Advances in Computational Biology: A Comprehensive Review of Emerging Technologies and Applications

[1]M. Vinitha, [2]S.K. Preethika
[1]Assistant Professor, [2]II M.Sc Mathematics
[1]Department of Mathematics,
[1]Dr.N.G.P. Arts and Science College, Coimbatore, India
[1]vinithamaths09@gmail.com, [2]preethiakshu04@gmail.com

doi.org/10.64643/JATIRV2I1-140110-001

Abstract- Artificial intelligence and machine learning developments applied to large biological datasets have significantly changed computational biology. This review methodically summarizes paradigm-shifting approaches that are transforming biological research in the areas of quantum computing, systems-level integration, variant detection, protein structure prediction, and foundation models. We review twenty papers published between 2021 and 2025, demonstrating the importance of computational techniques in deriving useful insights from high dimensional biological data. Data standardization, model transferability, algorithmic interpretability, and computational accessibility are still major issues. Quantum algorithm development, polypharmacology prediction, and rational protein engineering are future priorities. This review illustrates how computational methods radically alter bio logical research paradigms and facilitate the development of precision medicine.

Index-Terms- computational biology, artificial intelligence, machine learning, deep learning, protein structure prediction, drug discovery, bioinformatics, systems biology

I. INTRODUCTION

Over the past ten years, artificial intelligence, high-performance computing, and the exponential growth of biological data have all contributed to the revolutionary transformation of computational biology. Researchers can now decode complex biological systems with previously unheard-of accuracy and speed thanks to the field’s advanced algorithms, machine learning frameworks, and deep learning architectures.

This review looks at the state of the art in computational biology today, emphasizing cutting-edge approaches, creative uses, and new areas that are changing our knowledge of molecular, cellular, and organismal biological processes. The emergence of omics tech neologies transcriptomics, proteomics, metabolomics, and genomics has produced enormous datasets that are difficult for conventional statistical methods to handle. As a result, computational approaches are now essential for deriving significant biological insights from these intricate datasets. This review summarizes recent developments in computational biology in a variety of fields, with a focus on revolutionary methods such as integrative multi-omics analysis, quantum computing applications in drug discovery, foundation models for biological sequences, and deep learning-based protein structure prediction.

Unprecedented opportunities for biological discovery have been made possible by the convergence of computational power and algorithmic innovation. Computational screening has made it possible to complete tasks that previously required years of laboratory experimentation in a matter of hours or days. Certain facets of biological research have become more accessible as a result of this paradigm shift, but it has also brought forth new difficulties with regard to data standardization, model validation, and fair access to computational resources. Researchers, physicians, and legislators who want to use these potent tools to advance biomedical knowledge and enhance human health must have a thorough understanding of the state of computational biology today.

II. DEEP LEARNING PARADIGMS IN PROTEIN STRUCTURE PREDICTIONS

The advent of AlphaFold2, created by DeepMind researchers at Alphabet, caused a paradigm shift in the field of structural bioinformatics. Protein structure prediction was revolutionized by this innovative deep learning architecture, which in most cases achieved accuracy comparable to experimental structures. The root-mean-square deviation (RMSD) of AlphaFold2 was 0.8 ngstrms, an order of magnitude improvement over the previous best-performing method, which was 2.8 Angstroms. Innovative neural network designs, such as evolved transformer modules, equivariant attention mechanisms, and iterative refinement processes driven by pairwise features and multiple sequence alignments, are incorporated into the architecture.

It is impossible to exaggerate the importance of this accomplishment. Despite decades of research effort, protein structure prediction from amino acid sequences remained an un solvable computational problem. Traditional approaches including homology modeling, ab initio structure prediction, and hybrid methods required extensive computational resources and frequently failed for proteins lacking homologous templates. Through creative deep learning architecture design, AlphaFold2 overcame these constraints and achieved high accuracy across a variety of protein families, including difficult cases that have historically been prone to prediction failure.

Recent advancements, such as AlphaFold3, have broadened the coverage of structure prediction beyond only proteins to also include biomolecular complexes (nucleic acids, small molecules, and metal ions), using diffusion-based methods to refine predictions to produce structures that can occur in nature. Diffusion-based techniques let you use generative modelling to create an object, and then apply constraints of structure to produce the most biologically realistic model. The use of foundational models in conjunction with these predictive models has also improved the overall quality of the predictions, so that they could apply to a larger number of biological systems.

The Rose TTA Fold All-Atom technology represents the next significant advancement in Computational Structural Biology. This technology provides a unified frame work that will incorporate a variety of bio macromolecular Input Types (Proteins, Nucleic Acids, Ligands, and Covalent Modifications) to predict the structure and design multi-state protein complexes. The framework for the All-Atoms structure is based on the RoseTTAFold2, which uses the following 3 ways to process input data: 1D sequence, 2D pairwise distance matrix from homologous templates to represent 3D structure and the iterative refinement layers to improve the accuracy of the computational model. Improvements include a ’Protein Generator’, a new type of sequence space diffusion model, which allows you to generate new protein sequences guided by structural characteristics, which will help design thermostable proteins with desired functional properties

III. FOUNDATION MODELS AND SELF-SUPERVISED LEARNING IN BIO INFORMATICS

In bioinformatics, there has been a fundamental change with the advent of Foundation Models; they rely on massive amounts of biological data that were not previously labelled in order to learn about biological entities. There are several types of Foundation Models such as Language Foundation Models, Vision Foundation Models, Graph Foundation Models, and Multimodal Foundation Models. Foundation Models have been shown to perform exceptionally well in many downstream applications such as Genomic, Transcriptomic, Proteomic, and Drug Discovery and Single Cell Analysis.

Biological sequence language models offer genomic/proteomic researchers an effective, accurate way to perform their biological analysis. Pre-trained on large datasets of genomic sequences, DNABERT has demonstrated state-of-the-art accuracy in identifying regulatory elements (e.g., promoters, splice sites, and transcription factor binding regions). The long-range dependencies and hierarchical relationships inherent in the biological data that DNABERT captures were previously missed by feature-based analysis techniques. The other (RNA) models, RNA-FM and RNA-MSM, excel in accurately predicting RNA’s secondary and tertiary structures by leveraging self-supervised learning.

The development of Protein BERT, Protein-Evolutionary Scale Modeling (ESM), and similar models has created an efficient environment for performing many different kinds of protein predictions. Protein BERT has demonstrated performance approaching current ”state of the art” levels on multiple benchmark datagram datasets related to protein structure, post-translational modification, and biophysical characteristics, and it does so with substantially smaller and faster models than other methods. Protein BERT uses a combined pre-training strategy that includes language modeling along with Gene Ontology (GO) annotation prediction; this enables the model to learn functional characters tics in addition to learning sequence patterns. Through the process of few-shot learning, these models provide excellent generalization capabilities even when trained using small amounts of labeled data and are thus ideally suited for use in protein engineering and functional prediction tasks.

Foundation models have fundamentally transformed the way we compute by moving away from the traditional method of training models for only one task, to using models that are trained on data and can quickly adapt to new tasks. The ability of couple's foundation models with domain-specific expertise allows for even more flexibility in using foundation models for biomechanical research.

IV. ARTIFICIAL INTELLIGENCE APPLICATIONS IN VARIANT DETECTION AND GENOME ANALYSIS

Deep Learning techniques have revolutionized the manner in which genetic variation is identified through the analysis of sequenced data. Traditional variant calling tools, which are rule-based, depend on statistical limits set beforehand and manually designed features for reference. Due to these constraints, the effectiveness of such tools has been diminished in some regions of a genome, especially in relation to repetitive elements and highly polymorphic locations on chromosomes. Several AI-driven variant calling systems utilize both Convolutional Neural Networks (CNNs) and Deep Neural Networks (DNNs) to more effectively analyze sequencing data than previously possible. Examples of AI based variant calling systems include Deep Variant, DNA scope, Deep Trio, Clair, and Medaka.

Deep Variant displays remarkable similarity across a variety of species; Models developed using Human Genome data perform well when utilized to make calls on Mouse and other organism Genomes. This Models ability to perform well across all species suggests that Deep Learning is learning universal patterns of how Sequencing Errors occur and how Valid Variants look as opposed to learning unique characteristics of a particular species. Compared with other typologies of Variant Callers that use Artificial Intelligence (AI) and traditional (GATK and SAM) callers, Comparative Benchmarking Studies have shown that AI Variant Callers outperform traditional Calling Methods for multiple metrics: Sensitivity, Specificity, and Computational Efficiency. The greatest difference between AI and Traditional Methods occurs in Areas of the Genome that have been historically difficult to call with a high rate of False Positives or False Negatives.

Computational methods are now able to perform Pan genomics analysis and com parative genomic analysis of a much larger number of varieties and species than was previously possible by using a method that is based on single variants. Pan genome based machine learning models utilize graph based representations of genomes to capture genetic variation across many different species and populations. These methods allow the identification of genetic variants that are unique to specific populations and the functional genetic differences that result in various forms of susceptibility to certain diseases and response to drug therapies as well as adaptive traits. The ability of High Performance Computing combined with advanced computational techniques such as machine learning to analyses thousands of independent whole genomes simultaneously provides an unprecedented opportunity to perform very large-scale population genomics projects that are practical and reproducible.

V. SINGLE CELL TRANSCRIPTOMICS AND ADVANCED RNA ANALYSIS

Single-cell RNA sequencing has transformed the way we think about diversity among individual cells and how genes behave over time. With newer techniques like NASC-seq, we can measure newly made RNA through the use of chemical tagging along with computer algorithms that give us an estimate of how fast/how often ”bursts” of transcription (the process by which DNA is copied into RNA) occur in a cell. Using new statistical models to analyze these RNA counts allows us to gain unique insights into the dynamics of transcription and also how cells change from one state to another.

Single-cell analysis has its own set of computational challenges over and above those in bulk sequencing, including sparsity of single-cell data, inter-batch variability related to differences in sequencing methods and platforms, as well as feature extraction through dimensionality reduction to retain biologically relevant variability. Large knowledge bases of single-cell data sets learned by foundation models can better represent cellular data to help infer cell type classification, trajectory analysis, or learning novel cellular states. They generalize better than previous approaches to standard cell type classification primarily for varying tissue types or states of development and/or diseases.

More recently, the integration of single cell transcriptomics with complementary modalities, including chromatin accessibility assessed by ATAC-seq, protein abundance profiled by CITE-seq, and spatial location probed by spatial transcriptomics requires sophisticated computational frameworks for multi-modal data alignment and integration. Shared embedding spaces and deep generative models are emerging approaches to integrate across these modalities, enabling holistic views of cellular organization and regulatory mechanisms. These multi-modal approaches have revealed previously unknown cell types, developmental trajectories, and disease-associated cellular states.

VI. MACHINE LEARNING IN DRUG DISCOVERY AND REPURPOSING

Machine learning technology has significantly speeded up the discovery process in pharmaceutical research by facilitating the rapid screening of compound libraries, computer models estimating interactions between target proteins and medicines, and selection of the most promising candidate compound(s) among many others using artificial intelligence technology. Conventional high-through put screening involves the synthesis or preparation and subsequent experimental analysis of tens of millions of compounds, which cannot be practically achieved with current time and budget limitations.

Drug repurposing, which refers to finding new uses for existing approved drugs, is a particularly promising application area. Machine learning frameworks build heterogeneous biomedical graphs that integrate drug attributes, target proteins, disease links, and gene expression patterns. Graph neural networks, along with various topology-based approaches, can identify drug candidates through relationships within these graphs to make predictions for novel links between drugs and disease-related genes/proteins. More con temporary uses of AI, such as Deep Drug, integrate expertise in appropriate biomedicine related domains to improve predictions.

Computational metabolomics is another rapidly emerging area within the domain of drug development. Combining data from mass spectrometry and nuclear magnetic resonance techniques with machine learning techniques leads to the discovery of metabolic biomarkers, the prediction of molecular interactions, and the discovery of new metabolites of drugs. Multi-scale modeling approaches use molecular docking, quantum mechanics modeling, and machine learning methods to formulate drugs and predict the efficacy and toxicity of such drugs without entering the experimental phase.

VII. QUANTUM COMPUTING APPLICATIONS IN BIOLOGICAL SYSTEMS

Quantum computing is a revolutionary technological area on the threshold of breaking the limitation of computations, hindering the development of bio informatics. Quantum computing processors use optimization techniques, Variation Quantum Eigen solver, or quantum annealing, designed for complex computations not feasible on a computer. Bio informatics application areas of quantum computing include simulation, protein structures, and optimization of compound libraries.

POLARIS Quantum has engineered the first drug development platform based on quantum annealing, thereby converting drug design into optimization problems that can be addressed through the wave function analysis provided by quantum mechanics. The process allows for efficient exploration of vast chemical spaces for molecules with imperative pharmaceutical properties in significantly fewer computational steps than the traditional brute force method. Recent examples include successful quantum computer execution of drug design tasks like molecular docking and prediction of RNA secondary structure.

Integrating the capabilities of quantum computing with the existing classical high performance computing systems has made the resulting hybrid systems capable of tackling complex biological research questions. Currently, near term applications in the research area include the simulation of the binding of protein ligands, the modeling of the flexibility of molecules, and the prediction of the metabolism of drugs, but the existing limitations in the current state of the art in quantum computing will have to be overcome.

VIII. COMPUTATIONAL CRISPR/CAS9 OFF-TARGET PREDICTION

The CRISPR/Cas9 system is a recent addition to the realm of genomic research and holds tremendous potential as a therapeutic revolution. However, the problem of ”off target cutting,” or the potential for the CRISPR/Cas9 system to target and cleave genomic DNA away from the desired site of action, is a major issue that needs be resolved for this system to have potential as a therapeutic technique. Off-target cutting needs modern computational modeling that is capable of recognizing subtle regions within the genome with extremely high similarity for the guide RNA target and taking into consideration the specificity because of the PAM.

Modern computer-based solutions rely on three main approaches: sequence alignment based solutions involving Bowtie, BWA, for detection of regions of homologous genome locations; rule-based solutions involving machine learning, where features of nucleotide sequence and experimental evidence for scores of off-target locations are used; and deep learning approaches, where CRISPR-Net, a method involving recurrent neural networks, convolutional networks, aims for the location of term-based nucleotide sequence correlations. Deep learning solutions offer high sensitivity with high accuracy through machine training on a set of nucleotide sequence locations detected by experiment-based evidence.

The adenine base editors and cytosine base editors involve the use of advanced computational models for on-target conversion rates as well as the prevention of bystander editing and off-target events. The use of machine learning models based on the concepts of gradient boosting and deep conditional auto-regressive models can predict on-target conversion rates and bystander editing signatures.

IX. GRAPH NEURAL NETWORKS FOR BIOLOGICAL NETWORK ANALYSIS

Biomedical systems have inherent network structures; complex interactions happen at multiple scales ranging from protein-protein interactions to ecological networks. Graph Neural Networks form efficient models that can be applied for extracting and analyzing these networks. Graph Neural Networks process graph structured data; node embedding’s learn local and global graph structure information.

Applications in bio informatics include:

1. Prediction of diseases through fusion of multi-omics data and biological networks.
2. Discovery of drugs through link prediction to discover new target-drug interactions.
3. Discovery of biomarkers through node classification.
4. Graph generation models for predicting molecular properties.
5. GNN models for graph convolutional networks, graph attention networks, and message passing neural networks allow for biological reasoning which is unseen from biological sequences and structure alone.

New paradigms utilize more complex topologies of networks such as graph lets and hyper graphs, which describe interactions among more than two entities, because of the nature of biological systems, where functional entities are characterized by the synchronized action of more than two molecules. Topology-driven models enable the characterization of dysregulated sub-networks within disease states and condition-specific path alterations. Network analysis combined with machine learning methods has helped decode mechanisms of diseases, which were hidden from conventional methods of understanding.

X. COMPUTATIONAL APPROACHES TO EPI GENOMICS AND GENE REGULATION

Epi-genetic modifications such as DNA methylation and histone post-translational modifications play pivotal roles in gene regulation via mechanisms independent of DNA sequence and play critical roles in development and disease. The computational models using machine learning and deep learning algorithms facilitate genome-wide discovery of epigenetic loci associated with diseases. EWAS plus is a supervised machine learning algorithm with high accuracy as a binary classifier. It detects cytosine-guanine dinucleotides associated with disease using ensemble learning with either regularized logistic regression and/or gradient boosting decision trees trained on genome-wide association summary statistics.

Deep learning techniques such as convolutional neural networks and recurrent neural networks perform well in terms of finding patterns in epi genomic sequencing techniques such as chromatin immune precipitation-sequencing, assay for transposase accessible chromatin sequencing, and so on. Deep learning techniques overcome the limitations imposed by feature designs in machine learning techniques, where they learn features on their own from raw input data. Advanced deep learning models show impressive output for predicting transcription factors and chromatin state in various cell types and stages.

Incorporation of intensive epi genomic data, along with genomic and genetic data, has led to a clearer understanding of disease etiology and discovery of new therapeutic targets. Machine learning architectures that learn a common representation of given epi genomic data modalities have aided the discovery of 'epigenetic fine-mapping' signals delineating causal regulatory variants underlying disease association. Novel therapeutic targets have been discovered for cancer, neurological disease, and developmental disease using these methods.

XI. METAGENOMICS AND MICROBIOME COMPUTATIONAL ANALYSIS

Meta genomic analysis of complex microbial communities offers incomparable insights into microbiome composition and diversity, as well as microbiome functional capabilities. Model bio informatics pipelines involve several steps: quality processing and error correction; taxonomic classification by sequence alignment and marker genes; functional classification using homologies and alignment-free techniques; and systems biology analysis using pathway integration and metabolic models.

Genome-resolved metagenomics allows the reconstruction of entire or nearly entire microbial genomes from shotgun sequencing data, independent of culture. The technique involves connoting formation from short reads, with subsequent clustering based on composition, coverage, and tetra nucleotide signatures for metagenome-assembled genome formation. The technique relies on computer programs like Meta BAT, with graph-based algorithms for genome clustering, making it feasible to characterize previously non-cultivable bacteria using classic microbiological methods. Recent breakthroughs also allow for the assessment of genetically different variants in certain species at the strain level.

Metabolomics analysis enables computational metabolomics to unravel bioactive Meta bolites and metabolic networks of communities. The combination of genome-resolved Meta genomics and mass spectrometry and metabolomics analysis facilitates the connection of particular microbial taxa to the biosynthesis of bio medically significant metabolites. Metagenomics and metabolomics have transformed our strategies for understanding the role of the gut microbial community in human biology and disease and have uncovered microbial metabolites responsible for mediating immune homeostasis and metabolism.

XII. NATURAL LANGUAGE PROCESSING AND BIOMEDICAL TEXT MINING

The biomedical literature is an immense body of knowledge with decades of data. Exhaustively abstracting data from all these publications would be impossible. However, text mining and natural language processing assist in searching and abstracting data from biomedicine publications. The processes involve searching protein-protein interactions and gene disease correlations from PubMed abstract and full-text articles. The other examples of viable applications involve abstracting interactions of medications and predicting new gene functions with literature contexts.

Newly emerging approaches combine transformer-based language models trained on biomedical datasets such as PubMed BERT and Bio BERT with knowledge from biology domains. The transformer models address terms of biological concepts and their complicated relationships that are invisible in generic models of language. The method of applying variant interpretation combines variants and their phenotypic effects associated through clinical cases mined from reports of cases and association and function studies reported in bio-literate studies.

Multimodal models using both sequence data and text information facilitate reasoning about different data modes. ProtST and other models can leverage both protein sequences and their biomedical text descriptions to improve models of either type or gain biological knowledge. These models enable the discovery of new functions or interactions in proteins based on their characteristics and their descriptions using biomedical text information in the literature. Biomedical NLP use has enabled rapid discovery of knowledge, particularly new associations between genes and certain diseases.

XIII. MULTI-OMICS INTEGRATION AND SYSTEMS BIOLOGY

Contemporary biological science tends to produce increasing amounts of multi omics biological data from samples or populations of interest in biotechnology, such as genomics, transcriptomics, proteomics, metabolomics, and epigenomes. Multi omics biological data analysis through computers integrates different biological data for a comprehensive analysis of biological systems that go beyond single biological analysis in omics.

Computational multi-omics analysis techniques use various methodologies such as dimensionality reduction methods that identify key axes of variation across data modalities; network-based methods where molecules are represented as graph nodes and edges represent their relationship; and deep learning methods that learn joint latent spaces across data modalities. Such methods tackle various technical difficulties associated with multi omics datasets such as data sparsity, batch effects, and missing values. Deep learning based multi-omics analysis methods learn joint latent spaces of multiple data modalities. Such models learn an additional latent space that captures common characteristics of multiple data modalities.

Systems biology methods include modeling biological networks and processes to predict cell function and drug responses. System-based models, which simulate protein protein interactions, signaling pathways, and gene regulatory networks, assimilate known biological information with experimental data. Current systems biology methods have utilized machine learning to infer model parameters. This has minimized requirements for exhaustive biological characterization of kinetic parameters. Novel drug targets and individual patient responses to drugs have thus been predictively established by systems biology.

XIV. HIGH-PERFORMANCE COMPUTING AND EXASCALE INFRASTRUCTURE

Today, the size of available biological data calls for the need for exascale computing capabilities, with a computer performance of 1018 floating-point operations per second. Today, advanced silicon technology developed with the concept of Application Specific Integrated Circuits, Wafer Scale Engines, and GPU-accelerated Computing Nodes has ensured the provision of specialized hardware solutions for bio informatics tools. These solutions offer faster processing capabilities for genome sequence analysis, biomarker discovery, biomolecular simulation, and virtual screening of small molecules.

Cloud computing platforms make high computing power accessible for researchers without access to a supercomputing infrastructure. Para Bricks accelerates tasks like read alignment, variant calling, and base quality re computation by a factor of 30-50 using the GPU. These computing platforms make it feasible to process tens of thousands of genomes together, a process that was not feasible a few years ago.

The development of the digital twin technology for the generation of virtual replicas of biological systems, or organisms, is another upcoming use case based on supercomputing infrastructure. The models developed use large-scale biological data based on various omics technologies and help in the prediction of the systems performances in different scenarios. The digital twin may revolutionize personalized medicine in the future.

XV. CHALLENGES, LIMITATIONS AND FUTURE PERSPECTIVES

Although tremendous progress has been made, the field of computational biology is still faced with certain limitations that impede broader adoption and translation into the clinic. There are concerns regarding the quality, variability, and representation of results affected by batch effects, platform variability, and missing data. However, the generalization of machine learning models trained on particular groups or datasets tends to fail when applied to other settings, thus limiting the potential of the models across various ethnic groups. There is a lack of interpretability and explain ability of complex models like machine learning and deep learning.

Despite their power, the computational demands and energy usage of large models have become a challenge to their sustainability. This limits accessibility to these models. Training these models also demands a lot of computational power beyond what a research group could attain. This could lead to disparities in scientific breakthrough discoveries. The future should see more efficient models using fewer parameters.

Frontiers that are emerging and demanding continued innovation include rational de sign of multistate proteins with complex functionalities, predicting DDIs and poly pharmacology, incorporation of structural knowledge into sequences and functions for better protein design, and the development of efficient quantum algorithms for biologically relevant problems. The meeting of frontiers in computational biology and novel emerging technologies in quantum computing, microscopy, and DNA Storage offers a historic set of opportunities for fundamental discovery in the next ten years. It is critical that continued investment in infrastructure and training of interdisciplinary talent be made in order to fulfill the promise of computational biology in improving human health and our knowledge of biology.

XVI. CONCLUSION

Computational biology has fundamentally transformed biomedical research through advances in artificial intelligence, machine learning, and high-performance computing. This review establishes that computational methodologies are indispensable for extracting insights from complex biological datasets across protein structure prediction, drug discovery and genomics. The integration of foundation models, deep learning, and quantum computing has established new paradigms enabling researchers to address previously intractable biological problems with unprecedented efficiency.

Significant challenges persist: data standardization, model generalization, algorithmic interpretability, and equitable computational accessibility require continued attention. Future advancement necessitates developing efficient architectures, standardizing data formats, and fostering interdisciplinary collaboration.

Opportunities for revolutionary discoveries have never been so great, thanks to the merger of the field of computational biology with new technologies. New paradigms for research will be established in biology, as these new approaches become pervasive, pushing the frontiers of precision medicine.

REFERENCES

[1] Abdelwahab. O, El-Attar. M. M., El-Saadany. E. F, Artificial intelligence in variant calling: A review, Frontiers in Bioinformatics, Vol. 5, pp. 1-10, 2025.

[2] Accelerating Biology 2025 Conference, Compute to transcend: High-performance computing in biological sciences., C-DAC Research Initiative, 2025.

[3] Biotecnika, Top emerging trends in bioinformatics, 2025.

[4] Brandes. N, Ofer. D, Peleg. Y, Linial. M, Linial. N, ProteinBERT: A universal deep-learning model of protein sequences, Bioinformatics, Vol.8, pp. 2102-2110, 2022.

[5] Capela.J, Haddad. R, Cosentino. S, Kamieniarz. K, Frey. B. S, Comparative assessment of protein large language models, Bioinformatics, 2025.

[6] DeepDrug Research Team, DeepDrug as an expert guided and AI driven drug repurposing methodology for selecting the lead combination of drugs for Alzheimer's disease, Scientific Reports, Vol. 15, 2025.

[7] Garcia Alcalde. F, García López. F, Dopazo. J, Conesa. A, Bioinformatics tools for functional analysis of microarray and RNA sequencing data, Next-Generation Sequencing Technologies and Applications, Academic Press, pp. 405-426, 2024.

[8] Ghandikota. S. K., Radhakrishnan. N, Pawar. R, Kale. R. K, (2024). Application of artificial intelligence and machine learning in drug repurposing, Drug Discovery Today, 2024.

[9] Guo. F, Liu. Q, Han. Y, Cao. Y, Theodoris. C. V, Huang. J, et al, Foundation models in bioinformatics, PMC - PubMed Central, 2025.

[10] Huang. Y, Chu. L, Reilly. S. M, A machine learning approach to brain epigenetic analysis reveals disease-associated CpGs, Nature Communications, Vol. 12, 2021.

[11] Imperial College London Research Team, Towards using quantum computing to speed up drug development, Nature Computational Science, Imperial College News Release, 2023.

[12] Jumper. J, Evans. R, Pritzel. A, Green. T, Figurnov. M, Ronneberger. O, et al, Highly accurate protein structure prediction with AlphaFold, Nature, pp. 583-589, 2021.

[13] Junjun. R., Chen. H., Yang. X, Liu. S, A comprehensive review of deep learning-based variant calling and detection, Briefings in Functional Genomics, 2024.

[14] Kim. N, Zhang. Y, Li. X, Guo. T, Walker. A. W, Genome-resolved metagenomics: A game changer for microbiome research, Nature Reviews Microbiology, 2024.

[15] Krishna. R, Wang. J, Gipe. R. A, Lutz. N, Movshovitz-Attias. D, Leung. S, Generalized biomolecular modeling and design with RoseTTAFold All-Atom, pp. 384-394, 2024.

[16] Lisanza. S. L, Gershman. S. J, Baker. D, Brady. S. C., Multistate and functional protein design using RoseTTAFold, Nature Biotechnology, 2024.

[17] POLARIS Quantum Research Team, Harnessing the power of quantum computing for drug discovery, Nature Computational Science, 2023.

[18] Ramsköld. D, Lim. K, Shim. K, Mahmood. F, Zhu. J, Eng. C. H, Single-cell new RNA sequencing reveals principles of transcriptional bursting, Nature, Vol. 8, pp. 1486-1499, 2024.

[19] Yen. S, Li. X, Chen. J, Konkel. M. E, Hess. M, Metagenomics: A path to understanding the gut microbiome, Journal of Microbiology, 2021

[20] Zitnik. M, Agrawal. M, Leskovec. J, Current and future directions in network biology. Bioinformatics Advances, Vol. 1, 2024.