European Journal of Sustainable Development Research
2025, 9(3), em0298
e-ISSN: 2542-4742
https://www.ejosdr.com/

MODESTUM

Review Article

OPEN ACCESS

Bioinformatics tools in protein analysis: Structure prediction, interaction modelling, and function relationship

Taiwo Temitope Ogunjobi [1*], Ijeoma Chineme Okorie [2], Chimaobi Divine Gigam-Ozuzu [3], Jumoke Victoria Olorunleke [4], Felix Iyanu Ogunleye [5], Emmanuella Osaruese Irimoren [6], Dorcas Oyedolapo Atanda [7], Adaobi Mary-Ann Okafor [8], Chinyere Eucharia Agbo [8], Favour Onasokhare Okunbi [9], Otoh Dayo Umoren [10], Adoyi Daniel Adidu [11], Emmanuel Oluwadamilare Ojo [12]

[1] Department of Biochemistry, Faculty of Basic Medical Sciences, Ladoke Akintola University of Technology, Ogbomosho, Oyo State, NIGERIA
[2] Dapartment of Mathematics Sciences, Faculty of Science, NDA, Kaduna, Kaduna State, NIGERIA
[3] Department of Plant Science and Biotechnology, School of Natural and Applied, Science, University of Port Harcourt, Choba, Rivers State, NIGERIA
[4] Department of Pharmaceutical Microbiology, College of Pharmacy, Obafemi Awolowo University, Ile-Ife, Osun State, NIGERIA
[5] Department of Biomedical Technology, School of Basic Medical Science, Federal University of Technology Akure, Akure, Ondo State, NIGERIA
[6] Department of Anatomy, Faculty of Basic Medical Sciences, University of Benin, Benin City, Edo State, NIGERIA
[7] Department of Biochemistry, Faculty of Sciences, Lagos State University, Ojo, Lagos State, NIGERIA
[8] Department of Nutrition and Dietetics, Faculty of Agriculture, University of Nigeria Nsukka, Nsukka, Enugu State, NIGERIA
[9] Department of Microbiology, Faculty of Biological Sciences, Mountain Top University, Pakuro, Ogun State, NIGERIA
[10] Department of Biological Sciences, Faculty of Science, National Open University of Nigeria, Abuja, NIGERIA
[11] Department of Pharmacology and Therapeutics, Faculty of Basic Medical Sciences, University of Ibadan, Ibadan, NIGERIA
[12] Department of Biochemistry, Faculty of Sciences, Obafemi Awolowo University, Ile-Ife, Osun State, NIGERIA
***Corresponding Author:** ogunjobitaiwo95@gmail.com

Citation: Ogunjobi, T. T., Okorie, I. C., Gigam-Ozuzu, C. D., Olorunleke, J. V., Ogunleye, F. I., Irimoren, E. O., Atanda, D. O., Okafor, N. M.-A., Agbo, C. E., Okunbi, F. O., Umoren, O. D., Adidu, A. D., & Ojo, E. O. (2025). Bioinformatics tools in protein analysis: Structure prediction, interaction modelling, and function relationship. *European Journal of Sustainable Development Research, 9*(3), em0298. https://doi.org/10.29333/ejosdr/16340

ARTICLE INFO

Received: 09 Mar. 2024

Accepted: 01 Jun. 2024

ABSTRACT

Protein analysis has been completely transformed by the swift growth of bioinformatics, which has improved protein structure prediction, simulated interactions, and clarified functional interactions. To improve our knowledge of proteomics, this review carefully examines the application of diverse bioinformatics methods in protein analysis. We evaluate computational methods such as molecular dynamics simulations and machine learning algorithms critically, with an emphasis on their applicability to modeling protein-protein interactions and protein tertiary structure prediction. Our findings show that these methods are useful for predicting protein functions and interactions, which are important for drug discovery and development. We also talk about the important implications of these developments for our knowledge of complex biological systems and disease mechanisms at the molecular level. This review also provides insights into the existing and future potential of bioinformatics tools, emphasizing their vital role in revolutionizing protein analysis. We additionally offer future strategies to improve our knowledge and management of complex disorders, particularly highlighting the need for integrated, multi-scale approaches and additional research on underrepresented proteins.

Keywords: bioinformatics, protein analysis, structure prediction, function relationship, molecular modeling, protein-protein interactions

INTRODUCTION

Bioinformatics is a broad field having applications in biological sciences, such as finding new vaccines and medications, enhancing the functionality of dietary proteins, and comprehending protein interactions. Bioinformatics is the study and use of computer algorithms to analyze biological data, such as genetic information, protein amino acid sequences, and protein structures. Given its broad definition, it is helpful to categorize bioinformatics; among the important approaches to comprehending protein analysis in genomics, the categories of bioinformatics, are comprehending protein structure, interaction modeling, and function relationships (Bolyen et al., 2018).

The bioinformatics tools used in protein analysis employ various computer methods and algorithms to analyze proteins. They are necessary for the prediction of protein structure and for the creation of three-dimensional models. Additionally,

these technologies facilitate the modeling of protein-protein interactions, which aids in our understanding of complex biological processes (Xu et al., 2018b). By predicting and annotating functional sites, bioinformatics tools help clarify how protein structure and function relate. They combine many data sources, including genomics and proteomics, to explain protein analyses thoroughly. These tools are valuable in drug discovery, assisting in virtual screening and optimizing drug candidates. Overall, computational approaches for investigating protein structure, protein-protein interactions, and the complex interplay between structure and function are provided by bioinformatics tools (van Beusekom et al., 2018). The significance of bioinformatics in protein analysis lies in its ability to process, analyze, and interpret the vast amount of biological data associated with proteins. Here are some key points highlighting the significance of bioinformatics in protein analysis (Sun et al., 2018). Sequences, structures, and interaction networks connected to huge amounts of protein-related data can be handled using tools and methods from the field of bioinformatics. It makes varied protein data from many sources organized, retrievable, storable, and integrated, making it available for analysis (Gao et al., 2019). Bioinformatics plays a critical role in predicting protein structures, which is essential for understanding protein function, interactions, and drug discovery. Even in the lack of actual structures, bioinformatics tools may create three-dimensional models of proteins by employing computational techniques and algorithms (Lu et al., 2020).

To fully understand intricate cellular processes, one must have a thorough understanding of protein-protein interactions. By assisting in the identification of protein complexes, binding sites, and interaction networks, bioinformatics tools make it easier to anticipate and simulate protein-protein interactions. This knowledge helps research biological processes and develop specialized treatments (Singh & Singh, 2021). Bioinformatics tools assist in the annotation and prediction of protein function. By analyzing protein sequence and structure, these tools can identify conserved domains, functional sites, and motifs. This knowledge provides insights into protein activity, enzymatic function, and involvement in specific biological processes.

Bioinformatics tools mine and analyze protein-related data using computational algorithms and statistical techniques. These technologies analyze enormous datasets statistically, find patterns, and produce insightful results. This data-driven methodology aids in discovering novel links, comprehending protein evolution, and directing research that is hypothesis-driven (Vignani et al., 2019). The procedures of drug discovery and design depend heavily on bioinformatics technologies. They support molecular dynamics simulations, ligand docking, and virtual screening, facilitating the identification and improvement of prospective drug candidates. Additionally, bioinformatics techniques aid in the analysis of the structure-activity relationship and the prediction of how mutations would affect therapeutic efficacy (Zhang et al., 2018).

Protein structure prediction is of paramount importance in the field of molecular biology and bioinformatics. The three-dimensional structure of a protein is closely linked to its function. By predicting the structure, researchers can gain insights into how the protein carries out its specific biological activities, which is crucial for understanding cellular processes and disease mechanisms (Khalatbari et al., 2019). Protein structure prediction is a powerful tool that provides critical structural information necessary for understanding protein function. It offers valuable insights into the mechanisms underlying protein activity, interactions, and their involvement in various biological processes and disease states. The ability to predict protein structures is essential for advancing our knowledge of molecular biology and has numerous applications in biotechnology, medicine, and drug discovery. Protein structure prediction allows researchers to generate three-dimensional models of proteins, providing information about the arrangement of atoms, secondary structures, and active sites (AlQuraishi, 2020). This structural information is fundamental to understanding how a protein's shape and spatial arrangement enable specific biochemical functions. The identification of active sites, functional domains, and binding pockets within a protein's structure is aided by protein structure prediction. For the protein to function biologically, these areas are critical for mediating interactions with other molecules, such as substrates, cofactors, ligands, or other proteins (Kwon et al., 2020).

The identification of active sites, functional domains, and binding pockets within a protein's structure is aided by protein structure prediction. For the protein to function biologically, these areas are critical for mediating interactions with other molecules, such as substrates, cofactors, ligands, or other proteins (Volkov et al., 2022). Protein structure prediction is a crucial aspect of drug discovery. Understanding the structure of a target protein helps researchers identify potential drug-binding sites and design molecules that can interact with the protein to regulate its activity. This information accelerates the drug development process. The protein's three-dimensional structure often determines the active site of enzymes, where catalytic reactions take place. Accurately predicting enzyme structures aids in comprehending the catalysis mechanism and guides efforts to improve enzyme efficiency or design novel enzymes for industrial applications. Different diseases can be caused by protein mutations (Kuhlman & Bradley, 2019).

Researchers can learn how structural alterations impact protein function and affect disease pathology by predicting the structures of both a protein's normal and mutant variants. Proteins can be altered for certain purposes thanks to predictions about protein architecture. The creation of proteins with improved stability, altered binding affinities, or novel capabilities for biotechnological and medicinal applications is made possible by rational protein engineering, driven by structural data. The modeling of protein-protein interactions is aided by accurate protein structure prediction. Deciphering these interactions is essential for understanding physiological processes, signaling pathways, and the formation of multi-protein complexes (Chao & Byrd, 2018). Time- and money-consuming experimental techniques include X-ray crystallography and NMR spectroscopy. The process of experimental structure determination can be sped up and guided by the early models that protein structure prediction can offer. Functional annotation is needed due to the enormous amount of genomic data that sequencing projects have produced. To better understand the biological

functions of genes, protein structure prediction helps link gene sequences with possible functions (Zhang et al., 2019a). An effective method for guiding the investigation of protein interactions and disease causes is protein structure prediction, which also makes drug discovery and protein engineering easier. It is a crucial part of contemporary molecular biology and has wide-ranging effects on numerous scientific and medical disciplines (Chen et al., 2019).

Many biological processes depend on protein-protein interactions, making precise protein complex prediction critical to comprehending these processes at the molecular level. The precision of predicted protein complexes can be increased by including knowledge of protein-protein interactions in the structure prediction process. To anticipate protein-protein interactions, docking methods are frequently utilized. Based on the structures of individual proteins, these algorithms forecast the three-dimensional structure of protein complexes. These docking algorithms can be guided by experimental data on known protein-protein interactions, boosting their dependability and lowering false positives. Many biological processes depend on protein-protein interactions, making precise protein complex prediction critical to comprehending these processes at the molecular level (Weng et al., 2020). The precision of predicted protein complexes can be increased by including knowledge of protein-protein interactions in the structure prediction process. The identification of important functional regions within the structure of a protein depends on the structure-function relationship (Jiang et al., 2019). Active sites and binding pockets are examples of predicted functional sites that are important in catalysis, substrate binding, and chemical recognition. Understanding the mechanisms by which proteins carry out their many functions is made easier by accurately predicting protein structure with functional annotations. Understanding biological processes and disease pathways need this information (Masrati et al., 2021).

This paper is of significant important because it explores in-depth how bioinformatics techniques can be used to improve our understanding of protein structures, interactions, and functions–areas critical for advances in pharmaceutical development and medical research. This paper is novel because it provides a thorough analysis of modern computational methods including molecular dynamics simulations and machine learning algorithms and shows how these methods can be used to more precisely predict complex protein behaviors (Vignani et al., 2019). This paper closes two gaps in the literature. Firstly, it offers a comprehensive overview of the various applications of bioinformatics tools now being used in the proteomics field, which has been fairly dispersed in earlier research (van Beusekom et al., 2018). This puts a unified viewpoint front and center, making it possible to comprehend the strengths and weaknesses of the approaches used today. Secondly, It discusses the requirement for more multi-scale and integrated methods in protein analysis. By drawing attention to this, it not only highlights a gap in the field's existing understanding but also paves the way for further investigations that may produce more reliable and thorough models of protein behavior. This is especially crucial for the continuous attempts to create focused, efficient treatments and to comprehend diseases at the molecular level (Kuhlman & Bradley, 2019).

MATERIAL & METHOD

Protein Structure Prediction Methods

The interdisciplinary study topic of protein structure prediction has drawn interest from academics in many different fields, including biochemistry, medicine, physics, mathematics, and computer science. These researchers are working on the same structure prediction problem using a variety of research paradigms: biochemists and physicists study the laws governing protein folding; mathematicians, particularly statisticians, assume a probability distribution of protein structures given a target sequence and then determine the most likely structure; and computer scientists frame protein structure prediction as an optimization problem–finding the best solution (Schönherr et al., 2018). Since the latter half of the 20th century, more academics from various disciplines have focused their research on bio-related topics. Protein is one of the most common and complex macromolecules in living things, which attracts a lot of attention. Proteins differ from one another principally in terms of the amino acids they contain, which often results in differences in their spatial shape and structure and, consequently, in the biological tasks that they may carry out in cells. The process by which a protein folds from its one-dimensional sequence into a specific three-dimensional structure, however, is unknown (Kotowski et al., 2021). Contrary to the genetic code, which makes use of a triple-nucleotide codon in the sequence of nucleic acid to specify a single amino acid in a protein sequence, the relationship between a protein's sequence and its steric structure is known as the second genetic code.

Protein structure prediction is a complex computational task that involves several methods and approaches. There are primarily two categories of methods used in protein structure prediction: template-based modeling (homology modeling or comparative modeling (CM)) and de novo modeling (ab initio modeling) (Yan et al., 2020). A basic assumption is that proteins with similar sequences fold into similar 3D structures. In HM, the 3D structure of the protein is built commencing from structural information of evolutionarily-related sequence(s), whereas the more general names Template Based Model or comparative model denote that a template protein is used but that the template is not necessarily of related history or function to the target (Yan et al., 2020). TBM entails several processes, including homolog (template) discovery, target alignment to the template, structure creation, refinement, and validation. The hybrid approaches include components from both groups for increased accuracy (Rives et al., 2019).

Template-Based Modeling (Homology Modeling)

The basis of homology modeling is the idea that proteins with related sequences frequently have related structures and activities. It begins by locating a well-known protein structure (template) with the target protein's (query) strong sequence similarity. The sequences of the target protein and the template sequence are then aligned to establish analogous

locations. To create a 3D model of the target protein, the template protein's coordinates are transferred to the target protein (Houkes & Zwart, 2019). For homology modeling, a variety of software programs like MODELLER, SWISS-MODEL, and Phyre2 are frequently used. The threading techniques that return a complete 3D description for the target and comparison modeling both fall under the category of techniques known as template-based modeling. This category of protein structure modeling depends on the noticeable similarity between at least one known structure and the majority of the modeled sequence. Comparative modeling describes those template-based modeling scenarios when a full atom model is constructed in addition to selecting the fold from a pool of potential templates (Mura et al., 2019). In actuality, it means that the other members of the family can be modeled based on their alignment to the known structure if the structure of at least one protein in the family has been identified by experiments. It is feasible because a slight modification to the protein's 3D structure typically follows a little modification to the protein's sequence. Additionally, it is made easier by the fact that proteins from the same family's 3D structures are more conserved than their amino-acid sequences (Kim & Chung, 2020). As a result, structural similarity can typically be assumed between two proteins if similarity can be found at the sequence level. The fact that proteins only adopt a relatively small number of distinct folds, as well as the intensive mapping of the universe of potential folds by global structural genomics studies, have led to an increase in the applicability of template-based modeling (Runthala & Chowdhury, 2019).

There are benefits and drawbacks to template-based techniques for structure prediction. Usually, high-quality models similar to medium-resolution NMR solution structures or low-resolution X-ray crystallography are produced through comparative protein structure modeling. However, only sequences that can be securely mapped to known structures can be used with these algorithms. At the moment, the likelihood of discovering similar proteins with a known structure for a sequence randomly selected from a genome varies between 30% and 80%, depending on the genome. A minimum of one domain that can be linked to at least one protein with a known structure exists in about 70% of all known sequences (Wang & Yang, 2019). The proportion of experimentally determined protein structures that have been stored in the protein data bank (PDB) is more than an order of magnitude greater. As we shall see, in actual template-based modeling, information from general statistical observations or molecular mechanical force fields, in the form of various force restrictions, is always included and is independent of the template. The most effective strategies are a result of better force fields and search algorithms (Jang et al., 2020).

Steps Involved in Template-Based Modeling

Template selection

Finding an appropriate template protein with the target protein's considerable sequence similarity is the first step. The template should ideally span the entire length of the target protein and have a high sequence identity. A crucial phase in template-based modeling, sometimes referred to as homology modeling or comparative modeling, is template selection. Finding an appropriate template protein with a recognized 3D structure and a substantial amount of sequence similarity to the target protein (query) is required (Behl & Mishra, 2018). The selection of a suitable template has a significant impact on how well the homology modeling process goes (Peker et al., 2019). A thorough explanation of the template-choosing procedure is provided below.

1. **Sequence database search:** The sequence database search, such as BLAST (Basic Local Alignment Search Tool) or PSI-BLAST (Position-Specific Iterated BLAST), usually comes first in the template selection process. These algorithms compare the target protein's amino acid sequence to sequences in freely accessible databases like UniProt and the Protein Data Bank (PDB) (Gebert et al., 2019).

2. **Sequence identity threshold:** When choosing a template, sequence identity is an important consideration. It is more likely that target and template proteins will share comparable structures and functions the higher their sequence similarity. A sequence identity of between 30% and 40% is typically regarded as adequate for homology modeling (Hao et al., 2018).

3. **Coverage and alignment quality:** In addition to determining the sequence identity, it's critical to evaluate the target and template sequences' coverage and alignment quality. The target protein sequence should be covered by the alignment to the greatest extent possible, ideally with continuous lengths of aligned residues (Hiranuma et al., 2021).

4. **PDB template quality:** The PDB template's 3D structure must be of a high standard. Selecting a template with a high-resolution experimental structure and few mistakes or artifacts is essential. Assessing the caliber of the template structure can be done with the aid of structural validation programs like MolProbity or PROCHECK (Lensink et al., 2018).

5. **Biological relevance:** The chosen template must be compatible with the target protein's biology. Picking a template with the target's function or domain architecture in mind is recommended. Selecting a template from the same family as the target protein can improve the homology model's accuracy if the target protein is a member of that family of proteins (Salinas & Ranganathan, 2018).

6. **Consistency with biological knowledge:** The template choice should be in line with the functional annotations and current biological knowledge. An excellent basis for selection can be provided by experimental data pointing to a close link between the target protein and a particular template (Ban et al., 2019).

When choosing a template for template-based modeling, it is important to carefully consider factors such as sequence identity, coverage, alignment quality, structural quality, biological significance, and congruence with known facts. The choice of a suitable template has a considerable impact on the precision and dependability of the homology model, making it an essential step in the entire protein structure prediction process (Jia & Jernigan, 2021).

Sequence alignment

After the template has been chosen, an alignment of the target protein's amino acid sequence with the template's sequence is done. To ensure that related residues are aligned appropriately, the alignment determines analogous locations (also known as residues) in both sequences. A key component of template-based modeling, sometimes referred to as homology modeling or comparative modeling, is sequence alignment (Kondra et al., 2021). It entails matching the target protein's (query's) amino acid sequence with the sequence of a recognized template protein that exhibits a substantial degree of sequence similarity. To ensure that related residues are accurately matched, the alignment seeks to find equivalent places (also known as residues) in both sequences (Karasev et al., 2019).

Here is a thorough description of how the sequence alignment procedure works.

1. Database search for potential templates

Finding possible template proteins with recognized 3D structures is the first step in the sequence alignment procedure. This search is often carried out against freely accessible databases like the Protein Data Bank (PDB) or UniProt using sequence comparison techniques like BLAST (Basic Local Alignment Search Tool) or PSI-BLAST (Position-Specific Iterated BLAST) (Mohamed et al., 2018).

2. Sequence identity calculation

Calculating the degree of sequence similarity between each candidate template and the target protein comes next once potential templates have been discovered. The proportion of residues that are identical between the two sequences is measured as sequence identity. The likelihood of a protein's structure and function being comparable depends on the degree of sequence identity (Marino & Dell'Orco, 2019).

3. Multiple sequence alignments

Multiple sequence alignments may be produced, depending on the quantity of available templates and the diversity of the target protein's sequence. The sequences can be aligned using a variety of methods, including ClustalW, MUSCLE, or MAFFT (Yu et al., 2019).

4. Gap penalty & scoring

To align the residues between the target and template sequences, gaps (insertions or deletions) are inserted during sequence alignment. Each aligned pair of residues is given a score by the alignment algorithms using scoring matrices like the BLOSUM or PAM matrices. To ensure the best alignment of residues and reduce the number of gaps, gap penalties are used (Lin & Hsu, 2020).

5. Consensus sequence & alignment

The most prevalent amino acid is then assigned to each aligned location to create a consensus sequence. Building the homology model is guided by the consensus sequence (Chen et al., 2018).

6. Evaluation & quality assessment

Several measures, including sequence identity, alignment coverage, and alignment scores, are used to evaluate the quality of the sequence alignment. To produce an accurate homology model, a high-quality alignment is essential. Sequence alignment ensures the correlation between the target and template sequences, making it a vital step in template-based modeling. A precise alignment serves as the foundation for the remaining phases in creating the target protein's 3D homology model (Huang et al., 2018).

Protein-protein interaction analysis

The homology model can shed light on the interface of the interaction if the target protein is involved in protein-protein interactions. The model can be used to investigate protein-protein interactions and direct experimental research on protein complexes by examining the surface residues and determining potential binding partners (Jin et al., 2020). By comparing its structure to those of known proteins with related functions, the homology model can help with functional annotations. If the modeling template has a function that has been identified, this knowledge can be translated to the target protein, offering useful functional insights.The homology model can be utilized to verify theories on the role or method of action of the target protein. The model can be used, for instance, to examine the structural background and putative functional significance of a particular residue that is thought to be essential for a biological activity (Jha & Saha, 2020).

The homology model can be used for structure-based medication design and virtual screening. It is a useful tool for foretelling ligand-binding interactions, directing the design of new compounds, and evaluating the likelihood that ligands will become therapeutic candidates.

When the target and template proteins share more than 30% to 40% of their sequences, template-based modeling is especially useful. However, when the sequence similarity drops, the homology model's accuracy declines, making it difficult for proteins with low sequence identity to match known structures. To predict protein structures under these circumstances, other techniques including de novo modeling and hybrid approaches are used (Lin et al., 2021).

BIOINFORMATICS TOOLS FOR PROTEIN STRUCTURE PREDICTION

Phyre2: Protein Homology/AnalogY Recognition Engine

Phyre2 (Protein Homology/AnalogY Recognition Engine 2) is a popular web-based program for predicting and analyzing protein structure. The Söding Group at the University of Oxford created and maintains it. The Phyre2 server, which replaces the original Phyre server, provides more accuracy and more features (Nardo et al., 2018).

Features & functionalities

By finding related proteins in the Protein Data Bank (PDB) with known structures and building models using these templates, Phyre2 uses homology modeling to predict protein structures. It makes use of ab initio modeling to investigate alternative conformations and forecast stable structures when a suitable template is not readily available. Furthermore, Phyre2 recognizes protein folds, deduces structural and functional characteristics, provides functional annotations,

Figure 1. PHYRE homepage (login page) (http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index)

Figure 2. Submission page (http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index)

predicts domains, and offers tools for thorough structural investigation (Orbán-Németh et l., 2018; Zhou & Panaitiu, 2020).

Usage

Using Phyre2 is typically straightforward through its web-based interface:

1. **Submission of the sequence:** Through the Phyre2 website, users can submit an interesting protein sequence in FASTA format.
2. **Analysis and prediction:** Phyre2 will examine the sequence and try to identify homologous templates, either by ab initio or homology modeling. The user is shown the best forecast or predictions.
3. **Visualization and analysis:** Using the tools and features offered, the user can visualize and examine the predicted protein structures.

Step in Using Phyre2

1. **Visit PHYRE2 website:** Launch a web browser and go to PHYRE2 website. Visit (http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index) to access it (**Figure 1**).
2. **Create an account or log in (optional):** If you are a first-time user, you might need to create an account. Although registration is not always necessary, it can be helpful for keeping track of your contributions and outcomes. Log in if you already have an account (see **Figure 1**).
3. **Submit your sequence:** A sequence submission box can be found on the PHYRE2 homepage. There are several ways to submit your protein sequence (see **Figure 2**):
    a. **Uploading a file:** If you have your protein sequence in a file (FASTA format is advised), you may upload it from your computer by clicking the "choose file" button (see **Figure 2**).
    b. **Pasting the sequence:** In addition, you have the option of simply pasting your protein sequence into the available text box (see **Figure 2**).
4. **Select an analytical option:** Options for doing various analyses can be found below the sequence submission box. Among these choices are:
    a. You can choose whether or not to receive secondary structure predictions.
    b. Solvent accessibility prediction: This option will forecast solvent accessibility.
    c. PHYRE2 can recognize domains in your protein.
    d. Make these options unique to your research requirements. Typically, leaving the default values is sufficient for typical structure prediction.
5. **Submit your job:** Click the "submit job" button after uploading your sequence and choosing the appropriate analysis choices. Your request will begin to be processed by PHYRE2 (see **Figure 2**).
6. **Awaiting results:** Depending on the intricacy of your request and the traffic on the system, the analysis may take a while to finish. An anticipated completion time will be given by PHYRE2.
7. **Obtain and examine results:** After the study is finished, PHYRE2 will give you the findings, which will include anticipated 3D models of your protein and other pertinent data.
8. **Download results:** You have the option to download the results for additional analysis and investigation. The predicted structures and other output files from PHYRE2 are normally available via download links.
9. **Examine & analyze findings:** Carefully examine the anticipated protein structures and related information. In light of your research's goals and objectives, interpret the results.

Swiss-Model: Protein Structure Homology Modeling Server

The Swiss Model Server is a tool created to make protein structure modeling and prediction easier. Researchers and scientists in the field of structural biology frequently utilize it; it is a component of the Swiss Institute of Bioinformatics (SIB). Based on amino acid sequences, the service uses cutting-edge computational methods to create three-dimensional models of protein structures (Mrozek et al., 2019).

An automated system called SWISS-MODEL (http://swissmodel.expasy.org/) uses homology modeling techniques to model the 3D structure of a protein from its amino acid sequence. Since its establishment as the first completely automated server for protein structure homology modeling 20 years ago, SWISS-MODEL has been constantly expanded and enhanced. The server has an intuitive web interface that makes it possible for non-specialists to create 3D models of their chosen proteins using a standard web browser

without having to download or install any complicated molecular modeling software (Kandathil et al., 2022). SWISS-MODEL receives more than 0.9 million requests for protein models each year, or almost one model per minute, making it one of the most frequently used structure modeling web servers in the world. Its functionality has recently been greatly expanded: SWISS-MODEL now models the oligomeric structures of the target proteins and incorporates evolutionary conserved ligands like metal ions or essential cofactors (Seidl et al., 2022). Users can now easily search for suitable templates using sensitive Hidden Markov Models (HMM) searches against the SWISS-MODEL Template Library (SMTL), analyze alternate templates and alignments, perform structural superposition and comparison, explore ligands and cofactors in templates, and compare the resulting models using mean force potential-based model quality estimation tools (Makigaki & Ishida, 2019).

Swiss-Model Web Interface

Input

With SWISS-MODEL, model building can be started from a variety of beginning points: A protein amino acid sequence can be given in the simplest case either directly (raw one-letter sequence or FASTA format) or by referencing its UniProt accession code in which case SWISS-MODEL will automatically get the matching item from UniProt (Lu et al., 2022). An alternative method for specifying a target-template sequence alignment is to use a multiple-sequence alignment that includes the target, the template, and eventually other homologous sequences. At this stage, the user can either start the completely automated modeling pipeline or start the template selection step, which allows them to manually choose particular templates (de Medeiros et al., 2020).

1. The relationship of the discovered templates in the space of sequence similarity is displayed in an interactive chart. A filled red circle designates the target protein. Each template is represented as a blue circle, with a thick blue arc indicating target coverage (the target protein's N-terminus begins at the top of the circle and wraps around clockwise to form the circle's border). Evolutionarily related templates will group because the distance between them is proportional to the pairwise sequence similarity.
2. Information unique to the template will be seen when you click on a circle. By hovering your mouse over a collection of templates, you may also see and choose a group of related templates.
3. For a quick visual comparison of structural variations, the superposed structures of the chosen templates will be exhibited in 3D right away (**Figure 3**).

   (A) Coordinates, target-template alignment, modeling log, as well as quality evaluation data are all given for each model. Additionally supplied is information on the ligands, cofactors, and oligomeric structure.

   (B) By selecting the corresponding button (represented by the adjustable spanner icon), the target-template sequence alignment's color scheme can be changed to a different one. The model's structural representation updates at the same time as the model itself.

   (C) Model quality assessments assigned by QMEAN are originally used to color the models displayed in the interactive viewer. This enables quick differentiation between model regions with good modeling and those with bad modeling. The per-reside plot (A) and global score (Z-score) in respect to a collection of high-resolution PDB structures (D) represent local estimations of the model quality based on the QMEAN scoring function (**Figure 4**).

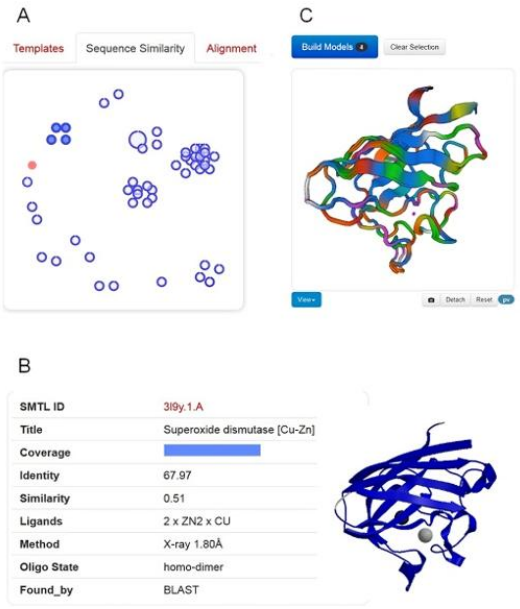

Figure 3. Templates selection and visualization (https://swissmodel.expasy.org/)

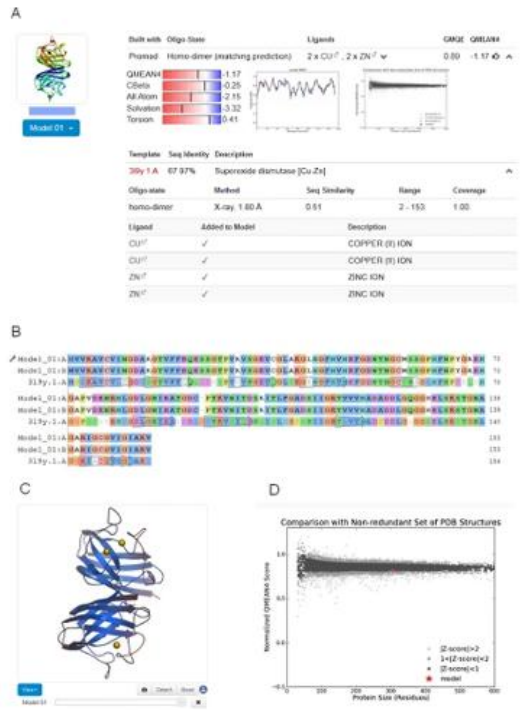

Figure 4. Modelling results (https://swissmodel.expasy.org/)

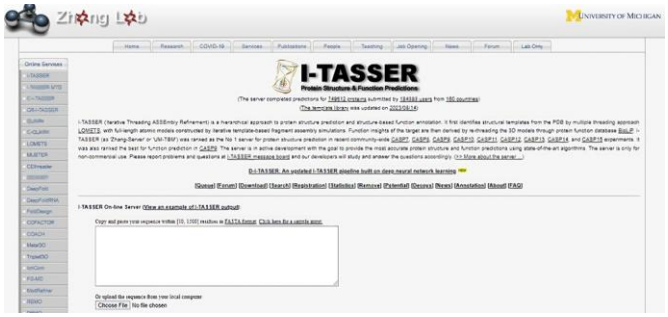

Figure 5. I-Tasser homepage (https://zhanggroup.org/I-TASSER)

Figure 6. Account setup page (https://zhanggroup.org/I-TASSER)

I-TASSER: Iterative Threading Assembly Refinement

Workflow & principles of I-TASSER

I-TASSER is a hierarchical system for structure-based function annotation and automated protein structure prediction. I-TASSER initially creates full-length atomic structural models from numerous threading alignments and iterative structural assembly simulations, followed by atomic-level structure refinement, starting with the amino acid sequence of the target proteins (Xu et al., 2018a). Based on sequence and structure profile comparisons, the biological functions of the protein, including ligand-binding sites, the enzyme commission number, and gene ontology terms, are then inferred from databases of known protein functions (Zhou et al., 2019). Both an online server and a standalone version of I-TASSER are offered without charge. This section explains how to develop structure and function predictions using the I-TASSER protocol, how to interpret the predictions, and alternate methods for enhancing the quality of I-TASSER modeling for targets that are distantly homologous and multi-domain proteins (Cheung & Yu, 2018).

Steps in using I-Tasser server

Please visit the website for the most current instructions.

Figure 5 shows the homepage.

Figure 6 shows account set-up page.

Here's a general outline of the steps:

1. **Access the I-TASSER server:** Go to the I-TASSER website (https://zhanglab.ccmb.med.umich.edu/I-TASSER/) to access the server interface. Figure 7 illustrates the submission form of I-TASSER with an example sequence.

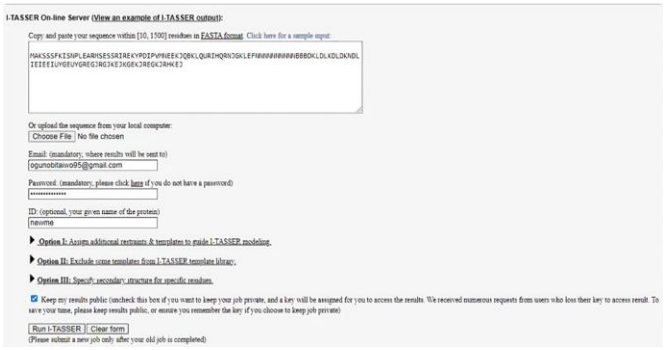

Figure 7. Submission form of I-TASSER with an example sequence (https://zhanggroup.org/I-TASSER/)

2. See Figure 7.
3. **Submit your protein sequence:** Enter or paste the amino acid sequence of the protein you want to predict the structure for into the provided text box.
4. **Provide an optional email address:** While not required, providing an email address can be useful to receive notifications when your job is completed and to access the results later.
5. **Submit your job:** Click the "Submit" button to initiate the structure prediction process. Your sequence will be sent to the I-TASSER server for analysis.
6. **Wait for completion:** The I-TASSER server will perform a series of tasks including threading, ab initio modeling, and model refinement to predict the protein's structure. The time it takes for the prediction to complete can vary depending on the server's workload and the complexity of the protein.
7. **Receive results:** Once the prediction is complete, you will receive an email notification (if you provided an email address) with a link to access your results. Alternatively, you can also access your results by entering your job ID on the I-TASSER website.
8. **Analyze results:** The results page will typically provide information about the predicted models, including their quality assessment scores, estimated accuracy, and more. You'll be able to download the predicted models and related data for further analysis.
9. **Model selection and refinement:** Analyze the predicted models and choose the one that appears to be the best representation of the protein's structure. You can further refine the selected model using various molecular modeling tools if needed.

I-Tasser protocol for protein structure & function prediction

Several other publications have provided descriptions of the I-TASSER protocol's specifics. I-TASSER, which begins with the amino acid sequence, first locates homologous structure templates (or super-secondary structural segments, if no homologous templates are available) from the PDB library using LOMETS, a meta-threading algorithm made up of numerous separate threading programs. The constantly aligned fragment structures removed from the LOMETS

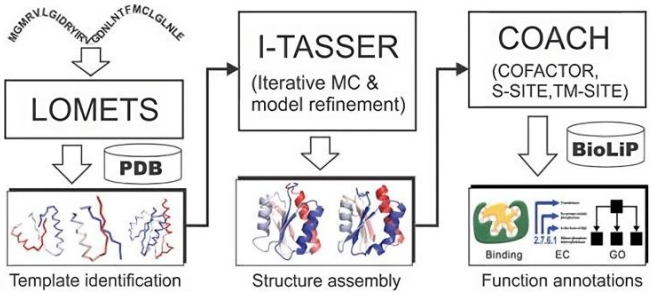

Figure 8. Protocol of I-Tasser (https://zhanggroup.org/I-TASSER/)

templates and super-secondary structure segments are then put back together to create the topology of the full-length models, with the structures of the unaligned regions being built entirely from scratch by ab initio folding based on replica-exchange Monte Carlo simulations (Milanetti et al., 2018). Through the clustering of the Monte Carlo simulation trajectory data, SPICKER determines the lowest-free-energy conformations. The structural models are refined through a second phase of structure reassembly starting from the SPICKER clusters, with full-atomic simulations using FG-MD and ModRefinerused to refine the low free-energy conformations (Tong et al., 2018) (**Figure 8**).

Automated methods for protein structure analysis and prediction are available on the Robetta server (http://robetta.bakerlab.org). Sequences supplied to the server are parsed into putative domains for structure prediction, and then structural models are created using de novo structure prediction or comparative modeling techniques (Rodrigues et al., 2019). Using BLAST, PSI-BLAST, FFAS03, or 3D-Jury, if a confident match to a protein with a known structure is discovered, it is used as a template for comparative modeling. Structure predictions are created using the de novo Rosetta fragment insertion method if no match is discovered. For RosettaNMR de novo structure determination, experimental nuclear magnetic resonance (NMR) constraints data can also be given together with a query sequence. The prediction of the effects of mutations on protein-protein interactions utilizing computational interface alanine scanning is another capability that is already available. Soon, the service will also provide access to the protein-protein docking and Rosetta protein design technologies (Mao et al., 2019).

The Baker Lab at the University of Washington created the renowned protein structure prediction pipeline known as Robetta. Robetta predicts protein structures from amino acid sequences using a variety of cutting-edge algorithms and methods. It has been frequently utilized to produce precise 3D models of proteins by structural biology researchers (Zhao et al., 2021a). The pipeline, its parts, and its applications are described in detail below.

Robetta protein structure prediction pipeline

1. **Modeling with templates (homology modeling)**
   a. Robetta starts by looking for proteins with comparable sequences to the target protein in the Protein Data Bank (PDB), a vast collection of experimentally verified protein structures.
   b. Robetta generates an initial model by aligning the target sequence to the template structure if suitable templates are discovered.
   c. After that, the pipeline uses molecular dynamics simulations to refine the model and optimize the shape of the structure.
2. **Folding from scratch**
   a. Robetta uses ab initio folding techniques to estimate the protein's structure from scratch if a suitable template is not provided.
   b. In order to find the protein's lowest-energy conformation, ab initio approaches examine various protein backbone and side chain conformations.
3. **Prediction of side chain positions**
   a. Robetta uses machine learning and energy-based techniques to anticipate, where the side chains will be located in the protein structure.
   b. A realistic protein structure can only be obtained with accurate side chain prediction.
4. **Model quality assessment:** Using a variety of criteria, such as energy calculations, structural validation tools, and compatibility with the input sequence, the pipeline evaluates the quality of the created models.
5. **Model refinement:** In order to maximize their overall geometry, energy, and clash-free interactions, the models created by Robetta are further refined.
6. **Loop modeling:** Robetta makes use of specific algorithms to represent missing or ambiguous loops or portions of the protein structure.

Applications & strengths of robetta

Biochemistry and molecular biology's prediction of protein structures is a key task with broad applications. Deciphering proteins' relationships, roles, and functions in numerous biological processes requires an understanding of their three-dimensional structures. In the area of protein structure prediction, Robetta is a prominent platform that stands out for its astounding precision and adaptability (Song et al., 2018). The uses of Robetta in biochemistry and biology are very broad. It is crucial for functional annotation, to start with. Robetta bridges the gap left by the lack of empirically established structures in many recently sequenced proteins by making precise structural predictions. These hypotheses provide a framework for annotating the functions of these proteins, illuminating their roles in both healthy cellular function and illness. Robetta is used extensively in the drug discovery process. Researchers can locate prospective therapeutic targets and create compounds that can interact with these targets according to the platform's precise protein structure predictions (Wang et al., 2019). This is especially useful in the field of structure-based drug design, as effective therapies require knowledge of the characteristics and shape of a protein's active site. Robetta excels due to its outstanding precision, durability, and adaptability. To increase its predictive potential, it makes use of a variety of cutting-edge computational techniques, such as ab initio modeling and homology modeling. This multifaceted strategy boosts the possibility of acquiring precise protein structures, especially in difficult situations like membrane proteins and proteins with multiple domains (Fukuda & Tomii, 2020). Robetta also keeps

changing to include the most recent developments in structural biology and bioinformatics. Robetta is a trustworthy and current resource in the constantly developing field of protein structure prediction because of this dedication to improvement, which guarantees users have access to cutting-edge prediction techniques. Its user-friendly interface also makes it accessible to scientists with different degrees of computational experience, promoting widespread use and inter-disciplinary cooperation in the scientific community (Zheng et al., 2019).

As a result, Robetta stands out as a reliable and accurate platform for this use. Protein structure prediction is a crucial tool in biochemistry with many applications. Its uses in structural biology, drug development, and functional annotation, along with its accuracy, adaptability, and constant progress make it an essential tool for researchers trying to solve the puzzles of protein structure and function (Larke et al., 2021).

Robetta's merits lay in its ability to integrate numerous approaches, including homology modeling and ab initio folding, to build accurate protein structures. It is flexible and adaptable to a variety of protein targets because to the mix of template-based and de novo prediction techniques. The predictions' accuracy can, however, differ based on variables like sequence similarity, template accessibility, and structural complexity, much like with all other methods for structure prediction (Wojtowicz et al., 2020).

PROTEIN-PROTEIN INTERACTION MODELING

Importance of Protein-Protein Interactions in Biological Processes

The workhorses of biological systems, proteins perform a wide range of vital tasks for life. However, they rarely carry out their operations alone. Instead, proteins frequently participate in complex molecular dances of which protein-protein interactions (PPIs) are the most prevalent and important. Numerous biological processes are built upon these interactions, which are crucial to cell signaling, enzyme activity, and even the control of gene expression (Rigoldi et al., 2018). For understanding the intricacy of life's inner workings, it is essential to comprehend the relevance of PPIs. PPIs play an important part in cell signaling and communication. To adjust to shifting environment, cells must react to outside cues like as hormones or neurotransmitters. Proteins may send and receive these signals thanks to PPIs, which relays important information inside the cell. For instance, serotonin and other neurotransmitters in the nervous system attach to neuronal receptors to start a chain reaction of PPIs that eventually affects mood, behavior, and other physiological functions (Bergenholm et al., 2018). The catalysts that power the chemical reactions required for life are enzymes. Many enzymes are made up of several protein subunits, which require precise interactions to work properly. PPIs make sure that these subunits combine at the proper moment and in the right orientation, promoting reactions that would not otherwise be energetically advantageous. Among other processes, such as DNA replication, cellular respiration, and metabolic pathways, this coordinated activity of proteins in enzyme complexes is essential (Seath et al., 2021).

PPIs are essential for the control of genes. For instance, proteins called transcription factors bind to particular DNA regions to regulate the production of genes. PPIs can modify their activity by joining forces with different proteins to create complexes. The timing and amount of gene expression are carefully regulated by this control, which affects cell destiny, differentiation, and responses to environmental signals (Li et al., 2018). PPIs make sure that these subunits combine at the proper moment and in the right orientation, promoting reactions that would not otherwise be energetically advantageous. Among other processes, such as DNA replication, cellular respiration, and metabolic pathways, this coordinated activity of proteins in enzyme complexes is essential (Ali et al., 2019).

PPIs are essential for the control of genes. For instance, proteins called transcription factors bind to particular DNA regions to regulate the production of genes. PPIs can modify their activity by joining forces with different proteins to create complexes. The timing and amount of gene expression are carefully regulated by this control, which affects cell destiny, differentiation, and responses to environmental signals. Researchers are concentrating more on PPIs as prospective therapeutic targets since they understand how important they are to biological processes (Liu et al., 2020). Specific PPIs can be targeted by small compounds or biologics to be enhanced or disrupted, influencing important disease-related pathways. Especially for complicated disorders, where single-target interventions might not be sufficient, this method shows promise for the development of more precise and efficient treatments. The fundamental units of complexity in life are protein-protein interactions. From basic biological functions to complex disease mechanisms and medication development, their importance is broad. Our ability to unravel the intricacies of biological systems and create creative ways to deal with the health concerns of our time improves along with our grasp of PPIs, which is still being further understood (Gouthami et al., 2022). As a result, the research of PPIs is at the cutting edge of contemporary biology, revealing the mysteries of life's most complex dance.

Various Approaches for Modeling Protein-Protein Interactions

Docking

Thus, the study of PPIs is at the cutting edge of modern biology, revealing the mysteries of life's most complex choreographies.Nearly all biological processes are governed by protein-protein interactions (PPIs), which coordinate processes like signaling, enzyme catalysis, and structural stability (Frezza & Lavery, 2019). For the purpose of developing new drugs, understanding structural biology, and obtaining knowledge of the complex mechanisms governing cellular regulation, it is crucial to comprehend the molecular specifics of these interactions. The computational method of docking has become a potent tool for modeling PPIs. We shall examine numerous docking approaches in this article, shining light on their methodologies, uses, and importance in understanding PPIs (Miller et al., 2020).

I. **Rigid body docking:** The simplest type of protein-protein interaction modeling is rigid body docking. It is assumed that during binding, proteins preserve their three-dimensional structures. In order to maximize the complimentary surface contacts between two proteins, this approach entails finding the best orientation and placement for one protein in relation to the other. When simulating interactions between clearly characterized protein domains, rigid body docking is very helpful since it can provide important details about potential binding locations and orientations (Siebenmorgen & Zacharias, 2020).

II. **Flexible docking:** Flexible docking recognises that proteins can undergo conformational changes following binding, in contrast to rigid body docking. This method enables small modifications to the protein structures to enhance binding affinity. Flexible docking techniques, such induced-fit and conformational ensemble docking, enable a more accurate representation of PPIs by taking into consideration the dynamic nature of proteins. This is especially important when researching interactions involving proteins that are inherently disordered or proteins that undergo significant conformational changes upon binding (De Paris et al., 2018).

III. **Energy-based docking:** The goal of energy-based docking methods is to determine the protein complexes' binding free energies. These techniques use scoring functions and force fields from molecular mechanics to determine the strength of the connection. They consider a number of variables, such as solvation energies, electrostatic interactions, and van der Waals forces. When assessing possible protein complexes and forecasting the stability of PPIs, energy-based docking offers a quantitative evaluation of the binding affinity (De Paris et al., 2018).

IV. **Data-driven docking:** Data-driven docking methods increase the precision of PPI models by utilizing experimental data from techniques like NMR spectroscopy, cryo-electron microscopy, or chemical cross-linking. The protein-protein interaction predictions are improved by including experimental limitations into the docking process. This method is especially useful for researching complicated and fleeting relationships that are difficult to model merely from structural data (Cava & Castiglioni, 2020).

V. **Machine learning in docking:** Recent developments in machine learning have also influenced docking techniques. Large datasets of well-known PPIs can be used to train machine learning algorithms that predict binding affinities, discover interaction hotspots, and speed up docking. The effectiveness and precision of docking simulations could be considerably improved by using these strategies (Bekker et al., 2020).

Applications & significance

In many scientific domains, docking is crucial. By foreseeing the interactions of tiny compounds with the target proteins, virtual screening employing docking can reveal prospective drug candidates in the drug discovery process. Docking is a technique used in structural biology to reveal the structural underpinnings of protein interactions and reveal disease processes. Additionally, docking is essential in systems biology for comprehending regulatory networks and signaling cascades (Cetin et al., 2020). The versatile and essential technology of docking is used to model protein-protein interactions. Rigid body docking, data-driven, and machine learning-based solutions are just a few of its numerous ways that provide a range of methods to meet the varied difficulties faced by PPIs. Docking is at the forefront of computational biology as our understanding of molecular interactions advances, offering useful insights into the intricate world of protein-protein interactions (Souza et al., 2021).

Molecular dynamics simulations

The fundamental building block of biological processes, protein-protein interactions (PPIs) control cellular activities, signaling networks, and structural stability. In order to fully understand the intricacies of life, it is essential to comprehend the dynamics of these interactions at the molecular level. A potent tool for modeling PPIs has emerged: molecular dynamics (MD) simulations, a computer method with physics and chemistry roots (Cezar et al., 2020). The many methods used in molecular dynamics simulations are examined in this article, along with their techniques, uses, and importance for understanding protein-protein interactions.

I. **Atomistic molecular dynamics simulations:** Atomistic MD simulations try to mimic how certain atoms and molecules behave throughout time. They compute the forces between atoms using classical force fields and rely on Newton's equations of motion. Atomistic MD offers a thorough perspective of the dynamic behavior of interacting proteins in the setting of PPIs. During binding events, proteins can be observed to move, interact, and alter conformation, providing information about the binding mechanisms and energy matrices (Harada, 2018).

II. **Coarse-grained molecular dynamics simulations:** The depiction of molecules is made easier by coarse-grained (CG) MD simulations, which combine many atoms into a single interaction site. Longer simulation timescales are possible as a result of the decreased computational complexity. While giving up some atomic-level information, CG MD can capture the crucial aspects of binding events in PPI research. Large protein complexes, protein folding, and the dynamics of inherently disordered proteins can all be studied using this method (Zhang et al., 2019b).

III. **Enhanced sampling techniques:** The timing restrictions of MD simulations, which can be a substantial difficulty in analyzing rare or complex PPI events, are intended to be solved through enhanced sampling strategies. Metadynamics,

replica exchange, and accelerated MD are some methods that make it easier to explore conformational space. Researchers can discover transient binding states and kinetic pathways that could be missed in conventional MD simulations by focusing the simulations on certain regions of interest (Truong & Li, 2018).

IV. **Quantum mechanics/molecular mechanics (QM/MM) simulations:** QM/MM simulations combine classical MD for the majority of the system with quantum mechanical calculations for a small region of interest (often the active site). This method enables realistic modeling of chemical processes and the electronic structure of reactants and products, which is particularly useful when researching enzymatic PPIs (Watanabe et al., 2020).

V. **Free energy calculations:** Calculations of free energy are intended to measure the energetics and binding affinity of PPIs. MD simulations are used to determine the free energy differences between bound and unbound states in techniques like umbrella sampling and thermodynamic integration. These calculations offer information on the kinetics and thermodynamics of PPIs, which is essential for developing new drugs and comprehending how biological processes work (Senior et al., 2020).

Applications & significance

MD simulations have a wide range of applications in the study of PPIs. They provide atomistic insights into binding mechanisms, binding pathways, and the role of water molecules in PPIs. MD simulations also help elucidate the structural dynamics of protein complexes, uncover transient intermediate states, and inform mutagenesis experiments. Furthermore, they play a pivotal role in drug discovery by predicting binding affinities, identifying potential drug candidates, and aiding in the rational design of novel therapeutics (Fang et al., 2019). Molecular dynamics simulations have revolutionized our ability to model and understand protein-protein interactions at the atomic level. The various approaches within MD, from atomistic to coarse-grained simulations, enhanced sampling techniques, QM/MM simulations, and free energy calculations, offer a diverse toolkit for studying PPIs across different timescales and levels of detail. As computational resources and methodologies continue to advance, MD simulations remain at the forefront of computational biology, enabling researchers to unravel the intricacies of protein-protein interactions and their role in the molecular machinery of life (Sejdiu & Tieleman, 2021).

Notable Software & Databases for Studying Protein-Protein Interactions

Access to specialist software tools and databases that offer a plethora of knowledge about protein interactions, structures, functions, and more is necessary for studying protein-protein interactions (PPIs) (Gemovic et al., 2019). Here are some significant programs and datasets that are frequently employed in the PPI research field:

1. STRING

Database: The STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) database contains details on both known and anticipated PPIs. It is extensive and widely utilized. To generate protein interaction networks for a variety of animals, it incorporates data from several sources, such as experimental evidence, co-expression, and text mining (Crosara et al., 2018).

2. BioGRID

Database: A curated collection of PPIs can be found in the Biological General Repository for Interaction Datasets (BioGRID). It includes interactions from both small-scale research and high-throughput trials. PPI data for different species can be searched for and retrieved using BioGRID's user-friendly interface (Zhao et al., 2021b).

3. Interologous interaction database (I2D)

Database: The PPIs that are conserved across species (interologs) are the topic of the specialist database I2D. It offers a user-friendly framework for examining preserved interactions and incorporates data from various PPI datasets (Nguyen et al., 2021).

4. Cytoscape

Software: Popular open-source software program Cytoscape is used to visualize and examine complicated networks, including PPI networks. In order to analyze networks, visualize them, and integrate them with different data sources, it offers a large variety of plugins and tools (Defoort et al., 2019).

5. STRING-DB

Software: The desktop version of the STRING database, STRING-DB, enables users to carry out comprehensive analysis of PPI networks on their own computers. It offers further customization options and sophisticated network analysis features (Sakhaee & Wilson, 2021).

6. Biological network gene ontology (BiNGO):

Plugin: A Cytoscape plugin called BiNGO was created specifically for PPI network analysis. Insights into the functional context of protein interactions are provided by helping to discover overrepresented Gene Ontology terms in a network (Yerneni et al., 2018).

7. MINT (Molecular INTeraction Database)

Database: MINT is a repository of PPIs with experimental support. It focuses on the interactions of proteins from Homo sapiens and gives comprehensive details on the experimental techniques used to find interactions (Bajpai et al., 2019).

For researchers looking into protein-protein interactions, this ecosystem of software tools and databases is very strong. They include a wide range of tools and features, such as the ability to explore functional annotations, visualize interaction networks, and support the identification of new PPIs. The tools and resources that best meet the interests and goals of each individual researcher are available for selection (Yim et al., 2018).

FUNCTIONAL ANNOTATION METHODS AND DATABASES

1. **UniProt:** One of the largest and most popular protein databases is the Universal Protein Resource (UniProt).

It offers a substantial database of protein sequences, functional annotations, and details on the structure and domains of proteins. The three primary parts of UniProt are UniProtKB, which provides curated protein sequences, UniRef, which provides clustered groupings of related sequences, and UniParc, which provides historical protein archive data. For researchers looking to investigate the links between proteins' functional properties, it acts as a fundamental resource (Davis et al., 2018). Website: https://www.uniprot.org/

2. **InterPro:** In order to anticipate protein domains, families, and functional locations, InterPro is an integrated resource that incorporates data from many sources. By utilizing data from Pfam, PROSITE, PRINTS, SMART, and other sources, it provides an in-depth understanding of how proteins operate. For scientists looking to gain a comprehensive grasp of how proteins interact functionally, InterPro is particularly useful (Zhang et al., 2020). Website: https://www.ebi.ac.uk/interpro/
3. **Gene ontology (GO):** A structured vocabulary for characterizing the functions of genes and proteins is offered by the Gene Ontology (GO) project. GO divides protein classification into three categories: cellular component, biological process, and molecular function. Each item in the ontology has a connection to a gene or protein, allowing researchers to systematically annotate and research the functions of proteins (Stacey et al., 2018). Website: http://geneontology.org/
4. **KEGG (Kyoto Encyclopedia of Genes and Genomes):** KEGG is a vast database that combines data from the genetic, chemical, and functional domains. It offers network diagrams, pathway maps, and functional annotations for proteins and genes. Researchers can investigate how proteins interact with other molecules in the setting of biological systems and pathways. Website: https://www.genome.jp/kegg/

MACHINE LEARNING TECHNIQUES FOR PREDICTIVE MODELING IN BIOINFORMATICS

In the discipline of bioinformatics, machine learning algorithms have become potent predictive modeling tools, providing a way to glean valuable insights from complicated biological data. Application of machine learning techniques has greatly improved bioinformatics, which entails the computational study of biological data. For tasks like anticipating protein shapes, locating disease indicators, and comprehending gene control, these methods are especially beneficial (Kumari et al., 2015). We will examine various machine learning techniques used in bioinformatics predictive modeling during this session. Sequence analysis is one of the main uses of machine learning in bioinformatics. Large datasets of DNA, RNA, or protein sequences can be used to train machine learning algorithms to discover patterns, motifs, and functional elements. Sequence-based classifiers, for instance, may foretell if a given DNA sequence contains a specific regulatory region or encodes a specific protein. These models can be used to annotate unidentified sequences because they rely on features retrieved from the sequences, like nucleotide or amino acid composition (Sartor et al., 2019).

Machine learning methods are extremely helpful for structural bioinformatics, especially when predicting protein structures. Protein 3D structures can be predicted with astonishing accuracy using techniques like AlphaFold. These models increase our understanding of protein function, relationships, and drug development by fusing deep learning with knowledge from existing protein structures. Functional annotation also heavily relies on machine learning. Machine learning algorithms can classify genes according to their functions in biological processes or disease pathways by examining gene expression data. As a result, potential therapeutic targets or illness biomarkers can be found. Furthermore, machine learning may combine many data sources, including genomic, transcriptomic, and proteomic data, to offer a comprehensive understanding of gene function (Li et al., 2020).

Another area, where machine learning excels is the prediction of protein-protein interactions. Comprehension cellular processes and signaling pathways requires a comprehension of these connections. To anticipate probable protein interactions, machine learning models can be trained on experimental data or features obtained from protein sequences and structures (Mucaki et al., 2019). This knowledge is crucial for dissecting intricate biological networks. Machine learning expedites the identification of potential drug candidates in the context of drug research. Machine learning methods are used in virtual screening to give high binding affinity molecules the highest priority. Virtual screening is the computational screening of compounds against therapeutic targets. This can result in the identification of new medicines and cuts down on the time and expense of trial screening (Ballard et al., 2021).

Additionally, tailored medicine benefits greatly from machine learning. Predictive models can assist in customizing treatment strategies by examining specific patient data, including genomes and clinical records. For instance, machine learning can forecast a patient's reaction to a particular cancer drug, assisting in treatment planning and enhancing results (Zhao et al., 2020). However, there are still issues with the use of machine learning in bioinformatics. Among the concerns that require careful study are those relating to data quality, model interpretability, and ethical dilemmas. The development of strong machine learning methods that can manage big data is a current research horizon as biological data continues to expand in scope and complexity.

Future Directions

Applications of bioinformatics in protein analysis are expected to grow increasingly complex and extensive as the field develops further. Combining machine learning (ML) and artificial intelligence (AI) methods with conventional bioinformatics instruments is one exciting avenue. The accuracy of protein structure prediction models may be greatly improved by this hybrid strategy, particularly for proteins that are inherently disordered and difficult to predict using existing

techniques. Moreover, the creation of more sophisticated protein-protein interaction modeling algorithms may shed light on the complex webs of biological functions and reveal new targets for treatment.

Future research should focus on integrating multi-omics data into bioinformatics analyses, which is another crucial area. Through the utilization of data from genomics, proteomics, metabolomics, and transcriptomics, scientists can acquire a comprehensive comprehension of protein functions and interactions in relation to whole biological systems. The development of precision medicine techniques and our understanding of complex diseases may both benefit from this multi-omics approach.

Finally, it is imperative that more diverse and underrepresented organisms be included in the bioinformatics toolkit. Investigating the proteomes of non-model organisms can reveal special proteins with unusual roles, providing fresh perspectives on evolutionary biology as well as possible uses in biotechnology and medicine. The future of bioinformatics in protein analysis holds the promise of solving the molecular puzzles of life and expanding the boundaries of science and medicine as computing power and data storage capacities keep rising.

CONCLUSIONS

Bioinformatics tools have become crucial resources in the field of protein analysis, spanning the prediction of protein structures, modeling of complex interactions, and clarification of function linkages. Our comprehension of proteins has reached previously unheard-of levels because to these technologies, which were created through the collaboration of biology, computer science, and mathematics. Structural biology has been transformed by the development of breakthroughs like AlphaFold, which demonstrate the extraordinary accuracy with which protein structures can be predicted. Such developments not only shed light on the complex protein architecture but also open up new avenues for the development of innovative therapies and the search for new drugs. Bioinformatics techniques have been important in deciphering the intricate network of cellular processes in the context of protein-protein interaction modeling. These models help identify new therapeutic targets and clarify illness processes by offering essential insights into the principles governing biological systems. Bioinformatics has also significantly improved the prediction and understanding of protein function connections. Bioinformatics tools enable researchers to maneuver the challenging terrain of genomics and proteomics, whether it be annotating newly sequenced proteins or figuring out their functions within biological contexts. The field of bioinformatics is still developing and broadening its frontiers as we stand at the nexus of biology and computational science. These techniques continue to be at the cutting edge of scientific advancement with the introduction of customized medicine, the acceleration of drug development, and the quest for greater understanding of the molecular foundation of life. Bioinformatics will continue to support academics and biochemists in the next years, assisting us in gaining a deeper comprehension of proteins and their roles. It is a journey characterized by creativity, teamwork, and a steadfast dedication to expanding our understanding of the complex world of proteins, one byte of data at a time.

Author contributions: TTO & CDG-O: contributed to writing sections on protein structure prediction methods & bioinformatics tools in protein analysis as well as to literature research, data collecting, & analysis; **ICO, ADA, & FIO:** contributed to data analysis & writing sections of machine learning techniques for predictive modeling in bioinformatics; **JVO & EOI:** aided in development & data analysis for bioinformatics algorithms for predicting protein structure; **DOA, CEA, NM-AO:** contributed to writing & data analysis of modeling of protein-protein interactions; & **FOO, EOO, & ODU:** contributed to writing sections of functional annotation methods and databases. All co-authors agree with the results and conclusions.

Funding: No funding source is reported for this study.

Acknowledgments: The authors would like to tahnk Department of Biochemistry at Ladoke Akintola University of Technology for their support.

Ethical statement: The authors stated that ethics committee approval was not required for the work, therefore it was exempted. The study involves data collection using online resources involving information freely available in the public domain that does not collect or store identifiable data. All related laws, rules, and regulations required for the study's implementation have been followed. The authors further stated that the article is the original study of the authors, and it has not been published elsewhere.

Declaration of interest: No conflict of interest is declared by the authors.

Data sharing statement: Data supporting the findings and conclusions are available upon request from corresponding author.

REFERENCES

Ali, A. M., Atmaj, J., Van Oosterwijk, N., Groves, M. R., & Dömling, A. (2019). Stapled peptides inhibitors: A new window for target drug discovery. *Computational and Structural Biotechnology Journal, 17*, 263-281. https://doi.org/10.1016/j.csbj.2019.01.012

AlQuraishi, M. (2020). A watershed moment for protein structure prediction. *Nature, 577*(7792), 627-628. https://doi.org/10.1038/d41586-019-03951-0

Bajpai, A. K., Davuluri, S., Tiwary, K., Narayanan, S., Oguru, S., Basavaraju, K., Dayalan, D., Thirumurugan, K., & Acharya, K. K. (2019). How helpful are the protein-protein interaction databases and which ones? *bioRxiv*. https://doi.org/10.1101/566372

Ballard, Z., Brown, C., Madni, A. M., & Ozcan, A. (2021). Machine learning and computation-enabled intelligent sensor design. *Nature Machine Intelligence, 3*(7), 556-565. https://doi.org/10.1038/s42256-021-00360-9

Ban, X., Lahiri, P., Dhoble, A. S., Li, D., Gu, Z., Li, C., Cheng, L., Hong, Y., Li, Z., & Kaustubh, B. (2019). Evolutionary stability of salt bridges hints its contribution to stability of proteins. *Computational and Structural Biotechnology Journal, 17*, 895-903. https://doi.org/10.1016/j.csbj.2019.06.022

Behl, A., & Mishra, P. (2018). Three-dimensional structure of Plasmodium falciparum knob associated heat shock protein 40 predicted by homology modeling method. *The Pharma Innovation Journal, 7*, 202-205.

Bekker, G.-J., Araki, M., Oshima, K., Okuno, Y., & Kamiya, N. (2020). Exhaustive search of the configurational space of heat-shock protein 90 with its inhibitor by multicanonical molecular dynamics based dynamic docking. *Journal of Computational Chemistry, 41*(17), 1606-1615. https://doi.org/10.1002/jcc.26203

Bergenholm, D., Liu, G., Holland, P., & Nielsen, J. (2018). Reconstruction of a global transcriptional regulatory network for control of lipid metabolism in yeast by using chromatin immunoprecipitation with lambda exonuclease digestion. *mSystems, 3*(4). https://doi.org/10.1128/msystems.00215-17

Bolyen, E., Ram Rideout, J., Chase, J., Anders Pitman, T., Shiffer, A., Mercurio, W., Dillon, M. R., & Caporaso,J. G. (2018). An introduction to applied bioinformatics: A free, open, and interactive text. *Journal of Open Source Education, 1*(5), 27. https://doi.org/10.21105/jose.00027

Cava, C, & Castiglioni, I. (2020). Integration of molecular docking and in vitro studies: A powerful approach for drug discovery in breast cancer. *Applied Sciences, 10*(19), 6981. https://doi.org/10.3390/app10196981

Cetin, B., Song, G. J., & O'Leary, S. E. (2020). Heterogeneous dynamics of protein-RNA interactions across transcriptome-derived messenger RNA populations. *Journal of the American Chemical Society, 142*(51), 21249-21253. https://doi.org/10.1021/jacs.0c09841

Cezar, H. M., Canuto, S., & Coutinho, K. (2020). DICE: A Monte Carlo code for molecular simulation including the configurational bias Monte Carlo method. *Journal of Chemical Information and Modeling, 60*(7), 3472-3788. https://doi.org/10.1021/acs.jcim.0c00077

Chao, F.-A., & Byrd, R. A. (2018). Protein dynamics revealed by NMR relaxation methods. *Emerging Topics in Life Sciences, 2*(1), 93-105. https://doi.org/10.1042/etls20170139

Chen, K.-H., Wang, T.-F., & Hu, Y.-J. (2019). Protein-protein interaction prediction using a hybrid feature representation and a stacked generalization scheme. *BMC Bioinformatics, 20*(1). https://doi.org/10.1186/s12859-019-2907-1

Chen, M., Lin, X., Lu, W., Schafer, N. P., Onuchic, J. N., & Wolynes, P. G. (2018). Template-guided protein structure prediction and refinement using optimized folding landscape force fields. *Journal of Chemical Theory and Computation, 14*(11), 6102-6116. https://doi.org/10.1021/acs.jctc.8b00683

Cheung, N. J., & Yu, W. (2018). De novo protein structure prediction using ultra-fast molecular dynamics simulation. *PLoS ONE, 13*(11), e0205819. https://doi.org/10.1371/journal.pone.0205819

Crosara, K. T. B., Moffa, E. B., Xiao, Y., & Siqueira, W. L. (2018). Merging in-silico and in vitro salivary protein complex partners using the STRING database: A tutorial. *Journal of Proteomics, 171*, 87-94. https://doi.org/10.1016/j.jprot.2017.08.002

Davis, R. G., Park, H.-M., Kim, K., Greer, J. B., Fellers, R. T., LeDuc, R. D., Romanova, E. V., Rubakhin, S. S., Zombeck, J. A., Wu, C., Yau, P. M., Gao, P., van Nispen, A. J., Patrie, S. M., Thomas, P. M., Sweedler, J. V., Rhodes, J. S., & Kelleher, N. L. (2018). Top-down proteomics enables comparative analysis of brain proteoforms between mouse strains. *Analytical Chemistry, 90*(6), 3802-3810. https://doi.org/10.1021/acs.analchem.7b04108

de Medeiros, A. D., Capobiango, N. P., da Silva, J. M., da Silva, L. J., da Silva, C. B., & dos Santos Dias, D. C. F. (2020). Interactive machine learning for soybean seed and seedling quality classification. *Scientific Report, 10*(1). https://doi.org/10.1038/s41598-020-68273-y

De Paris, R., Vahl Quevedo, C., Ruiz, D. D., Gargano, F., de Souza, O. N. (2018). A selective method for optimizing ensemble docking-based experiments on an InhA Fully-Flexible receptor model. *BMC Bioinformatics, 19*(1). https://doi.org/10.1186/s12859-018-2222-2

Defoort, J., Van de Peer, Y., & Carretero-Paulet, L. (2019). The evolution of gene duplicates in angiosperms and the impact of protein-protein interactions and the mechanism of duplication. *Genome Biology and Evolution*. https://doi.org/10.1093/gbe/evz156

Fang, G., Annis, I. E., Elston-Lafata, J., & Cykert, S. (2019). Applying machine learning to predict real-world individual treatment effects: Insights from a virtual patient cohort. *Journal of the American Medical Informatics Association, 26*(10), 977-988. https://doi.org/10.1093/jamia/ocz036

Frezza, E., & Lavery, R. (2019). Internal coordinate normal mode analysis: A strategy to predict protein conformational transitions. *The Journal of Physical Chemistry B, 123*(6), 1294-1301. https://doi.org/10.1021/acs.jpcb.8b11913

Fukuda, H., & Tomii, K. (2020). DeepECA: An end-to-end learning framework for protein contact prediction from a multiple sequence alignment. *BMC Bioinformatics, 21*(1). https://doi.org/10.1186/s12859-019-3190-x

Gao, M., Zhou, H., & Skolnick, J. (2019). DESTINI: A deep-learning approach to contact-driven protein structure prediction. *Scientific Reports, 9*, 3514. https://doi.org/10.1038/s41598-019-40314-1

Gebert, D., Jehn, J., & Rosenkranz, D. (2019). Widespread selection for extremely high and low levels of secondary structure in coding sequences across all domains of life. *Open Biology, 9*(5). https://doi.org/10.1098/rsob.190020

Gemovic, B., Sumonja, N., Davidovic, R., Perovic, V., & Veljkovic, N. (2019). Mapping of protein-protein interactions: Web-based resources for revealing interactomes. *Current Medicinal Chemistry, 26*(21), 3890-3891. https://doi.org/10.2174/0929867325666180214113704

Gouthami, K, Veeraraghavan, V, Rahdar, A, Bilal, M, Shah, A, Rai, V., Gurumurthy, D. M., Ferreira, L. F. R., Amrico-Pinheiro, J. H. P., Murari, S. K., Kalia, S., & Mulla, S. I. (2022). WITHDRAWN: Molecular docking used as an advanced tool to determine novel compounds on emerging infectious diseases: A systematic review. *Progress in Biophysics and Molecular Biology*. https://doi.org/10.1016/j.pbiomolbio.2022.10.001

Hao, W., Wang, Y., & Liang, W. (2018). Slice-based building facade reconstruction from 3D point clouds. *International Journal of Remote Sensing, 39*(20), 6587-6606. https://doi.org/10.1080/01431161.2018.1463113

Harada, R. (2018). Simple, yet efficient conformational sampling methods for reproducing/predicting biologically rare events of proteins. *Bulletin of the Chemical Society of Japan, 91*(9), 1436-1450. https://doi.org/10.1246/bcsj.20180170

Hiranuma, N., Park, H., Baek, M., Anishchenko, I., Dauparas, J., & Baker, D. (2021). Improved protein structure refinement guided by deep learning based accuracy estimation. *Nature Communications, 12*(1). https://doi.org/10.1038/s41467-021-21511-x

Houkes, W., & Zwart, S. D. (2019). Transfer and templates in scientific modelling. *Studies in History and Philosophy of Science, 77*, 93-100. https://doi.org/10.1016/j.shpsa.2017.11.003

Huang, L.-C, Ross, K. E., Baffi, T. R., Drabkin, H., Kochut, K. J., Ruan, Z., D'Eustachio, P., McSkimming, D., Arighi, C., Chen, C., Natale, D. A., Smith, C., Gaudet, P., Newton, A. C., Wu, C., & Kannan, N. (2018). Integrative annotation and knowledge discovery of kinase post-translational modifications and cancer-associated mutations through federated protein ontologies and resources. *Scientific Reports, 8*, 6518. https://doi.org/10.1038/s41598-018-24457-1

Jang, W. D., Lee, S. M., Kim, H. U., & Lee, S. Y. (2020). Systematic and comparative evaluation of software programs for template-based modeling of protein structures. *Biotechnology Journal, 15*(6). https://doi.org/10.1002/biot.201900343

Jha, K., & Saha, S. (2020). Amalgamation of 3D structure and sequence information for protein-protein interaction prediction. *Scientific Reports, 10*(1). https://doi.org/10.1038/s41598-020-75467-x

Jia, K., & Jernigan, R. L. (2021). New amino acid substitution matrix brings sequence alignments into agreement with structure matches. *Proteins, 89*(6), 671-682. https://doi.org/10.1002/prot.26050

Jiang, M., Li, Z., Bian, Y., & Wei, Z. (2019). A novel protein descriptor for the prediction of drug binding sites. *BMC Bioinformatics, 20*(1). https://doi.org/10.1186/s12859-019-3058-0

Jin, S., Chen, M., Chen, X., Bueno, C., Lu, W., Schafer, N. P., Lin, X., Onuchic, J. N., & Wolynes, P. G. (2020). Protein structure prediction in CASP13 using AWSEM-suite. *Journal of Chemical Theory and Computation, 16*(6), 3977-3988. https://doi.org/10.1021/acs.jctc.0c00188

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. *PNAS, 119*(4). https://doi.org/10.1073/pnas.2113348119

Karasev, D., Sobolev, B., Lagunin, A., Filimonov, D., & Poroikov, V. (2019). Prediction of protein–ligand interaction based on the positional similarity scores derived from amino acid sequences. *International Journal of Molecular Sciences, 21*(1), 24. https://doi.org/10.3390/ijms21010024

Khalatbari, L., Kangavari, M. R., Hosseini, S., Yin, H., & Cheung, N.-M. (2019). MCP: A multi-component learning machine to predict protein secondary structure. *Computers in Biology and Medicine, 110*, 144-155. https://doi.org/10.1016/j.compbiomed.2019.04.040

Kim, J.-Y., & Chung, H. S. (2020). Disordered proteins follow diverse transition paths as they fold and bind to a partner. *Science, 368*(6496), 1253-1257. https://doi.org/10.1126/science.aba3854

Kondra, S., Sarkar, T., Raghavan, V., & Xu, W. (2021). Development of a TSR-based method for protein 3-D structural comparison with its applications to protein classification and motif discovery. *Frontiers in Chemistry, 8*. https://doi.org/10.3389/fchem.2020.602291

Kotowski, K., Smolarczyk, T., Roterman-Konieczna, I., & Stapor, K. (2021). ProteinUnet–An efficient alternative to SPIDER3-single for sequence-based prediction of protein secondary structures. *Journal of Computational Chemistry, 42*(1), 50-59. https://doi.org/10.1002/jcc.26432

Kuhlman, B., & Bradley, P. (2019). Advances in protein structure prediction and design. *Nature Reviews Molecular Cell Biology, 20*(11), 681-697. https://doi.org/10.1038/s41580-019-0163-x

Kumari, P., Nath, A., & Chaube, R. Identification of human drug targets using machine-learning algorithms. *Computers in Biology and Medicine, 56*, 175-181. https://doi.org/10.1016/j.compbiomed.2014.11.008

Kwon, Y., Shin, W.-H., Ko, J., & Lee, J. (2020). AK-score: Accurate protein-ligand binding affinity prediction using an ensemble of 3D-convolutional neural networks. *International Journal of Molecular Sciences, 21*(22), 8424. https://doi.org/10.3390/ijms21228424

Larke, M. S. C., Schwessinger, R., Nojima, T., Telenius, J., Beagrie, R. A., Downes, D. J., Oudelaar, M., Truch, J., Graham, B., Bender, M. A., Proudfoot, N. J., Higgs, D. R., & Hughes, J. R. (2021). Enhancers predominantly regulate gene expression during differentiation via transcription initiation. *Molecular Cell, 81*(5), 983-997.e7. https://doi.org/10.1016/j.molcel.2021.01.002

Lensink, M. F., Velankar, S., Baek, M., Heo, L., Seok, C., & Wodak, S. J. (2018). The challenge of modeling protein assemblies: the CASP12-CAPRI experiment. *Proteins, 86*(S1), 257-273. https://doi.org/10.1002/prot.25419

Li, C., Cesbron, F., Oehler, M., Brunner, M., & Höfer, T. (2018). Frequency modulation of transcriptional bursting enables sensitive and rapid gene regulation. *Cell Systems, 6*(4), 409-423.e11. https://doi.org/10.1016/j.cels.2018.01.012

Li, Z., Huang, Q., Chen, X., Wang, Y., Li, J., Xie, Y., Dai, Z., & Zou, X. (2020). Identification of drug-disease associations using information on molecular structures and clinical symptoms via deep convolutional neural network. *Frontiers in Chemistry, 7*. https://doi.org/10.3389/fchem.2019.00924

Lin, H.-N., & Hsu, W.-L. (2020). GSAlign: An efficient sequence alignment tool for intra-species genomes. *BMC Genomics, 21*(1). https://doi.org/10.1186/s12864-020-6569-1

Lin, T.-T., Yang, L.-Y., Lu, I.-H., Cheng, W.-C., Hsu, Z.-R., Chen, S.-H., & Lin, C.-Y. (2021). AI4AMP: An antimicrobial peptide predictor using physicochemical property-based encoding method and deep learning. *mSystems, 6*(6). https://doi.org/10.1128/msystems.00299-21

Liu, Z., Miller, D., Li, F., Liu, X., & Levy, S. F. (2020). A large accessory protein interactome is rewired across environments. *Elife, 9*. https://doi.org/10.7554/elife.62365

Lu, J., Chen, D., Wang, G., Kiritsis, D., & Torngren, M. (2022). Model-based systems engineering tool-chain for automated parameter value selection. *IEEE Transactions on Systems, Man, and Cybernetics, 52*(4), 2333-2347. https://doi.org/10.1109/tsmc.2020.3048821

Lu, W., Bueno, C., Schafer, NP., Moller, J., Jin, S., Chen, X., Chen, M., Gu, X., de Pablo, J. J., & Wolynes, P. G. (2020). OpenAWSEM with Open3SPN2: A fast, flexible, and accessible framework for large-scale coarse-grained biomolecular simulations. *bioRxiv*. https://doi.org/10.1101/2020.09.07.285759

Makigaki, S, & Ishida, T. (2019). Sequence alignment using machine learning for accurate template-based protein structure prediction. *bioRxiv*. https://doi.org/10.1101/711945

Mao, W., Ding, W., Xing, Y., & Gong, H. (2019). AmoebaContact and GDFold as a pipeline for rapid de novo protein structure prediction. *Nature Machine Intelligence, 2*(1), 25-33. https://doi.org/10.1038/s42256-019-0130-4

Marino, V., & Dell'Orco, D. (2019). Evolutionary-conserved allosteric properties of three neuronal calcium sensor proteins. *Frontiers in Molecular Neuroscience, 12*. https://doi.org/10.3389/fnmol.2019.00050

Masrati, G., Landau, M., Ben-Tal, N., Lupas, A., Kosloff, M., & Kosinski J. (2021). Integrative structural biology in the era of accurate structure prediction. *Journal of Molecular Biology, 433*(20), 167127. https://doi.org/10.1016/j.jmb.2021.167127

Milanetti, E., Trandafir, A. G., Alba, J., Raimondo, D., & D'Abramo, M. (2018). Efficient and accurate modeling of conformational transitions in proteins: The case of c-src kinase. *The Journal of Physical Chemistry B, 122*(38), 8853-8860. https://doi.org/10.1021/acs.jpcb.8b07155

Miller, E, Murphy, R, Sindhikara, D, Borrelli, K, Grisewood, M, Ranalli, F, Dixon, S., Jerome, S., Boyles, N., Day, T., Ghanakota, P., Mondal, S., Rafi, S. B., Troast, D. M., Abel, R., & Friesner, R. (2020). A reliable and accurate solution to the induced fit docking problem for protein-ligand binding. *ChemRxiv*. https://doi.org/10.26434/chemrxiv.11983845.v1

Mohamed, E. M., Mousa, H. M., & Keshk, A. E. (2018). Comparative analysis of multiple sequence alignment tools. *International Journal of Computer Science and Information Technologies, 10*(8), 24-30. https://doi.org/10.5815/ijitcs.2018.08.04

Mrozek, D., Suwała, M., & Małysiak-Mrozek, B. (2019). High-throughput and scalable protein function identification with Hadoop and Map-only pattern of the MapReduce processing model. *Knowledge and Information Systems, 60*(1), 145-178. https://doi.org/10.1007/s10115-018-1245-3

Mucaki, E. J., Zhao, J. Z. L., Lizotte, D. J., & Rogan, P. K. (2019). Predicting responses to platin chemotherapy agents with biochemically-inspired machine learning. *Signal Transduction and Targeted Therapy, 4*(1). https://doi.org/10.1038/s41392-018-0034-5

Mura, C., Veretnik, S., & Bourne, P. E. (2019). The Urfold: Structural similarity just above the superfold level? *Protein Science, 28*(12), 2119-2126. https://doi.org/10.1002/pro.3742

Nardo, A. E., Añón, M. C., & Parisi, G. (2018). Large-scale mapping of bioactive peptides in structural and sequence space. *PLoS ONE, 13*(1), e0191063. https://doi.org/10.1371/journal.pone.0191063

Nguyen, M. Q., von Buchholtz, L. J., Reker, A. N., Ryba, N. J. P., & Davidson, S. (2021). Single nucleus transcriptomic analysis of human dorsal root ganglion neurons. *bioRxiv*. https://doi.org/10.1101/2021.07.02.450845

Orbán-Németh, Z., Beveridge, R., Hollenstein, D. M., Rampler, E., Stranzl, T., Hudecz, O., Doblmann, J., Schlögwlhofer, P., & Mechtler, K. (2018). Structural prediction of protein models using distance restraints derived from cross-linking mass spectrometry data. *Nature Protocols, 13*(3), 478-494. https://doi.org/10.1038/nprot.2017.146

Peker, N., Garcia-Croes, S., Dijkhuizen, B., Wiersma, H. H., van Zanten, E., Wisselink, G., Friedrich, A. W., Kooistra-Smid, M., Sinha, B., Rossen, J. W. A., & Couto, N. (2019). A comparison of three different bioinformatics analyses of the 16S-23S rRNA encoding region for bacterial identification. *Frontiers in Microbiology, 10*. https://doi.org/10.3389/fmicb.2019.00620

Rigoldi, F., Donini, S., Redaelli, A., Parisini, E., & Gautieri A. (2018). Review: Engineering of thermostable enzymes for industrial applications. *APL Bioengineering, 2*(1). https://doi.org/10.1063/1.4997367

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., & Fergus, R. (2019). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *bioRxiv*. https://doi.org/10.1101/622803

Rodrigues, C. H. M., Myung, Y., Pires, D. E. V., & Ascher, D. B. (2019). mCSM-PPI2: predicting the effects of mutations on protein-protein interactions. *Nucleic Acids Research, 47*(W1), W338-W344. https://doi.org/10.1093/nar/gkz383

Runthala, A., & Chowdhury, S. (2019). Refined template selection and combination algorithm significantly improves template-based modeling accuracy. *Journal of Bioinformatics and Computational Biology,17*(02), 1950006. https://doi.org/10.1142/s0219720019500069

Sakhaee, N., & Wilson, M. C. (2021). Information extraction framework to build legislation network. *Artificial Intelligence and Law, 29*(1), 35-58. https://doi.org/10.1007/s10506-020-09263-3

Salinas, V. H., & Ranganathan, R. (2018). Coevolution-based inference of amino acid interactions underlying protein function. *Elife, 7*. https://doi.org/10.7554/elife.34300

Sartor, R. C., Noshay, J., Springer, N. M., & Briggs, S. P. (2019). Identification of the expressome by machine learning on omics data. *PNAS, 116*(36), 18119-18125. https://doi.org/10.1073/pnas.1813645116

Schönherr, R., Rudolph, J. M., & Redecke, L. (2018). Protein crystallization in living cells. *Journal of Biological Chemistry, 399*(7), 751-772. https://doi.org/10.1515/hsz-2018-0158

Seath, C. P., Trowbridge, A. D., Muir, T. W., & MacMillan, D. W. C. (2021). Reactive intermediates for interactome mapping. *Chemical Society Review, 50*(5), 2911-2926. https://doi.org/10.1039/d0cs01366h

Seidl, P., Renz, P., Dyubankova, N., Neves, P., Verhoeven, J., Wegner, J. K., Segler, M., Hochreiter, S., & Klambauer, G. (2022). Improving few- and zero-shot reaction template prediction using modern Hopfield networks. *Journal of Chemical Information and Modeling, 62*(9), 2111-2120. https://doi.org/10.1021/acs.jcim.1c01065

Sejdiu, B. I., & Tieleman, D. P. (2021). ProLint: A web-based framework for the automated data analysis and visualization of lipid–protein interactions. *Nucleic Acids Research, 49*(W1), W544-W550. https://doi.org/10.1093/nar/gkab409

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Zidek, A., Nelson, A. W. R., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D. T., Silver, D., Kavukcuoglu, K., & Hassabis, D. (2020). Improved protein structure prediction using potentials from deep learning. *Nature, 577*(7792), 706-710. https://doi.org/10.1038/s41586-019-1923-7

Siebenmorgen, T., & Zacharias, M. (2020). Efficient refinement and free energy scoring of predicted protein-protein complexes using replica exchange with repulsive scaling. *Journal of Chemical Information and Modeling, 60*(11), 5552-5562. https://doi.org/10.1021/acs.jcim.0c00853

Singh, P., & Singh, N. (2021). Role of data mining techniques in bioinformatics. *International Journal of Applied Research in Bioinformatics, 11*(1), 51-60. https://doi.org/10.4018/ijarb.2021010106

Song, J., Li, F., Takemoto, K., Haffari, G., Akutsu, T., Chou, K.-C, & Webb, G. I. (2018). PREvaIL, an integrative approach for inferring catalytic residues using sequence, structural, and network features in a machine-learning framework. *Journal of Theoretical Biology, 443*, 125-137. https://doi.org/10.1016/j.jtbi.2018.01.023

Souza, P. C. T., Limongelli, V., Wu, S., Marrink, S. J., & Monticelli, L. (2021). Perspectives on high-throughput ligand/protein docking with Martini MD simulations. *Frontiers in Molecular Biosciences, 8*. https://doi.org/10.3389/fmolb.2021.657222

Stacey, R. G., Skinnider, M. A., Chik, J. H. L., & Foster, L. J. (2018). Context-specific interactions in literature-curated protein interaction databases. *BMC Genomics, 19*(1). https://doi.org/10.1186/s12864-018-5139-2

Sun, P., Tan, X., Guo, S., Zhang, J., Sun, B., Du, N., Wang, H., & Sun, H. (2018). Protein function prediction using function associations in protein–protein interaction network. *IEEE Access, 6*, 30892-30902. https://doi.org/10.1109/access.2018.2806478

Tong, Q., Xue, L., Lv, J., Wang, Y., & Ma, Y. (2018). Accelerating CALYPSO structure prediction by data-driven learning of a potential energy surface. *Faraday Discuss, 211*, 31-43. https://doi.org/10.1039/c8fd00055g

Truong, D. T., & Li, M. S. (2018). Probing the binding affinity by jarzynski's nonequilibrium binding free energy and rupture time. *The Journal of Physical Chemistry B, 122*(17), 4693-4699. https://doi.org/10.1021/acs.jpcb.8b02137

van Beusekom, B., Joosten, K., Hekkelman, M. L., Joosten, R. P., &Perrakis, A. (2018). Homology-based loop modeling yields more complete crystallographic protein structures. *IUCrJ, 5*(5), 585-594. https://doi.org/10.1107/s2052252518010552

Vignani, R., Liò, P., & Scali, M. (2019). How to integrate wet lab and bioinformatics procedures for wine DNA admixture analysis and compositional profiling: Case studies and perspectives. *PLoS ONE, 14*(2), e0211962. https://doi.org/10.1371/journal.pone.0211962

Volkov, M., Turk, J.-A., Drizard, N., Martin, N., Hoffmann, B., Gaston-Mathé, Y., & Rognan, D. (2022). On the frustration to predict binding affinities from protein–ligand structures with deep neural networks. *Journal of Medicinal Chemistry, 65*(11), 7946-7958. https://doi.org/10.1021/acs.jmedchem.2c00487

Wang, H., & Yang, W. (2019). Toward building protein force fields by residue-based systematic molecular fragmentation and neural network. *Journal of Chemical Theory and Computation, 15*(2), 1409-1417. https://doi.org/10.1021/acs.jctc.8b00895

Wang, T., Qiao, Y., Ding, W., Mao, W., Zhou, Y., & Gong, H. (2019). Improved fragment sampling for ab initio protein structure prediction using deep neural networks. *Nature Machine Intelligence, 1*(8), 347-355. https://doi.org/10.1038/s42256-019-0075-7

Watanabe, G., Eimura, H., Abbott, N. L., & Kato T. (2020). Biomolecular binding at aqueous interfaces of Langmuir monolayers of bioconjugated amphiphilic mesogenic molecules: A molecular dynamics study. *Langmuir, 36*(41), 12281-12287. https://doi.org/10.1021/acs.langmuir.0c02191

Weng, G., Gao, J., Wang, Z., Wang, E., Hu, X., Yao, X., Cao, D., & Hou, T. (2020). Comprehensive evaluation of fourteen docking programs on protein–peptide complexes. *Journal of Chemical Theory and Computation, 16*(6), 3959-3969. https://doi.org/10.1021/acs.jctc.9b01208

Wojtowicz, W. M., Vielmetter, J., Fernandes, R. A., Siepe, D. H., Eastman, C. L., Chisholm, G. B., Cox, S., Klock, H., Anderson, P. W., Rue, S. M., Miller, J. J., Glaser, S. M., Bragstad, M. L., Vance, J., Lam, A. W., Lesley, S. A., Zinn, K., & Garcia, K. C. (2020). A human IgSF cell-surface interactome reveals a complex network of protein-protein interactions. *Cell, 182*(4), 1027-1043.e17. https://doi.org/10.1016/j.cell.2020.07.025

Xu, X., Yan, C., & Zou, X. (2018a). MDockPeP: An ab-initio protein-peptide docking server. *Journal of Computational Chemistry, 39*(28), 2409-2413. https://doi.org/10.1002/jcc.25555

Xu, Y., Wang, S., Hu, Q., Gao, S., Ma, X., Zhang, W., Shen, Y., Chen, F., Lai, L., & Pei, J. (2018b). CavityPlus: A web server for protein cavity detection with pharmacophore modelling, allosteric site identification and covalent ligand binding ability prediction. *Nucleic Acids Research, 46*(W1), W374-W379. https://doi.org/10.1093/nar/gky380

Yan, L., Sun, W., Lu, Z., & Fan, L. (2020). Metagenomic next-generation sequencing (mNGS) in cerebrospinal fluid for rapid diagnosis of Tuberculosis meningitis in HIV-negative population. *International Journal of Infectious Diseases, 96*,270-275. https://doi.org/10.1016/j.ijid.2020.04.048

Yerneni, S., Khan, I. K., Wei, Q., & Kihara, D. (2018). IAS: Interaction specific GO term associations for predicting protein-protein interaction networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 15*(4), 1247-1258. https://doi.org/10.1109/tcbb.2015.2476809

Yim, S., Yu, H., Jang, D., & Lee, D. (2018). Annotating activation/inhibition relationships to protein-protein interactions using gene ontology relations. *BMC Systems Biology, 12*(S1). https://doi.org/10.1186/s12918-018-0535-4

Yu, C.-H., Qin, Z., Martin-Martinez, F. J., & Buehler, M. J. (2019). A self-consistent sonification method to translate amino acid sequences into musical compositions and application in protein design using artificial intelligence. *ACS Nano, 13*(7), 7471-7482. https://doi.org/10.1021/acsnano.9b02180

Zhang, C., Zheng, W., Freddolino, P. L., & Zhang, Y. (2018). MetaGO: Predicting gene ontology of non-homologous proteins through low-resolution protein structure prediction and protein-protein network mapping. *Journal of Molecular Biology, 430*(15), 2256-265. https://doi.org/10.1016/j.jmb.2018.03.004

Zhang, M. M., Beno, B. R., Huang, R. Y.-C., Adhikari, J., Deyanova, E. G., Li, J., Chen, G., & gross, M. L. (2019a). An integrated approach for determining a protein–protein binding interface in solution and an evaluation of hydrogen–deuterium exchange kinetics for adjudicating candidate docking models. *Analytical Chemistry, 91*(24), 15709-15017. https://doi.org/10.1021/acs.analchem.9b03879

Zhang, P., Shen, L., & Yang, W. (2019b). Solvation free energy calculations with quantum mechanics/molecular mechanics and machine learning models. *The Journal of Physical Chemistry B, 123*(4), 901-908. https://doi.org/10.1021/acs.jpcb.8b11905

Zhang, Y., Aryee, A. N. A., & Simpson, B. K. (2020). Current role of in silico approaches for food enzymes. *Current Opinion in Food Science, 31*, 63-70. https://doi.org/10.1016/j.cofs.2019.11.003

Zhao, B., Katuwawala, A., Oldfield, C. J., Dunker, A. K., Faraggi, E., Gsponer, J., Kloczkowski, A., Malhis, N., Mirdita, M., Obradovic, Z., Söding, J., Steinegger, M., Zhou, Y., & Kurgan, L. (2021a). DescribePROT: Database of amino acid-level protein structure and function predictions. *Nucleic Acids Research, 49*(D1), D298-D308. https://doi.org/10.1093/nar/gkaa931

Zhao, L., Ciallella, H. L., Aleksunes, L. M., Zhu, H. (2020). Advancing computer-aided drug discovery (CADD) by big data and data-driven machine learning modeling. *Drug Discovery Today, 25*(9), 1624-1638. https://doi.org/10.1016/j.drudis.2020.07.005

Zhao, T., Liu, J., Zeng, X., Wang, W., Li, S., Zang, T., Peng, J., & Yang, Y. (2021b). Prediction and collection of protein-metabolite interactions. *Briefings in Bioinformatics, 22*(5). https://doi.org/10.1093/bib/bbab014

Zheng, W., Wuyun, Q., Li, Y., Mortuza, S. M., Zhang, C., Pearce, R., Ruan, J., & Zhang, Y. (2019). Detecting distant-homology protein structures by aligning deep neural-network based contact maps. *PLOS Computational Biology, 15*(10), e1007411. https://doi.org/10.1371/journal.pcbi.1007411

Zhou, J., Panaitiu, A. E., & Grigoryan, G. (2020). A general-purpose protein design framework based on mining sequence–structure relationships in known protein structures. *PNAS, 17*(2), 1059-1068. https://doi.org/10.1073/pnas.1908723117

Zhou, P., Wang, J., Wang, M., Hou, J., Lu, J. R., & Xu, H. (2019). Amino acid conformations control the morphological and chiral features of the self-assembled peptide nanostructures: Young investigators perspective. *Journal of Colloid and Interface Science, 548*, 244-254. https://doi.org/10.1016/j.jcis.2019.04.019