1

Introduction

1.1 Background and Motivation

Proteins are the molecular machines that power life itself. Every cell in a living organism contains a vast array of proteins, each responsible for specific tasks, from facilitating chemical reactions to structural integrity, and regulating gene expression. The study of proteins is essential for understanding the fundamental processes of life, ranging from cellular metabolism to disease pathology. At the molecular level, proteins are composed of long chains of amino acids that fold into specific three-dimensional structures, a process known as protein folding (Ptitsyn, 1991; Richardson and Richardson, 1992). The unique shape of a protein determines its functionality, as only a specific conformation allows it to interact with other molecules, catalyze biochemical reactions, and maintain cellular processes (Figure 1.1).

However, despite the critical role of proteins in cellular function, a major challenge in molecular biology remains: understanding how proteins achieve their three-dimensional shapes and how mutations in these structures can lead to diseases. For decades, researchers have attempted to predict protein structures based on their amino acid sequences, but this task has proven to be extraordinarily complex. The sequence of amino acids in a protein is like a string of letters in an alphabet, yet the way these letters arrange themselves into a specific shape is governed by intricate physical and chemical interactions that are not immediately obvious from the sequence alone.

In the past, the understanding of protein structures relied heavily on experimental methods such as X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and cryo-electron microscopy (cryo-EM). These techniques can provide high-resolution information on the structure of proteins, but they are time-consuming, expensive, and often require high-quality samples, which are not always available. Moreover, they struggle to capture the dynamic nature of proteins, which constantly change shape during their interactions with other molecules. These challenges have led researchers to seek out computational approaches that can predict protein structure from sequence, simulate protein dynamics, and investigate the effects of mutations on protein function (Figure 1.2).

Machine Learning in Protein Science: Efficient Prediction of Protein Structures and Properties,
First Edition. Jinjin Li and Yanqiang Han.

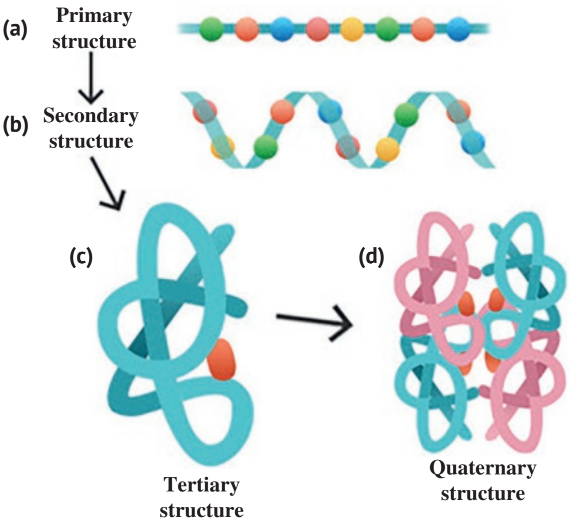

Figure 1.1 (a) The primary structure of a protein can be understood as a linear string. (b) The secondary structure refers to how the peptide chain undergoes twists, folds, and other transformations based on the string of the primary structure, forming a local three-dimensional structure. (c) The tertiary structure is the process of splicing multiple secondary structures together and folding them into a complete three-dimensional protein structure. (d) A quaternary structure refers to the combination of multiple tertiary molecules into a complex.

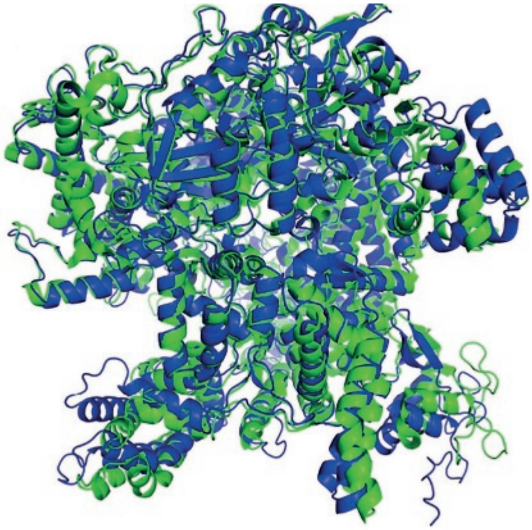

Figure 1.2 The three-dimensional structural model of proteins is usually predicted by bioinformatics software based on the amino acid sequence of proteins or analyzed through experimental methods such as X-ray crystallography, nuclear magnetic resonance (NMR), or cryo-electron microscopy. Different colors represent different secondary structures of proteins.

Computational protein biology has seen immense progress in recent years. The development of new algorithms and the exponential growth of computational power have paved the way for the application of more efficient techniques. Among the most groundbreaking advancements in this field is the application of machine learning (ML) and artificial intelligence (AI) to predict protein structures and functions (Jumper et al., 2021; Rives et al., 2021). The ability to predict a protein's structure from its sequence without the need for experimental data has been one of the "holy grails" of computational biology. ML models, particularly those based on deep learning techniques, have shown immense promise in this area, outperforming traditional methods in accuracy and speed. One of the most notable breakthroughs in this domain is AlphaFold, a deep learning algorithm developed by DeepMind. AlphaFold's ability to predict protein structures with near-experimental accuracy has revolutionized the field and demonstrated the potential of AI-driven approaches in protein science (Figure 1.3).

The success of AlphaFold (Jumper et al., 2021), which has been heralded as a major milestone in structural biology, highlights the potential of ML to solve long-standing problems in computational biology. AlphaFold uses deep neural networks trained on vast datasets of known protein structures to predict the three-dimensional structure of proteins based on their amino acid sequences. The algorithm has achieved unprecedented levels of accuracy, solving the protein folding problem for a wide range of proteins with remarkable precision. AlphaFold's success has provided a glimpse into the future of protein research, where ML models can be used not only to predict protein structure but also to simulate protein function, understand the effects of mutations, and design novel proteins with desired properties.

Despite the significant strides made in protein structure prediction, there remain several challenges that need to be addressed. While AlphaFold's algorithm is capable of predicting the structure of individual proteins, the prediction of

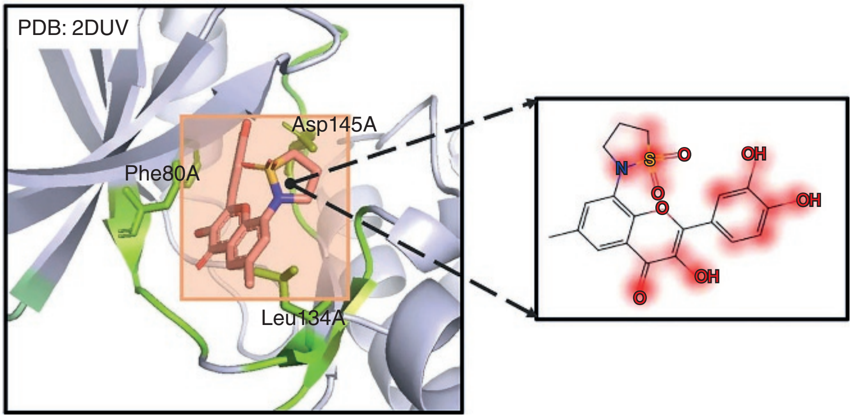

Figure 1.3 The detailed structure of the binding sites between one drug molecule and a protein molecule demonstrates how drugs interact with proteins, which is crucial for drug design and understanding protein function.

protein–protein interactions (PPIs), protein–ligand binding, and the dynamic behavior of proteins in complex biological environments is still an open problem. These processes are crucial for understanding cellular signaling pathways, enzyme catalysis, and drug design (Krasner, 1972). In particular, predicting how proteins interact with one another and how their structures change in response to different conditions is a complex task that requires a deeper understanding of the molecular forces at play. Moreover, protein interactions often occur in crowded cellular environments, making it difficult to model these interactions accurately using traditional computational methods (Zheng et al., 2020).

Furthermore, the impact of mutations on protein structure and function remains a significant challenge. Mutations in DNA can lead to changes in the amino acid sequence of a protein, which in turn may alter its structure and function. Some mutations can lead to loss of function, while others may result in gain of function, causing diseases such as cancer, neurodegenerative disorders, and genetic diseases. Being able to predict the effects of mutations on protein structure and function is crucial for understanding disease mechanisms and developing therapeutic strategies. Although ML models have shown promise in predicting the effects of mutations, there is still much to be done in terms of improving the accuracy and robustness of these predictions.

In addition to structure and mutation prediction, protein function annotation remains one of the most important challenges in bioinformatics. While the genome sequencing revolution has provided us with vast amounts of sequence data, the function of many proteins remains unknown. The process of assigning a biological function to a protein based on its sequence is known as function annotation. Traditionally, function annotation has relied on experimental techniques, such as gene knockout experiments, to determine the role of a protein in a biological context. However, these methods are time-consuming and expensive. Computational methods, particularly those based on ML, have the potential to accelerate the process of function annotation by predicting the biological role of a protein based on its sequence, structure, or interaction with other molecules.

The need for accurate, high-throughput methods for protein function annotation has become even more urgent in the context of personalized medicine. With the increasing availability of genomic data, there is a growing demand for tools that can predict how genetic variations in individuals affect protein function. The ability to link specific genetic mutations to disease-causing proteins can provide valuable insights into the molecular basis of disease and guide the development of targeted therapies. In this regard, ML has the potential to revolutionize the way we approach drug discovery and personalized medicine by enabling the rapid identification of disease-related proteins and the design of therapies that target these proteins.

The integration of quantum mechanical calculations into protein research represents another promising avenue for improving the accuracy of protein predictions. Quantum mechanics, which describes the behavior of matter at the atomic and subatomic levels, provides a powerful framework for modeling the interactions between atoms and molecules. By applying quantum mechanical methods to protein systems, researchers can gain a deeper understanding of the forces that govern protein folding, stability, and interactions. Quantum mechanical calculations are particularly useful for studying the detailed electronic structure of proteins, including the behavior of electrons and the formation of chemical bonds. However, these calculations are computationally expensive and often require specialized software and hardware. As a result, they have been limited to small systems or simplified models. The challenge lies in developing methods that combine the accuracy of quantum mechanical calculations with the scalability needed to model large, complex proteins.

In recent years, the combination of quantum mechanics and ML has emerged as a promising strategy to overcome the computational limitations of traditional quantum mechanical methods (Peral-García et al., 2024). By using ML algorithms to predict the parameters required for quantum mechanical calculations, researchers can improve the efficiency and accuracy of these methods. For example, deep learning techniques have been used to predict the electronic structure of molecules, enabling researchers to simulate the behavior of proteins more efficiently. This hybrid approach has the potential to revolutionize the field of protein modeling by making high-level quantum mechanical calculations accessible for larger and more complex systems.

As protein research continues to evolve, new computational techniques are allowing for more nuanced simulations and predictions, accelerating our ability to explore protein behavior. One of the key areas of innovation is in the development of more efficient algorithms for protein structure prediction, PPIs, and protein–ligand binding. Traditional molecular dynamics simulations, which track the movement of atoms in a protein over time, have been essential in understanding protein folding and dynamics. However, these simulations are often limited by computational cost and the difficulty of modeling large protein complexes accurately over long time scales. Emerging hybrid techniques that combine molecular dynamics with ML models hold promise for overcoming these barriers. By using data-driven approaches to predict the behavior of molecular systems, these models can enhance the efficiency of simulations and provide deeper insights into protein functionality.

Moreover, the need for high-resolution predictions in the study of protein dynamics is particularly evident in understanding the mechanisms behind diseases caused by protein misfolding. Diseases like Alzheimer's, Parkinson's, and cystic fibrosis have been linked to specific protein misfolding events. These diseases are often the result of mutations that destabilize the protein structure, leading to aggregation, loss of function, or toxic gain-of-function effects. Predicting the structural consequences of mutations and understanding the resulting changes in protein behavior are crucial for identifying therapeutic targets. ML methods have the potential to revolutionize this process by identifying patterns in the sequence–structure–function relationships of proteins that were previously invisible to conventional methods.

At the same time, the integration of protein structure prediction with personalized medicine offers a unique opportunity to transform healthcare. As more individuals undergo genome sequencing and provide detailed information about their genetic predispositions, the need for predictive models that can link specific genetic variations to diseases becomes increasingly important. The ability to predict how individual mutations will affect protein folding, stability, and interactions could allow for more precise and personalized treatments, reducing the trial-and-error approach that is common in drug development today. ML driven techniques could enable the creation of drug candidates tailored to an individual's specific genetic makeup, optimizing treatment efficacy while minimizing adverse effects.

Beyond prediction, another important area of protein research is the de novo design of proteins with desired properties. Protein engineering has traditionally focused on modifying existing proteins for industrial or therapeutic purposes. However, the ability to design entirely new proteins from scratch offers an exciting frontier in biotechnology and synthetic biology (Huang et al., 2016). Advances in ML and computational modeling have made it increasingly feasible to design novel proteins with tailored functions, such as enzymes that catalyze specific reactions or therapeutic proteins that bind to disease-causing agents. The application of generative models and reinforcement learning algorithms to protein design is an area of active research, with the goal of creating proteins with unprecedented capabilities for use in medicine, agriculture, and environmental sustainability.

One of the most ambitious goals of protein science is to bridge the gap between computational predictions and experimental validation. While algorithms like AlphaFold have made substantial progress in accurately predicting protein structures, there is still a need for experimental validation to confirm these predictions in a laboratory setting. The integration of computational approaches with experimental data is crucial for improving the reliability of predictions and translating computational models into practical applications. For example, high-throughput experimental techniques, such as cryo-EM and mass spectrometry, could provide complementary data to refine computational models and validate predictions. This synergy between computational and experimental methods could pave the way for more rapid advancements in drug discovery, protein engineering, and our understanding of fundamental biological processes.

The integration of quantum mechanics with ML in protein research is also driving significant advancements in the field. Quantum mechanical calculations, which provide a detailed and accurate description of atomic interactions, have long been recognized for their potential to improve our understanding of protein behavior. However, these calculations are computationally expensive, limiting their applicability to small systems or simplified models. To address this, researchers have begun using ML algorithms to accelerate quantum mechanical calculations and extend them to larger, more complex systems. This hybrid approach allows for the accurate simulation of protein folding, stability, and interactions, facilitating the study of proteins that were previously too large or too dynamic to model with traditional quantum mechanics.

The application of deep neural networks to protein research is another area that has garnered significant attention. Deep learning models, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have shown promise in tasks such as protein structure prediction (Karwasra et al., 2024), mutation effect prediction, and function annotation. These models excel at recognizing patterns in large datasets, which is particularly useful for identifying relationships between protein sequences and their corresponding structures or functions. The ability to train these models on vast amounts of data has led to breakthroughs in the prediction of protein functions, particularly in cases where experimental data is limited or unavailable.

ML also holds the potential to improve drug discovery by identifying novel drug candidates (Vamathevan et al., 2019; Catacutan et al., 2024). In traditional drug discovery pipelines, researchers often screen thousands of compounds to find those that interact with a target protein. However, this process is time-consuming and costly. ML algorithms can significantly speed up this process by predicting the binding affinity of compounds to a target protein before experimental screening. These models can analyze vast chemical libraries and identify potential drug candidates based on their predicted interaction with the protein of interest. This approach not only saves time and resources but also enables the discovery of compounds with higher potency and fewer side effects.

Another exciting development in protein research is the application of graph-based ML models to protein design and analysis (Akid et al., 2024; Ingraham et al., 2019). Proteins can be represented as graphs, with nodes corresponding to amino acids and edges representing interactions between them. Graph neural networks (GNNs) have been used to predict protein structure, function, and interactions by learning from the graph representations of proteins. These models offer a new way to capture the complex relationships between amino acids in a protein and can be applied to tasks such as protein folding, PPI prediction, and protein–ligand docking (Réau et al., 2023; Knutson et al., 2022).

As computational power continues to grow and ML models become more sophisticated, the future of protein science looks increasingly promising. The integration of AI-driven approaches with quantum mechanical calculations, protein structure prediction, and protein design will likely accelerate the pace of discovery and open up new possibilities for therapeutic interventions. The ability to predict and design proteins with tailored functions will not only revolutionize drug discovery but also transform fields such as materials science, agriculture, and environmental sustainability (Figure 1.4).

In the near future, the collaboration between experimentalists and computational biologists will be key to advancing our understanding of protein systems. While computational predictions provide invaluable insights into protein structure and function, experimental validation remains essential for confirming these predictions and translating them into practical applications. By combining the power of AI and quantum mechanics with experimental techniques, the field of protein research is poised to make major breakthroughs in understanding the molecular basis of life, improving human health, and addressing global challenges (Figure 1.5).

The rise of computational tools in protein research has been fueled by several factors, including advancements in computational power, the availability of large biological datasets, and the increasing recognition of the importance of protein systems in health, disease, and biotechnology. Historically, the study of proteins has relied heavily on experimental methods such as X-ray crystallography, NMR, and cryo-EM, each of which has its limitations. For example, X-ray crystallography requires proteins to be crystallized, which is often challenging for membrane proteins or large complexes. NMR, on the other hand, is limited by the size of the proteins that can be studied and requires high concentrations, making it less suitable for some biological systems. Cryo-EM has emerged as a powerful method for studying large,

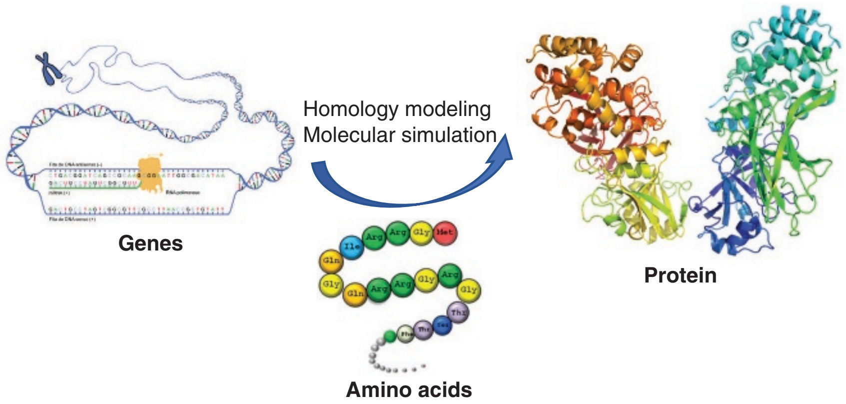

Figure 1.4 This image illustrates the bioinformatics analysis process from genes to protein structures. On the left is a diagram of genes, representing DNA sequences that contain genetic information. The middle section displays the amino acid sequence, which is the basic unit transcribed from DNA into mRNA and then translated into proteins during gene expression. On the right is the three-dimensional structure of the protein, predicted through homology modeling and molecular simulation techniques. Homologous modeling is a method of predicting unknown protein structures based on known protein structures, while molecular simulation is used to study the dynamic behavior of proteins at the atomic level. This process is of great significance for understanding the function of proteins and designing new drugs.

dynamic proteins, but it requires specialized equipment and is still relatively time-consuming. These experimental approaches, though invaluable, are often expensive, labor-intensive, and time-consuming.

Computational approaches, in contrast, provide a complementary toolset for protein research. The ability to simulate the behavior of proteins and predict their structures, functions, and interactions without the need for expensive experimental setups has opened new possibilities in protein science. Computational models, such as molecular dynamics simulations and quantum mechanical calculations, offer insights into protein folding, stability, and function on time scales and at resolutions that were previously unattainable. Moreover, these models can be used to test hypotheses about protein behavior before conducting costly experiments, saving both time and resources.

ML has become an increasingly powerful tool in this computational arsenal. By training algorithms on large datasets of protein sequences, structures, and functions, ML models can identify complex patterns that traditional methods cannot easily detect. These models have demonstrated remarkable success in predicting protein structures (e.g., AlphaFold), estimating the effects of mutations on protein function, and identifying novel protein–ligand interactions for drug discovery. The use of ML also allows researchers to develop predictive models that can generalize across a wide variety of proteins, making it easier to study proteins with unknown functions or to explore new avenues for drug development. One of the key advantages of ML is its ability to integrate diverse sources of data – ranging from genomic sequences to experimental structural data – into a unified model, enabling researchers to make more accurate predictions and derive new biological insights.

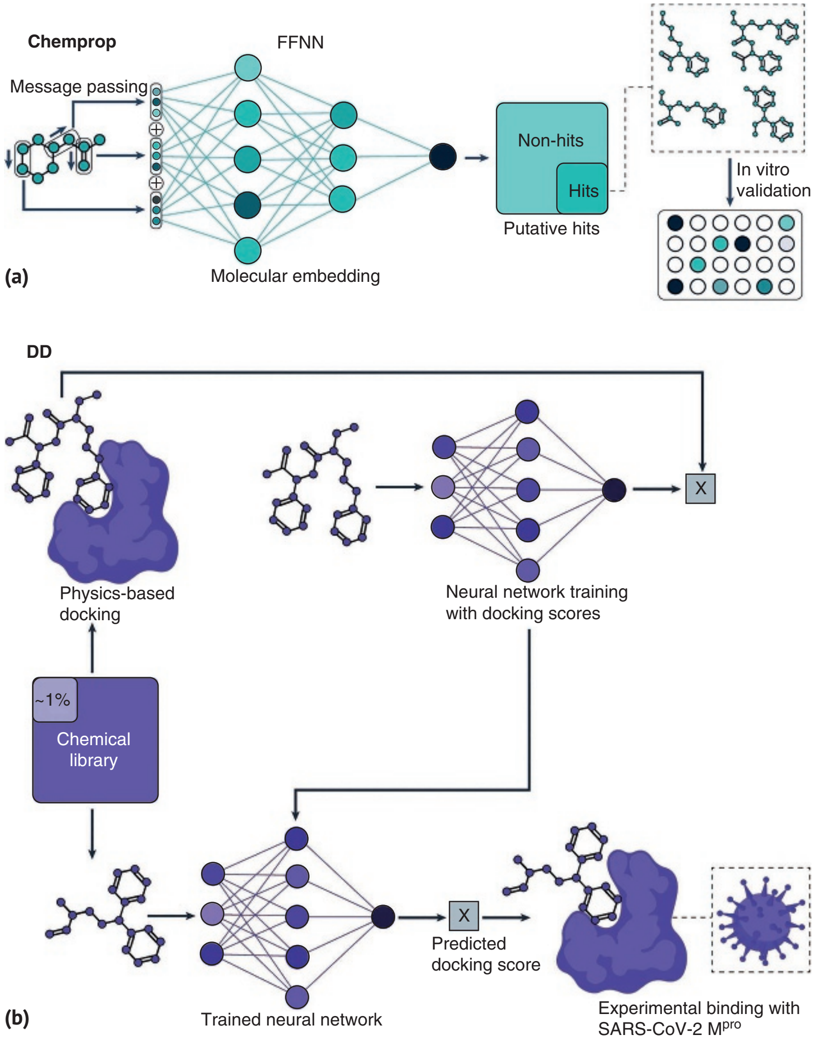

Figure 1.5 This illustration showcases the application of machine learning in drug design. (a) The role of message-passing neural networks in hit discovery. Through a series of messaging steps, molecular embeddings rich in contextual information are generated, capturing the connectivity among atoms. Upon aggregation, these vectors are fed into a feedforward neural network (FFNN) to yield predicted property values. Based on user-defined thresholds, compounds may subsequently be prioritized for in vitro potency validation. (b) Deep Docking (DD) synergizes physics-based docking methods with deep learning to evaluate vast chemical libraries, uncovering their potential to bind with specific proteins. Initially, a fraction of the chemical library (approximately 1%) undergoes physics-based docking, with the resulting scores utilized for training the FFNN, which then swiftly predicts docking scores for the remaining compounds (around 99%).

The application of ML and computational methods is particularly important in the context of personalized medicine. As we move toward an era of precision healthcare, where treatments are tailored to an individual's genetic makeup, the ability to predict how genetic variations will affect protein function is becoming crucial. Genetic mutations can have profound effects on protein structure and function, potentially leading to diseases or altered responses to drugs. By leveraging large-scale genomic and proteomic data, ML algorithms can identify which mutations are likely to be pathogenic, how they affect protein stability, and what therapeutic strategies may be effective in mitigating these effects. This opens up the possibility of designing personalized treatments that are more effective and have fewer side effects.

In addition to the practical benefits for medicine, the computational study of protein systems has significant implications for biotechnology, agriculture, and environmental sustainability. For example, proteins play a key role in many industrial processes, including the production of biofuels, enzymes, and pharmaceuticals. The ability to design proteins with specific functions – such as enzymes that catalyze desired chemical reactions or proteins that bind to toxic compounds – could greatly enhance the efficiency of these processes. In agriculture, proteins that are resistant to pathogens or pests could be engineered to improve crop yields and reduce the need for chemical pesticides. In the environmental sector, proteins could be designed to break down pollutants or to capture carbon, contributing to efforts to combat climate change (Figure 1.6).

The need for accurate predictions in the field of protein systems extends beyond traditional protein folding and structure prediction. Protein interactions are essential for nearly every biological process, from signaling pathways to immune responses. Understanding how proteins interact with each other and with small molecules (such as drugs or other ligands) is fundamental to drug discovery, systems biology, and synthetic biology. Proteins often work in large, dynamic networks, where the binding of one protein can have cascading effects on the entire system. Predicting these interactions, especially in complex biological systems, has been a long-standing challenge in computational biology. Recent advances in ML and graph-based methods, however, offer new ways to model PPIs, protein–ligand interactions, and protein–small molecule binding. These methods can leverage the structural information of individual proteins and the network of interactions to predict how proteins will behave in various biological contexts.

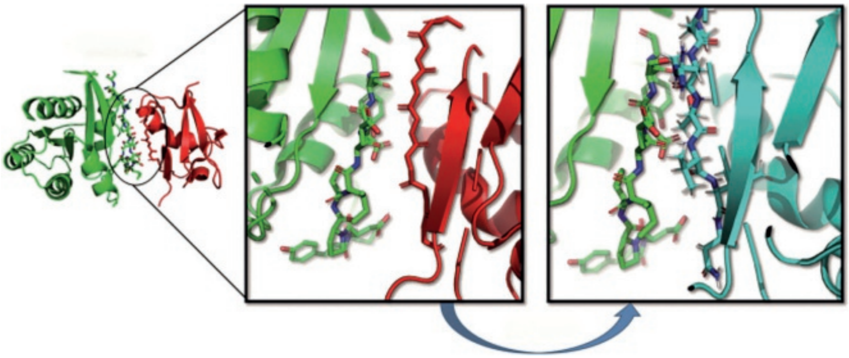

Figure 1.6 This image shows the dynamic changes in protein structure. This change is crucial for understanding the function, stability, and how proteins interact with other molecules. By studying these dynamic changes, scientists can better understand the working mechanisms of proteins in living organisms.

The convergence of ML, quantum mechanical calculations, and experimental biology is poised to revolutionize the field of protein research. While ML models are powerful in making predictions, they are still limited by the quality and quantity of data they are trained on. Experimental methods, however, provide the real-world data needed to validate and refine these predictions. Integrating these two approaches, experimental data and computational models, could significantly accelerate the discovery of new protein functions, the development of new drugs, and the design of novel proteins for biotechnological applications. For instance, experimental techniques such as X-ray crystallography, NMR, and cryo-EM can generate high-resolution structural data that can be used to train ML models and improve the accuracy of structure prediction. Similarly, high-throughput screening methods can generate large datasets that can be used to train models for drug discovery, protein–ligand binding prediction, and mutation effect prediction.

One particularly exciting area of research is the application of quantum mechanical calculations in protein science. Quantum mechanics provides a level of detail that is unmatched by classical simulations, offering insights into the atomic-level interactions between molecules. However, quantum mechanical calculations are computationally expensive, especially when applied to large, complex proteins. Recent developments in ML and hybrid approaches, which combine quantum mechanics with data-driven models, promise to make quantum mechanical simulations more accessible and applicable to larger protein systems. This approach, often referred to as deep learning-assisted full-system quantum mechanical (FQM) calculations (Li et al., 2021), has the potential to improve the accuracy of protein simulations, enabling researchers to study proteins in more detail than ever before.

As the field progresses, it is clear that the future of protein research will be shaped by the continued integration of computational and experimental approaches. The growing availability of large biological datasets, advances in computational algorithms, and the increasing power of ML models will accelerate our understanding of protein systems. Furthermore, the application of quantum mechanical calculations, deep learning, and transfer learning will expand the range of proteins that can be studied, providing new insights into their functions and interactions. The ultimate goal of this research is to harness the power of computational models to predict and design proteins with desired functions, opening up new possibilities for drug discovery, protein engineering, and biotechnology.

In summary, the study of protein systems is a dynamic and rapidly evolving field that stands at the intersection of biology, chemistry, physics, and computer science. The integration of computational tools with experimental techniques is transforming our understanding of proteins and their roles in biological processes. As computational power continues to increase and new algorithms are developed, the potential for breakthrough discoveries in medicine, biotechnology, and environmental science grows exponentially. The future of protein research holds great promise, with ML, quantum mechanical calculations, and protein design leading the way toward a deeper understanding of the molecular basis of life.

Bibliography

Hajer Akid, Kirsley Chennen, Gabriel Frey, Julie Thompson, Mounir Ben Ayed, and Nicolas Lachiche. Graph-based machine learning model for weight prediction in protein–protein networks. *BMC Bioinformatics*, 25(1): 349, November 2024. ISSN 1471-2105. https://doi.org/10.1186/s12859-024-05973-6.

Denise B. Catacutan, Jeremie Alexander, Autumn Arnold, and Jonathan M. Stokes. Machine learning in preclinical drug discovery. *Nature Chemical Biology*, 20(8): 960–973, August 2024. ISSN 1552-4469. https://doi.org/10.1038/s41589-024-01679-1.

Po-Ssu Huang, Scott E. Boyken, and David Baker. The coming of age of de novo protein design. *Nature*, 537(7620): 320–327, September 2016. ISSN 1476-4687. https://doi.org/10.1038/nature19946.

John Ingraham, Vikas Garg, Regina Barzilay, and Tommi Jaakkola. Generative models for graph-based protein design. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873): 583–589, August 2021. ISSN 1476-4687. https://doi.org/10.1038/s41586-021-03819-2.

Ritu Karwasra, Kushagra Khanna, Kapil Suchal, Ajay Sharma, and Surender Singh. Chapter 13 - protein structure prediction with recurrent neural network and convolutional neural network: a case study. In Khalid Raza, Debmalya Barh, Deepak Singh, and Naeem Ahmad, editors, *Deep Learning Applications in Translational Bioinformatics*, volume 15 of *Advances in Ubiquitous Sensing Applications for Healthcare*, pp. 211–229. Academic Press, January 2024. https://doi.org/10.1016/B978-0-443-22299-3.00013-X.

Carter Knutson, Mridula Bontha, Jenna A. Bilbrey, and Neeraj Kumar. Decoding the protein–ligand interactions using parallel graph neural networks. *Scientific Reports*, 12(1): 7624, May 2022. ISSN 2045-2322. https://doi.org/10.1038/s41598-022-10418-2.

Joseph Krasner. Drug-protein interaction. *Pediatric Clinics of North America*, 19(1): 51–63, February 1972. ISSN 0031-3955. https://doi.org/10.1016/S0031-3955(16)32666-9.

Wei Li, Haibo Ma, Shuhua Li, and Jing Ma. Computational and data driven molecular material design assisted by low scaling quantum mechanics calculations and machine learning. *Chemical Science*, 12(45): 14987–15006, 2021. https://doi.org/10.1039/D1SC02574K.

David Peral-García, Juan Cruz-Benito, and Francisco José García-Peñalvo. Systematic literature review: Quantum machine learning and its applications. *Computer Science Review*, 51: 100619, February 2024. ISSN 1574-0137. https://doi.org/10.1016/j.cosrev.2024.100619.

O.B. Ptitsyn. How does protein synthesis give rise to the 3D-structure? *FEBS Letters*, 285(2): 176–181, 1991. ISSN 1873-3468. https://doi.org/10.1016/0014-5793(91)80799-9.

Manon Réau, Nicolas Renaud, Li C Xue, and Alexandre M J J Bonvin. DeepRank-GNN: a graph neural network framework to learn patterns in protein–protein interfaces. *Bioinformatics*, 39(1): btac759, January 2023. ISSN 1367-4811. https://doi.org/10.1093/bioinformatics/btac759.

David C. Richardson and Jane S. Richardson. The kinemage: a tool for scientific communication. *Protein Science*, 1(1): 3–9, 1992. ISSN 1469-896X. https://doi.org/10.1002/pro.5560010102.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences of the United States of America*, 118(15): e2016239118, April 2021. https://doi.org/10.1073/pnas.2016239118.

Jessica Vamathevan, Dominic Clark, Paul Czodrowski, Ian Dunham, Edgardo Ferran, George Lee, et al. Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery*, 18(6): 463–477, June 2019. ISSN 1474-1784. https://doi.org/10.1038/s41573-019-0024-5.

Shuangjia Zheng, Yongjian Li, Sheng Chen, Jun Xu, and Yuedong Yang. Predicting drug–protein interaction using quasi-visual question answering system. *Nature Machine Intelligence*, 2(2): 134–140, February 2020. ISSN 2522-5839. https://doi.org/10.1038/s42256-020-0152-y.