards enhancing tasar silkworm breeding, sericulture practices, and biotechnological innovations across the country. The identification of vitellogenin, a female-specific protein present in the hemolymph, marked a significant milestone in insect biochemistry, first reported in Hyalophora cecropia by [39] This discovery paved a smoother pathway for deeper molecular insect studies. Subsequently, the mapping of the silkworm genome[40][41] enabled the application of advanced proteomic technologies to silkworm research. In this context, two-dimensional electrophoresis (2D-PAGE) combined with mass spectrometry has proven particularly effective for analyzing this protein’s expression in various tissues such as silk glands, fat body, skeletal muscle, and haemolymph [42][43] The development of the silk gland, which undergoes significant growth during the fourth and fifth larval instars, is largely affected by environmental factors such as light, diet, temperature, and humidity .In a key contribution to silk protein characterization, [44] successfully isolated, purified, and analysed proteins from the cocoon peduncle of Antheraea mylitta. Further advancements were made in this field )[45], who identified 169 distinct proteins in *Bombyx mori* cocoon silk using shotgun LC-MS/MS techniques. In addition to the primary silk proteins, fibroin and sericin, studies have revealed the presence of enzymes, protease inhibitors, and proteins of yet-unknown function. The resulting mass spectrometry data were analysed using MaxQuant software (version 1.3.0.1) [46], while the database searches were conducted against the integrated silkworm proteome resource containing 35,379 protein sequences from NCBI and SilkDB [47]. In parallel, amino acid sequence analysis of tasar silkworm proteins has provided critical insights into their structure, function, and biochemical properties. Structural predictions can be achieved through homology modelling when similar known protein structures are available, or via ab initio methods in the absence of such templates. These computational approaches aid in understanding biological functions such as enzymatic activity, ligand-binding affinity, subcellular localization, post-translational modifications (e.g., phosphorylation, glycosylation), and interactions with other proteins or small molecules like hormones and drugs. To support these predictions, several robust bioinformatics tools are employed. BLAST and HMMER facilitate sequence alignment and motif detection, while SWISS-MODEL and I-TASSER assist in 3D structural modelling. Functional annotations can be predicted using Functome, subcellular localization through SubLoc, and post-translational modifications via NetPhos. This review highlights a comparative analysis of these protein prediction methods, emphasizing both their advantages and limitations in advancing silkworm proteomics.

**RESULTS AND DISCUSSION**

One of the major challenges in protein structure and function prediction for the Tasar silkworm lies in the limited availability and quality of genomic and proteomic data. This scarcity hinders the training and validation of computational prediction models. As of March 2021, while the Protein Data Bank (PDB) housed around 180,000 protein structures [48] databases like UniProt and TrEMBL had catalogued approximately 207 million protein sequences by the end of 2020 (UniProt Consortium, 2021). The vast disparity between available sequence and structure data creates a bottleneck for accurate model training.

Proteins from the Tasar silkworm are often complex, which adds to the difficulty of structural prediction. The lack of closely related species with well-characterized proteins further limits the accuracy of homology-based models. In particular, predicting reliable 3D structures for large and intricate silk proteins such as fibroin and sericin remains a substantial challenge. Additionally, assigning specific functions to predicted protein sequences is often inconclusive, and post-translational modifications (PTMs) like phosphorylation and glycosylation—crucial to protein function—are difficult to predict with current tools. Another layer of complexity arises in the prediction of protein–protein interactions, which are essential to understanding biological processes. The sheer volume of protein sequences demands considerable computational resources, both for storage and analysis. Moreover, existing algorithms may fail to capture the full spectrum of protein behaviour, and no universally accepted protocols currently exist for validating predictions. Variations in input parameters and methodology further affect the consistency and reproducibility of outcomes. Function prediction is further complicated by moonlighting proteins, which perform multiple unrelated biochemical or biophysical roles [49][50] Some enzymes, for instance, have dual roles—acting catalytically and also regulating transcription or translation through interactions with DNA, RNA, or transcription factors [51]

To address some of these challenges, machine learning approaches have been integrated into protein prediction frameworks. These models consider factors such as the frequency of Gene Ontology (GO) terms, amino acid sequence patterns, domain and motif presence, and various biophysical attributes [52]. A recent review by [53] offers a comprehensive overview of machine learning methodologies in protein function prediction, including feature selection techniques, algorithm types, model implementations, and evaluation strategies. The complexity increases with multi-domain proteins and longer sequences, which tend to have higher radii of gyration, making structural modeling more difficult [54][55]. In cases where no structural homologs with significant sequence similarity are found in PDB, alternative strategies like threading or fold recognition are employed. These methods attempt to align target sequences with known structural folds, exploiting the evolutionary conservation of protein structures [56] [57][58][59][60]There is also growing recognition of the functional importance of microproteins—those with fewer than 100 amino acids—which may serve regulatory, structural, or other roles [61] Community initiatives like the Critical Assessment of Functional Annotation (CAFA) have played a pivotal role in benchmarking algorithm performance and have reported steady improvements over the last decade [62][63][64]

**CONCLUSION**

Protein prediction in the Tasar silkworm is an emerging and rapidly evolving field that holds great promise for advancing our understanding of this economically and ecologically significant insect. Through the application of cutting-edge computational techniques and integrative strategies, researchers are now able to predict protein structures, functions, and properties with increasing accuracy. Potential applications of this research span improved silk production, novel biomaterials, and future biopharmaceutical developments. To realize these benefits, future research should prioritize: such as Generation of high-quality genomic and proteomic datasets; Development of advanced AI and machine learning-based prediction models; Integration of multi-omics data (genomics, transcriptomics, proteomics); Experimental validation of computational predictions; Exploration of industrial and biotechnological applications.

Interestingly, polypeptide studies suggest that shorter sequences can introduce more complexity and lower confidence in structural prediction [65][66]. This is due to the surface elements responsible for physicochemical interactions and binding often not being directly involved in the folding process, which means they may not significantly affect pharmacological or toxicological behavior [67] In conclusion, this comprehensive review of protein prediction in *Antheraea mylitta* provides critical insights and lays a robust foundation for future investigations. Continued advancements in this field are essential for unlocking the full biological and industrial potential of Tasar silk proteins.

**Table-1: Highlights of different approaches to protein structure prediction in a tabular form**

|  |  |  |
| --- | --- | --- |
| Year | Discoverer/Author | Type of Approach |
| 1969 | Braune et al. | Comparative modelling (Summers and Karplus, 1989)[68] |
| 1974 | Richard Corey and Irving Kuntz | First protein folding algorithm (DREIDING) (Niazi et al., 2024) [69] |
| 1991 | Bowie et al.[32] | Threading (Xu et al., 2007) [70] |
| 1994 | Moult et al.[37] | CASP(Critical Assessment of Structure Prediction) (Moult et al., 1995)  |
| 1995 | Sali et al.[71] | MODELLER  |
| 1997 | Altschul et al.[19] | PSI-BLAST(Position Specific Iterative Basic Local Alignment Search Tool)  |
| 1997 | Baker et al. | Rosetta (Simons et al. 1999)[[72] |
| 1999 | Jones[58] | PSIPRED(Position Specific Iterative Blast based secondary structure prediction)  |
| 2004 | Zhou and Zhou[73] | SPARKS  |
| 2005 | Zhang et al.[74] | I-TASSER(Iterative Threading ASSEmbly Refinement) (Zhang, 2008) |
| 2005 | XU et al.[75] | QUARK |
| 2018 | Senior et al.[76] | Alphafold |
| 2019 | Ingraham et al.[77] | NEMO(NF-Kappa-B Essential Modulator)  |
| 2019 | AlQuraishi et al.[78] | RGN(Recurrent Geometric Network)  |
| 2019 | Mao et al.[79] | GDFold(Gradient Decent Folding)  |
| 2020 | Yang et al.[80] | trRosetta(transform restrained Rosetta)  |
| 2021 | Ju et al.[81] | CopulaNet |
| 2021 | Google Deep mind | Alphafold 2 (Jumper et al, 2021) [82] |
| 2021 | Baek et al.[83] | RoseTTA fold |
| 2022 | Brandes et al.[84] | Protein BERT |
| 2022 | Chowdhury et al.[85] | RGN2(Recurrent Geometric Network 2)  |
| 2022 | Yu et al.[86] | Profold single(Protein fold single)  |
| 2022 | Wu et al.[87] | Omegafold |
| 2022 | Kandthil et al. [88] | DMP fold 2 |
| 2022 | Lin et al. [89] | ESM fold (Evolutionary Scale Modeling)  |
| 2024 | Google DeepMind and Isomorphic Labs | Alphafold 3 (Abramson et al., 2024)[90] |

**Table 2: Current Prediction Methods: (A)Prokaryotic localization predictors**

**(B) Eukaryotic localization predictors**

|  |
| --- |
| **(A)Prokaryotic localization predictors** |
| Prediction Method | Proposer or Discoverer | Year of Discovery |
| PSORT I | Nakai K. And Kanehisa[91] | 1991 |
| PSORT B | Nakai K. and Kanehisa [92] | 2001 (Gardy et al, 2003)[93] |

|  |
| --- |
| **(B) Eukaryotic localization predictors** |
| Prediction Method | Proposer or Discoverer | Year of Discovery |
| PSORT II | Nakai K. and Horton P.[94] | 1999 |
| i PSORT | Bannai et al.[95] | 2001 (Bannai et al., 2002) |
| Target P | Emanulsson et al.[96] | 2000 |
| ESLPred | Bhasin and Raghava[97] | 2004 |

**Table-3: Prokaryotic and Eukaryotic localization predictors**

|  |  |  |
| --- | --- | --- |
| Prediction Method | Proposer or Discoverer |  Year of discovery |
| NNPSL | Reinhardt and Hubbard[98]  | 1998 |
| Subloc(Subcellular Localization Prediction)  | Hua and Sun [99] | 2001 |

* Data Availability Statement: Not Applicable
* Funding Statement: Not Applicable
* Conflict of Interest Disclosure: There is no Conflict of Interest between the Authors
* Ethics approval statement: Not Applicable

Disclaimer (Artificial intelligence): Author(s) hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc.) and text-to-image generators have been used during the writing or editing of this manuscript.

COMPETING INTERESTS DISCLAIMER:

Authors have declared that they have no known competing financial interests OR non-financial interests OR personal relationships that could have appeared to influence the work reported in this paper.**References**

1. Akai, H. (2000). Cocoon filament characters and post cocoon technology. International Journal of wild silkmoth silk, 255-259
2. Rao, N. A. (2007). Forest ecology in India Foundation Press
3. Thangavelu, K. (1992). Population ecology of *Anthreaea mylitta* Drury (Lepidoptera:Saturniidae). Wild silkmoths 92,99-104
4. Srivastava, A. K., P. K., Naqvi, A. H., Sinha, A. K., Singh, B. M. K., Sinha B. R. R. P., & Thangavelu, K. (2003) Descriptor of tropical tasar silkworm *Anthreaea mylitta* Drury ( Lepidoptera:Saturniidae). Central Tasar Research and Training Institute, Ranchi, India
5. Kavane, R. P., & Sathe, T. V. (2009). On a new variety (Kolhapurensis) of *Anthreaea* *mylitta* from India. Biological Forum. An International Journal, 1(2), 45-47
6. Kumar A, Chordia N(2017) Role of bioinformatics in biotes Rev Biosci 12(1) :116
7. Voet, D., Voet J. G. (1990). Biochemistry ( 2nd Ed.) John Wiley and sons
8. Dorn, M., Breda. A., & de Souza. O. N. (2008). A hybrid method for the protein structure prediction problem. In A. L. C. Bazzan M. Craven & N. F. Martins ( Eds.) Advances in Bioinformatics and Computational Biology:BSB 2008( Lecture Notes in Computer Science, vol 5167,pp 56-67).Springer Berlin, Heidelberg
9. Brown, S., & Head Gordon, T. (2003) Coarse grained sequences for protein folding and design. Proceedings of the National Academy of Sciences USA, 100.(19), 10712-10717 <https://doi.org/10.1073/pnas.1931882100>
10. Csaba, G., Birzele, F., &Zimmer, R. (2009) Systematic comparison of SCOP and CATH: a new gold standard for protein structure analysis. BMC structural Biology, 9,23 <https://doi.org/10.1186/1472-6807-9-23>
11. He, Y., Chen, Y., Alexander, P. A., Bryan, P. N., &Orban, J. (2012) Mutational tipping points for switching protein folds and functions. Structure, 20(2), 283-291.https://doi. org/10.1016/j.str.2011.11.018
12. Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, H., &Phillips, D. C. (1958). “A three dimensional model of the myoglobin molecule obtained by X-ray Analysis. “Nature, 181.(4610), 662-666.doi:10.1038/181662ao
13. Burley, S. K., Berman, H. M., Bhikadiya, C., Bi, C., Chen, L., Di Costanzo, L., & Westbrook, J. D. (2019) RCSB protein Data Bank :biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. Nucleic Acids Research, 47(D1), D464-D474.
14. Snyder, D. A., Chen, Y., Denissova, N. G., Acton, T., Aramini, J. M., Ciano, M., Montelione, G. T. (2005) Comparison of NMR Spectral Quality and Success in Crystallization Demonstrate that NMR and X-ray Crystallography Are complementary methods for small protein structure Determination Journal of the American Chemical Society 127(47), 16505-16511 <https://doi.org/10.1021/ja053564h>
15. Orlova, E. V., & Sainik, H. R. (2011) Structural analysis of macromolecular assemblies by electron microscopy Chemical Reviews, 111(12), 7710-7748
16. Rabel, J. J, Abrishami, V., & Meier, E. V. (2010). Preparation of 2 D crystals of membrane proteins for high resolution electron crystallography data collection. Journal of Structural Biology, 171(1), 70-79.
17. Peitsch, M. C. (1997). Large scale protein modelling and model repository. In T. Gaasterland, P. Karp, K. Karplus, C. Ouzounis, & C. Sander(Eds). Proceedings of the fifth International Conference on Intelligent Systems for Molecular Biology ( Vol. 5,pp-234-236) AAAI press
18. Sanchez. R., & Sali, A. (1997). Advances in comparative protein- structure modelling. Current opinion in Structural Biology, 7(2), 206-214.doi-10.1016/S0959-440 (97) 80027-9
19. Altschul, S. F., Gish, W., Miller, W. Myers, E. W., &Lipman, D. J. (1990). Basic Local Alignment Search Tool Journal of Molecular Biology. 215(3) , 403-410.https://doi.org/10.1016/S0022-2836(05) 80360\_2.
20. Pearson, W. R. (1990). Rapid and sensitive sequence comparison with FASTP and FASTA. Methods in Enzymology, 183,63-98.https:doi.org/10.1016/0076-6879(90) 83007-V
21. Thompson, J. D. Higgins, D. G., S Gibson, T. J. (1994) CLUSTAL W:Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position- specific gap penalties and weight matrix choice. Nucleic acids Research 22(22), 4673-4680 https:doi.org/10.1093/nar/22.22.4673
22. Johnson , M. S., Srinivasan, N., Sowdhamini, R., &Blundell, T. L. (1994). Knowledge based protein modelling. Critical Review in Biochemistry and Molecular Biology 29(1), 1-68 <https://doi.org/10.3109/10409239409086801>
23. Sali, A., &Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. Journal of molecular Biology, 234(3), 779-815,.https://doi.org/10.1006/jmbi.1993.1626
24. Peitsch, M. C. (2000). About Swiss- model and the Swiss PDB viewer :An environment for comparative protein modelling. Electrophoresis, 21(1), 85-91.https://doi.org/10.1002/SICI 1522-2683(20000101)
25. Dayringer, H. E., Tramentano, A., Sprang, S. R., &Fletterick R. J. (1986). Interactive program for visualization and modelling of proteins, nucleic acids and small molecules Journal of Molecular Graphics, 4(2), 82-87.https://doi.org/10.101610263-7855(86) 80003-5
26. Bates, P. A., & Stenberg, M. J. E. (1999). Model building by comparison at CASP3: Using expert Knowledge and computer automation. Proteins structure, Function and Bioinformatics, Suppl3, 47-54 [https://doi.org/10.1002/(SICI)](https://doi.org/10.1002/%28SICI%29) 1097-0134(1999) 37:3
27. Fiser, A., Do, R. K., & Sali, A. (2000) modelling of loops in protein structutes. Protein Science, 9(9) 1753-1773.doi-10.1110/ps.9.9.1753
28. Novotny , J., Bruccoleri, R. E., & Karplus. M. (1988). Empirical free energy calculations:A blind test and further applications. Proteins :Structure, Function and Bioinformatics 4(1), 19-30 <https://doi.org/10.1002/prot.340040104>
29. Baumann, G., Frommel, C., & Sander, C. (1989).” Polarity as a criterion in protein design.”Protein Engineering, 2,329-334
30. Sippl, M. J. (1993) Recognition of errors in three dimensional structures of proteins. Proteins:Structure, Function and Genetics, 17(4), 355-362.https://doi.org/10.1002/prot.340170404
31. Hendlich, M. Lackner, P., Weitckus, S. Floeckner, H., Froschauer, R., Gottsbacher, K., Tanner, W., &Froschauer, R., (1990) Identification of native protein folds amongst a large number of incorrect models. The role of surface accessibility patterns. Jornal of Molecular Biology, 216.(1), 167-180.https:doi.org/10.10161S0022-2836(99) 80006-1
32. Bowie, J. U., Luthy, R., &Eisenberg, D. (1991) “A method to identify protein sequences that fold into a known three – dimensional structure. “Science, 253,164-170(1991)
33. Shi J. Blundell, T. L. &Mizuguchi, K. (2001) “FUGUE:Sequence-structure homology recognition using environment- specific substitution tables and structure- dependent gap penalties ’’Journal of Molecular Biology,310, 243-257(2001)
34. Han, K. F. &Baker, D. (1995) -Recurring local sequence motifs in proteins. Journal of Molecular Biology, 251(1), 176-187
35. Shortle, D., Simons, K. T., Ruczinski, I. Fox, B. A. Bystroff C., & Baker, D. (1998) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scroning functions. Journal of molecular Biology, 268(1), 209-225.
36. Simons, K. T., Bonneau, R., Puczinski , I., & Baker, D. (1999). Ab initio protein structure prediction of CASP III targets using ROSETTA. Proteins: Structure Function, and Genetics, 37(S3), 171-176.https:// doi.org.10.1002/(SICI) 1097-0134(1999) 37:3+/
37. Moult, J., Fidelis, K., Kryshtafovych, A., & Tramontano, A. (2018) Critical Assessment of methods of protein structure prediction (CASP) -Round XII. Proteins:Structure, Function and Bioinformatics, 86(Suppl1).7-15
38. Torrisi, M., Kaleel, M., & Pollastri, G. (2018) Porter 5:Fast, State of the art ab initio prediction of protein secondary structure in 3 and 8 classes. Bio Rxiv. Preprint <https://doi.org/10.1101/2018.06.18.160403>
39. Telfer, W. H., & Williams, C. M. (1953). Immunological studies on insect metamorphosis. I. Qualitative and quantitative description of the blood antigens of the cercropia silkworm. Journal of General Physiology, 36(3), 389-413
40. Xia etal. (2004) “A Draft sequence for the Genome of the Domesticated silkworm ( *Bombyx* *mori*) ’’Science, 306(5703), 1937-1940
41. Xia, Q., Li. S., Feng. Q. (2008) “From genome to proteome : great progress in the domesticated silkworm ( *Bombyx mori*) ” ABBS, 40(07), 601-607.
42. Hou Y. etal., 2007,Zhang P. etal(2007) Studies on middle and posterior silk glands of silkworm ( *Bombyx* *mori*) using two-dimensional electrophoresis and mass spectrometry .Insect Biochemistry and Molecular Biology 37(5), 486-496
43. Zhang P., Aso, Y., Jikuys, H., Kusakabe T., Lee, J. M., Kawaguchi, Y., Yamamoto, K., Banno, Y., & Fuji, H. (2007) Proteomic profiling of silkworm skeletal muscle proteins during larval – pupal metamorphosis. Journal of proteome Research, 6(6), 2295-2303.https://doi.org/10.1021/pr0700714
44. Dash, R., Mukherjee, S. & Kundu, S. C. (2006). Isolation, Purification and Characterization of silk protein sericin from cocoon peduncle s tropical tasar silkworms *Anthreaea mylitta*. International Journal of Biological macromolecules, 38(3-5) 255-258.doi.org:10.1016/j.ijibiomac.2006.03.00
45. Deng, Z, Zhao, P., Wang. C., Zhang, Y., Chen, J., Wang, X., Lin Y., &Xia, Q. 2013.Comparative proteomics reveal diverse functions and dynamic changes of *Bombyx* *mori* silk proteins spun from different development stages. Journal of proteome Research. 12(11), 5213-5222 <https://doi.org/10.1021/pr4005772>
46. Cox, J., & Mann, M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b-range mass accuracies and proteome – wide protein quantification. Nature Biotechnology 26(12), 1367-1372.https://doi.org/10.1038/nbt.1511
47. Zhang, Y., Zhao, P., Dong, Z., Wang, D., Guo, P., Guo, X., Zhang, W., &Xia, Q. (2015) Comparative proteome analysis of multi layer cocoon of the silkworm Bombyx mori. PLoS ONE 10(4), e0123403 https:doi.org/10.1371/journal.pone.0123403
48. Berman, H. M., Westbrook, J., Feng. Z., Gilliland, G., Bhat, T. N. Weissig, H., Shindyalov, I. N., &Bourne, P. E. (2000) The protein Data Bank. Nucleic acids research, 28(1), 235-242.https://doi.org/10.1093/nar/28.1.2235
49. Jeffery.C.J.(1999). Moonlighting proteins Trends in Biochemical Sciences. 24(1), 8.11.https://doi.org/10.1016/s0968-0004(98) 01335-8
50. Jeffery, C. J. (2017). Protein moonlighting What is it and why is it important? Philosophical Transactions of the Royal Society B. Biological Sciences, 373(1738) 20160523.https://doi.org/10.1098/rstb.2016.0523
51. Commichau, F. M. & Stulke, J. (2015) Trigger enzymes: Coordination of metabolism and virulence gene expression. Microbiology Spectrum, 3(4). <https://doi.org/10.1128/> microbiolspec. MBP-0010-2014
52. You, R., Zhang, Z., Xiong, Y., Sun, F., Mamitsuka, H., & Zhu, S., (2018) G O Labeler : Improving sequence based large scale protein function prediction by learning to rank. Bioinformatics, 34(14), 2465-2473. [https://doi.org/10.10931 bioinformatics/ bty130](https://doi.org/10.10931%20bioinformatics/%20bty130)
53. Beneta Valentino, R., Ebejer, J. P., & Valentino, G. (2023) machine learning using neural networks for metabolomic pathway analyses methods in Molecular Biology, 2553,395-415 <https://doi.org/10.1007/978-1-0716-2617-7-17>
54. Tang, X, Liu, Y., Zhang, W., etal(2022) Progressive assembly of multi domain protein structures from cryo- EM density maps. Nature methods, 19(12), 1472-1480
55. Lobanov, M. Y., Boqatyreva, N. S., & Galzitskaya, O. V. (2008). Radius of gyration as an indicator of protein structure compactness. Molecular Biology, 42(4), 623-628.https://doi.org/10.1134/s0026893308040195
56. Bowie, J. U., Luthy, R., & Eisenberg, D., (1991). A method to identify protein sequences that fold into a known three – dimensional structure. Science, 253(5016), 164-170.https:// doi. Org /101126/science1853201
57. Jones, D. T. Taylor, W. R., &Thornton, J. M. (1992). A new approach to protein fold recognition. Nature, 358(6381), 86-89.https://doi.org/10.1038/358086a0
58. Jones, D. T. (1999). Gen THREADER: an efficient and reliable protein fold recognition method for genomic sequences. Journal of Molecular Biology, 287(4), 797-815, <https://doi.org/10.1006/jmbi.1999.2638>
59. Chothia, C., (1992) One thousand families for the molecular biology. Nature. 357(6379), 543-544 <https://doi.org/10.1038/3575-43a0>
60. Zhang, Y., &Skolnick, J. (2005). TM-align:A protein structure alignment algorithm based on the TM- score. Nucleic acids research, 33(7), 2302-2309.https://doi.org/10.1093/ nar/gki524
61. Brunet, M. A., Leblanc, S., &Roucou, X. (2020). Reconsidering proteomic diversity with functional investigation of small ORFs and alternative ORFs. Experimental cell Research, 393(1), 112057 <https://doi.org/10.1016/j.yexcr.2020.112057>
62. Radivojac, P., Clark, W. T., Iron, T. R., Schnoes, A. M., Wittkop, T., Sokolov, A., & Obradovic, Z., (2013) A large scale evaluation of computational protein function prediction. Nature methods, 10(3) 221-227. https:// doi. Org//10.1038/nmeth.2340
63. Jiang, Y., Clark, W. T. Friedman, C., & Radivojac, P. (2016). An expanded evaluation of protein function prediction methods shows improvement in accuracy Genome Biology, 17,184.https:// doi.org/10.1186/s13059-016-1002-4
64. Zhou, N, Jiang, Y., Bergquist, T. R., Lee, A. J., Kacsoh, B. Z., Crocker, A. W.,… . &Radivojac, P. (2019) The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. Genome Biology, 20,244. <https://doi.org/10.1186/s13059-019-1835-8>
65. AlQuraishi, M. (2019) End - to- end differentiable learning of protein structure cell systems, 8(4), 292-301.e3 . <https://doi.org/10.1016/j.cels.2019.03.006>
66. Ismi, D. P., Pulungan, R., &Afiahayati.(2022) Deep learning for protein secondary structure prediction: Pre and post – Alphafold. Computational and Structural Biotechnology Journal, 20,6271-6286 . <https://doi.org/10.1016/j.csbj.2022.11.012>
67. Zmasek, C. M., & Godzik, A. (2011). Strong functional patterns in the evolution of eukaryotic genomes revealed by the reconstruction of ancestral protein domain reportoires. Genome Biology, 12,R4.https://doi.org/10.1186/gb-2011-12-1-r4
68. Summers, N. L., & Karplus, M. (1989). Construction of side-chains in homology modelling: Application to the C-terminal lobe of rhizopuspepsin. Journal of Molecular Biology, 210(4), 785–811. [https://doi.org/10.1016/0022-2836(89)90109-5](https://doi.org/10.1016/0022-2836%2889%2990109-5)
69. Niazi, S., Mariam, Z., & Paracha, R. (2024). Limitations of protein structure prediction algorithms in therapeutic protein development. BioMedInformatics, 4(1), 98–112. <https://doi.org/10.3390/biomedinformatics4010007>
70. Xu, Y., Liu, Z., Cai, L., Xu, D. (2007). Protein Structure Prediction by Protein Threading. In: Xu, Y., Xu, D., Liang, J. (eds) Computational Methods for Protein Structure Prediction and Modeling. Biological and Medical Physics, Biomedical Engineering. Springer, New York, NY. <https://doi.org/10.1007/978-0-387-68825-1_1>
71. Sali A, Shakhnovich E, Karplus M. How does a protein fold? Nature 1994:369:248-51

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. <https://doi.org/10.1093/nar/25.17.3389>

1. Simons, K. T., Bonneau, R., Ruczinski, I., & Baker, D. (1999). Ab initio protein structure prediction of CASP III targets using ROSETTA. Proteins: Structure, Function, and Genetics, 37(S3), 171–176. https://doi.org/10.1002/(SICI)1097-0134(1999)37:3+<171::AID-PROT21>3.0.CO;2-Z
2. Zhou, H., & Zhou, Y. (2004). Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. Proteins: Structure, Function, and Bioinformatics, 55(4), 1005–1013. <https://doi.org/10.1002/prot.2016>
3. Zhang, Y. (2008). I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics, 9*, 40.
4. Xu Y., Liu, Z., Cai, L., Xu, D. (2007). Protein Structure Prediction by Protein Threading. In: Xu, Y., Xu, D., Liang, J. (eds) Computational Methods for Protein Structure Prediction and Modeling. Biological and Medical Physics, Biomedical Engineering. Springer, New York, NY. <https://doi.org/10.1007/978-0-387-68825-1_1>
5. Senior, A. W., Evans, R., Jumper, J., & Green, T. (2020). Improved protein structure prediction using potentials from deep learning. Nature, 577(7792), 706-710
6. Ingraham, J., Garg, V. K., Barzilay, R., & Jaakkola, T. (2019). Learning protein structure with a differentiable simulator. In International Conference on Learning Representations. <https://openreview.net/forum?id=Byg3y3C9Km>
7. AlQuraishi, M. End-to-End Differentiable Learning of protein structure. Cell Syst. 2019,8,292-30.e.3
8. Mao, W., Ding, X., Yuan, Y., & Xu, J. (2019). AmoebaContact and GDFold as a pipeline for rapid de novo protein structure prediction. Nature Machine Intelligence, 1(8), 490–496. <https://doi.org/10.1038/s42256-019-0130-4>
9. Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S., & Baker, D. (2020). Improved protein structure prediction using predicted interresidue orientations. Proceedings of the National Academy of Sciences, 117(3), 1496–1503. <https://doi.org/10.1073/pnas.1914677117>
10. Ju, F., Wang, Z., & Xu, J. (2021). CopulaNet: Learning residue co-evolution directly from multiple sequence alignment for protein structure prediction. Nature Communications, 12, 2535. <https://doi.org/10.1038/s41467-021-22869-8>
11. Jumper, J., Evans, R., Pritzel, A., & Green, T. (2021). Highly accurate protein structure prediction with Alphafold. Nature, 596(7873), 583-589.
12. Baek, M., Di Maio, F. Anischenko, I., & Zhang, J. (2021). Accurate prediction of protein structures and interactions using a three track neural network. Science, 373(6557), 871-876
13. Brandes, N., Ofer, D., Peleg, Y., Rappoport, N., & Linial, M. (2022). ProteinBERT: A universal deep-learning model of protein sequence and function. Bioinformatics, 38(8), 2102–2110. <https://doi.org/10.1093/bioinformatics/btac020>
14. Chowdhury R, Bouaua N. Biswas S, Floristean C., Kharkare A, Roye K, et al. Single-sequence protein structure prediction using a language model and deep learning. National Biotechnology 2022:40;1617-23
15. Yu, J., Wang, S., & Xu, J. (2022). Accurate and efficient protein sequence design through learning concise local structure representations. Bioinformatics, 39(3), btad122. <https://doi.org/10.1093/bioinformatics/btad122>
16. Wu, Y, Q., Chen, Y. Shen, Y.; Cao, Y., Hu, G., Cui, W., Gai, J., Zheng, W. Recent Progress of Protein Tertiary Structure Prediction. Molecules 2024
17. Kandathil, S. M., Greener, J. G., Lau, A. M., & Jones, D. T. (2022). Ultrafast end-to-end protein structure prediction enables high-throughput exploration of uncharacterized proteins. Proceedings of the National Academy of Sciences, 119(6), e2113348119. <https://doi.org/10.1073/pnas.2113348119>
18. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., … & Sercu, T. (2022). Evolutionary-scale prediction of atomic-level protein structure with a language model. Science, 376(6594), 1327–1333. <https://doi.org/10.1126/science.ade2574>
19. Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., Bodenstein, S. W., Evans, D. A., Hung, C.-C., O’Neill, M., Reiman, D., Tunyasuvunakool, K., Wu, Z., Žemgulytė, A., Arvan hiiiiti, E., Beattie, C., ... Jumper, J. M. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature, 630*(8025), 493–500. <https://doi.org/10.1038/s41586-024-07487-w>
20. Nakai, K. and Kanehisa, M., Expert system for predicting protein localization sites in Gram-negative bacteria, PROTEINS: Structure, Function, and Genetics 11, 95-110 (1991).
21. Nakai, K. and Kanehisa, M., A knowledge base for predicting protein localization sites in eukaryotic cells, Genomics 14, 897-911 (1992).
22. Nakai, K. and Horton, P., PSORT: a program for detecting the sorting signals of proteins and predicting their subcellular localization, Trends Biochem. Sci, 24(1) 34-35 (1999).
23. Gardy, J. L., Spencer, C., Wang, K., Ester, M., Tusnády, G. E., Simon, I., Hua, S., deFays, K., Lambert, C., Nakai, K., & Brinkman, F. S. (2003). PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria. Nucleic acids research, 31(13), 3613–3617. <https://doi.org/10.1093/nar/gkg602>
24. Bannai, H., Tamada, Y., Maruyama, O., Nakai, K., & Miyano, S. (2002). Extensive feature detection of N-terminal protein sorting signals. Bioinformatics, 18(2), 298–305. <https://doi.org/10.1093/bioinformatics/18.2.298>
25. Emanuelsson, O., Nielsen, H., Brunak, S., & von Heijne, G. (2000). Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. Journal of molecular biology, 300(4), 1005–1016. <https://doi.org/10.1006/jmbi.2000.3903>
26. Bhasin, M., & Raghava, G. P. (2004). ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. Nucleic acids research, 32(Web Server issue), W414–W419. <https://doi.org/10.1093/nar/gkh350>
27. Reinhardt, A., & Hubbard, T. (1998). Using neural networks for prediction of the subcellular location of proteins. Nucleic Acids Research, 26(9), 2230–2236. <https://doi.org/10.1093/nar/26.9.2230>
28. Hua, S., & Sun, Z. (2001). Support vector machine approach for protein subcellular localization prediction. Bioinformatics (Oxford, England), 17(8), 721–728. Doi <https://doi.org/10.1093/bioinformatics/17.8.721>