REVIEW PAPER

Machine Learning in Computational Biology: Emerging Applications in Drug Discovery and Drug Delivery Systems

Shatrupa Singh[1], Anistha[1], Kanchan Yadav[1], Nilesh Yadav[1] and Megha Bansal[2*]

[1]Department of Biotechnology, Delhi Technological University, Delhi, India.
[1]Department of Biotechnology, Delhi Technological University, Delhi, India.
[1]Department of Biotechnology, Delhi Technological University, Delhi, India.
[2]Department of Biotechnology, Graphic Era University, Dehradun, India.

Corresponding Author:
Megha Bansal

ABSTRACT

Machine learning methodologies have been introduced into computational biology, which have radically changed the paradigms in biological research over the last decade. This paper analyzes the primary applications of machine learning in various fields of computational biology, including protein structure prediction, genomic sequence analysis, drug discovery, and systems biology. We evaluate the theoretical basis of these applications, their implementation, and the problems associated with applying computational learning algorithms to biological data. Special focus is given to deep learning structures that have proven exceptionally successful in the modelling of complex biological systems. We also examine recent developments, such as explainable artificial intelligence in medical biology, federated learning for conducting privacy-preserving medical studies, and the use of machine learning models to integrate multi-omics data. This thorough examination can present researchers with information on the existing methodologies, coupled with the promotion of prospective studies.

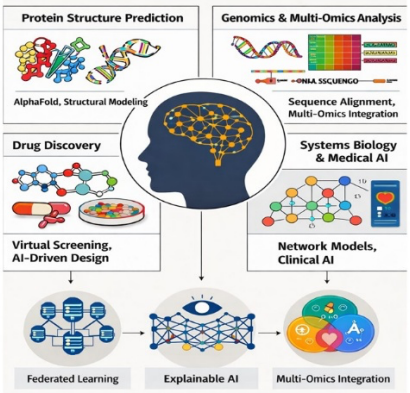
Keywords: Machine learning, computational biology, deep learning, protein structure prediction, drug discovery

How to cite this article: Shatrupa Singh, Anistha, Kanchan Yadav, Nilesh Yadav, Megha Bansal., Machine Learning in Computational Biology: Emerging Applications in Drug Discovery and Drug Delivery Systems..Int J Drug Deliv Technol. 2026;16(12s): 107-120. DOI: 10.25258/ijddt.16.12s.12

Source of support: Nil.

Conflict of interest: None

Graphical Abstract



INTRODUCTION

Biological systems are highly complex systems that involve intricate interactions between molecules, regulatory mechanisms, and emergent properties that cannot be easily analyzed using standard tools. The massive increase in biological information produced by high-throughput technologies has not only made it possible to perform computational analyses but also required more complex analysis models that can identify any meaningful pattern in large and diverse datasets [1]. Machine learning in this regard has become an invaluable asset, extending immense benefits in pattern recognition, predictive modelling, and knowledge inferences in the biological field.

Machine learning has not been applied to biology, and one of the first usages of neural networks was to predict protein secondary structure in the 1990s [2]. In recent years, the

*Author for Correspondence: megha.bt@geu.ac.in

complexity of algorithms, computational hardware, and large repositories of biological data have led to a revival in this interdisciplinary area. Modern machine learning methods have been able to overcome previously difficult-to-solve problems, as in the case of the near-experimental accuracy of AlphaFold when predicting protein structures [3].

This review summarizes existing information on the use of machine learning in computational biology, discussing known procedures and new developments. We divide our discourse into large areas of application without losing sight of both the algorithmic principles and practical issues. In our analysis, we have found that machine learning has replaced its functions as a tool of analysis and is now becoming a part of the biological research pipeline, which governs experimental design, hypothesis generation, and mechanistic understanding.

1.1 Scope and Organization

This is a review of supervised, unsupervised and reinforcement learning paradigms that are implemented in a biological environment. It is a questioning program in the spectrum of molecular to systems-level biology, including genomics, proteomics, drug discovery, and medical diagnostics. The paper is structured in a chronological order, building the theoretical concepts, underpinning the arguments to a state-of-the-art application, and finally, an evaluation of the current limitations and future research directions. Each part contrasts theoretical foundation with practical case studies, hence providing the reader with a comprehensive knowledge of both methodological dogmas as well as practical applications of those to biological use.

Statement of Significance

Category	Description
Problem or Issue	Biological systems exhibit extraordinary complexity with intricate molecular interactions that traditional computational approaches struggle to model. The exponential growth of high-throughput biological data demands sophisticated analytical frameworks capable of extracting meaningful patterns from large, heterogeneous datasets while maintaining biological interpretability.
What is Already Known	Machine learning has demonstrated success in specific biological applications including protein secondary structure prediction, genomic classification, and basic sequence analysis. Existing reviews typically focus on individual application domains or specific algorithmic approaches, lacking comprehensive integration across computational biology's diverse landscape.
What this Paper Adds	This review provides a unified framework connecting supervised, unsupervised, and reinforcement learning paradigms across molecular to systems-level biology. It critically evaluates state-of-the-art applications including AlphaFold2, transformer-based regulatory models, and federated learning while addressing persistent challenges in interpretability, generalization, and data quality. The work uniquely bridges theoretical foundations with practical implementations and identifies emerging directions including foundation models, multi-omics integration, and causal inference methodologies.
Who Would Benefit	Computational biologists seeking comprehensive methodological guidance; experimental biologists interested in leveraging machine learning for hypothesis generation; drug discovery researchers exploring AI-driven design; clinical investigators applying precision medicine approaches; and interdisciplinary teams requiring integrated perspectives on algorithm selection, validation strategies, and translational considerations for diverse biological research contexts.

2. Fundamental Machine Learning Paradigms in Biology

2.1 Supervised Learning Approaches

Supervised learning is the foundation of a multiplicity of biological studies, in which algorithmic models are trained on annotated datasets to provide a mapping between input features and predefined output labels. In the biological context, supervised learning has been proposed to be especially effective in classification tasks, such as disease diagnosis, gene functional annotation, and prediction of protein-ligand binding affinities [4] (Table 1).

Support vector machines (SVMs) are one of the first successful applications of supervised learning to biological data. Their ability to work in high-dimensional spaces makes them particularly appropriate to genomic studies, where the number of variables can be much larger than the number of samples. The SVMs have been beneficial in a continuum of applications where they are utilized in the classification of cancer using gene-expression profiles or the location of transcription-factor binding-sites [5]. The

SVM's kernel trick can be used to model nonlinear relationships without necessarily transforming the features, an attribute that is especially useful when it comes to modelling complicated biological interactions.

The ensemble approaches include random forests and gradient-boosting machines, which are a combination of multiple decision trees to achieve a high level of predictive power. The methods are good at capturing nonlinear relationships and feature interactions, and retaining interpretability as well as feature-importance measures. Random forests have also been used in the context of biological studies by predicting variant effect, where heterogeneous genomic annotations are combined and used to assess the functional consequences of genetic variants [6]. The inherent potential of tree-based models to support missing data points and mixed types of variables, which are common to biological data, fits the nature of biological data that often has missing or heterogeneous information.

2.2 Deep Learning Architectures

Deep neural networks have also played an important role in computational biology by allowing end-to-end learning of hierarchical representations on raw data. Convolutional neural networks (CNNs) have been demonstrated to be useful in the analysis of biological data in a sequential format, such as DNA and protein sequences, which autonomously identify motifs and higher-order patterns without any manual feature engineering [7]. Another added benefit of CNNs is that they can be applied across species and hence are used to discover regulatory factors that are located in different genomic areas.

The recurrent neural networks (RNNs) and their variation, the long short-term memory (LSTM) networks, are effective at modelling sequential dependencies among biological data. They have been used to model protein sequences, in which structural and functional characteristics are regulated by long-range interactions between amino acids. The ability of the LSTM to remember over long sequences removes the difficulty of identifying relations of hundreds of residues in proteins or thousands of nucleotides in regulatory regions [8].

Transformer architecture signifies a shift in the paradigm of sequence modelling since it uses self-attention-based learning to learn relationships between sequence elements regardless of their relative positions. Transformers have also shown outstanding performance in protein language modelling in biology; pre-trained transformers trained on millions of protein sequences have been shown to gain evolutionary and structural constraints, which can be fine-tuned to downstream tasks [9]. The explicability of transformer attention mechanisms explains the sequence positions with the highest contribution to predictions, providing the researchers with practical hypotheses about the compositions of functional elements.

2.3 Unsupervised Learning and Dimensionality Reduction

The annotation of biological data is often insufficient, and unsupervised learning methods are commonly required as one of the essential aspects of exploratory analysis and pattern discovery. The use of clustering algorithms, including k-means, hierarchical, and density-based clustering, is regularly used to classify cell types in single-cell RNA-sequencing datasets, to define disease subtypes and to cluster genes with similar expression patterns, a practice that has been in use for many years [10].

High-throughput biological measurements are also plagued by the curse of dimensionality, which is usually overcome using dimensionality-reduction tools. PCA offers a linear transformation process, which retains the highest amount of variance, and as such, it is used to visualize and interpret very complex data structures. Non-linear methods, such as t-stochastic neighbor embedding (t-SNE) and uniform manifold approximation and projection (UMAP), have become the techniques of choice to visualize high-dimensional single-cell experiments, uncovering the patterns of cellular heterogeneity and developmental trajectories [11].

Autoencoders, along with their probabilistic versions, variational autoencoders (VAEs), are deep-learning methods of dimensionality reduction that are trained to learn compact representations of biological data without structural distortions. VAEs have, in particular, been useful in modelling single-cell gene expression because they minimize technical and biological confounding signals and can realistically generate them, allowing the expression profiles to be recreated [12].

Table 1: Machine Learning Paradigms and Their Applications in Computational Biology

Paradigm	Key Algorithms	Biological Applications	Advantages	Limitations
Supervised Learning	SVMs, Random Forests, CNNs, Transformers	Disease diagnosis, variant classification, protein-ligand binding prediction	High accuracy with labelled data, interpretable feature importance	Requires extensive labelled datasets, prone to overfitting

Unsupervised Learning	K-means, Hierarchical clustering, PCA, t-SNE, UMAP, VAEs	Cell type identification, disease subtyping, dimensionality reduction	No labels required, discovers hidden patterns	Difficult to validate, subjective interpretation
Deep Learning	CNNs, RNNs, LSTMs, GANs, Attention mechanisms	Protein structure prediction, genomic sequence analysis, and drug design	Learns hierarchical features, handles complex patterns	Black-box nature, computationally intensive, requires large datasets
Reinforcement Learning	Q-learning, Policy gradients, Actor-Critic	Protein engineering, de novo drug design, experimental design	Optimizes sequential decisions, learns from feedback	Requires well-defined reward functions, sample inefficient

3. Genomics and Sequence Analysis

3.1 Variant Calling and Interpretation

The determination and identification of genomic variants are key issues in population genetics and precision medicine. The methods based on machine learning have enhanced the accuracy of variant calling through the combination of several sources of evidence, including read-alignment quality, base-quality scores, and local sequence context. Convolutional neural networks and similar deep-learning structures have shown that patterns of sequencing errors are highly complex, and these types of models can learn better than a traditional probabilistic model [13].

The functional effect of variants is a more challenging problem to predict, which requires the combination of evolutionary conservation, structural context, and functional annotations to determine variant pathogenicity. The prediction strategies that fuse over two of the individual predictors in a multifactorial approach are more accurate compared to single-tool methods, thus making variant interpretation a multifactorial task. Recent advances have used deep learning to combine heterogeneous annotations of the genome with good performance on distinguishing between pathogenic and benign variations [14].

3.2 Regulatory Genomics

Complex interactions between transcription factors, chromatin architecture and DNA sequence motifs regulate the transcription. Current innovations in machine learning are exploitable to forecast regulatory activity based on nucleotide sequence, where the convolutional neural networks show notable performance in defining sequence motifs of transcription factor binding sites and chromatin accessibility [15].

DeepSEA and Basset represent essential examples of the use of deep-learning methods in the field of regulatory genomics; they use primary DNA sequence to directly predict chromatin signatures and transcription factor occupancy. Such architectures derive hierarchical descriptions of regulatory grammar, thus describing discrete motifs and higher-order combinations of the motifs. Their high predictive accuracy can be used in in silico experiments of mutagenesis to enable a systematic analysis of the implications of sequence variants on regulatory competence [16].

The reconstruction of gene-expression modalities on the basis of both the sequence and the epigenome is a new configuration of research. Programs like Enformer are based on the idea of transformer models to predict transcriptional output on whole genomic loci, including hundreds of kilobases. These models have mechanistic explanations of transcriptional governance by incorporating long-range enhancer-promoter contacts [17].

3.3 Single-Cell Genomics

The inventions of single-cell sequencing technologies have significantly advanced our understanding of heterogeneity at the single-cell level, generating datasets that contain millions of single-cells with thousands of features of gene-expression annotated. These large-scale and sparse high-dimensional data require advanced machine-learning algorithms to investigate the data. One of the key analytical goals is the cell-type assignment (supervised and semi-supervised learning algorithms) with cellular identities based on the expression profiles [18].

Nevertheless, trajectory inference models based on unsupervised learning paradigms can be used to infer development trajectories and cell differentiation pathways using single-cell data. They combine dimensionality reduction with graph-based approaches to cell pseudo-temporal ordering, thus explaining dynamic processes of cells based on allegedly fixed snapshots. Most recent deep-learning progress, based on variational autoencoders, has led to a better ability to infer trajectories, where the latent representations learned are more faithful to the biological variability that they model [19].

Fixing the batch effects and ensuring strong data integration are daunting issues when integrating single-cell datasets that are obtained with different experiments or technologies. Adversarial training-based and optimal-transport-theory-based machine-learning approaches have proven to be effective in aligning such data, as well as in maintaining inherent biological diversity. They can be used to conduct a complete cross-study and cross-platform analysis [20].

4. Protein Structure and Function Prediction

4.1 Secondary and Tertiary Structure Prediction

One of the most important achievements of machine learning in computational biology is related to predictive protein structure. Initial neural-network models of secondary-structure prediction had relatively low accuracy and were unable to provide the long-range interactions required to form tertiary structures [21]. The further development of deep-learning models incorporating evolutionary data, including the assistance of multiple-sequence alignments, was a decisive breakthrough in the accuracy of structural prediction.

AlphaFold2 is a very important system, in which an attention-based neural network predicts protein structures with an accuracy equivalent to empirical approaches [22]. This architecture takes evolutionary coupling information and, at the same time, optimizes the predicted structure and the associated models of multiple-sequence alignments. A differentiable formulation of this kind allows a subtle connection between sequence, evolutionary context, and resultant structure. The functions of AlphaFold2 have significant implications for structural biology; the software can be used to clarify protein structures that other research strategies cannot access and speed up drug discovery procedures.

RoseTTAFold is another paradigm of deep-learning structure prediction, which demonstrates how different architecture paradigms can be used to reach high-accuracy results using both evolutionary and geometrical characteristics [23]. Access to structural information has been made democratic by these technological advances, and structural repositories of predicted structures have been compiled that contain large parts of the known proteome.

4.2 Protein Function Prediction

Protein functional annotation is more than just structural prediction; it requires the combination of sequence, structural, and evolutionary information to infer molecular functions, biological processes, and subcellular locations. Traditional methods using homology have often found it challenging to face homologs that are too distant or proteins that have no properly characterized homologs. The machine-learning approaches that address these limitations are able to generate complex sequence-to-function relationships based on large collections of annotated corpora.

The ability of deep learning architectures to learn functional properties has a proven ability to abstract primary sequence features of proteins. Protein language models, which are similar to those used in natural language processing, learn representations that are constrained by both functional and structural aspects through unsupervised pre-training on large datasets of sequences [24]. These representations are fine-tuned to perform task-specific prediction tasks with state-of-the-art performance on an array of functional annotation problems, including prediction of Enzyme Commission numbers, Gene Ontology term assignment and inference of subcellular localization.

Graph neural networks (GNNs) have emerged as powerful tools in structure-directed functional prediction, which predicts proteins as graphs, with each vertex representing an amino acid residue, and each edge reflecting spatial adjacency or interaction potentials. GNNs continuously update the information about the local structural neighborhoods; thus, they are structurally sensitive to functional sites. These types of methodologies have proven quite successful in catalytic residues, ligand-binding sites, and protein-protein interaction interfaces prediction [25].

4.3 Protein Design and Engineering

The reverse problem of protein design is being tackled with the increasing application of machine learning, where the goal is to produce sequences with desired three-dimensional geometry or with desired functional properties. Various forms of generative models have been trained to sample new protein sequences generated by learned probability distributions (variational autoencoders (VAEs) and generative adversarial networks (GANs)). This means that it is now easier to explore protein sequence space beyond natural diversity [26].

Recently, approaches that integrate deep learning with reinforcement learning have been used to optimize proteins to meet desired properties. These methods formulate protein engineering as a series of decision-making problems, where an agent is trained to serve proposed mutations that are beneficial, via contact with a scoring or experimental feedback. These approaches have been effective in optimizing proteins to achieve stable proteins, binding affinity, and enzyme activity [27].

5. Drug Discovery and Development

5.1 Molecular Property Prediction

Optimization of the molecular properties, including potency, selectivity, pharmacokinetics, and toxicity, is a core issue of the drug-discovery process. These properties can be predicted directly by supervised machine-learning models, and these models can be trained solely based on the underlying molecular structure, accelerating optimization

of leads through prioritization of candidate compounds. Special effectiveness has been shown in this direction using graph neural networks, which encode molecules as graph data, atomic nuclei as nodes, and chemical bonds as edges [28] (Figure 1).

The message-passing neural network (MPNN) structures generalize traditional graph-convolutional operations into the molecular space and allow obtaining very complex structure-property correlations. MPNNs derive latent representations through neighbourhood-aggregation processes, obtaining representations of the local chemical environment, as well as of the global molecular properties. It has been empirically demonstrated that MPNNs can achieve the state-of-the-art performance in predicting solubility, toxicity, and biological activity in heterogeneous chemical spaces [29].

Multi-task learning paradigms take advantage of the similarities between chemically related prediction problems, and thus improve generalization in situations where the data on every single task is sparse. Transfer learning strategies, i.e. training large-scale, pre-trained models on task-specific endpoints, have in practice provided better predictive performance compared to models trained on small amounts of task-specific data. These methodologies would particularly be beneficial in drug-discovery cases, where the evidence on the subject of novel therapeutic targets is often limited [30].

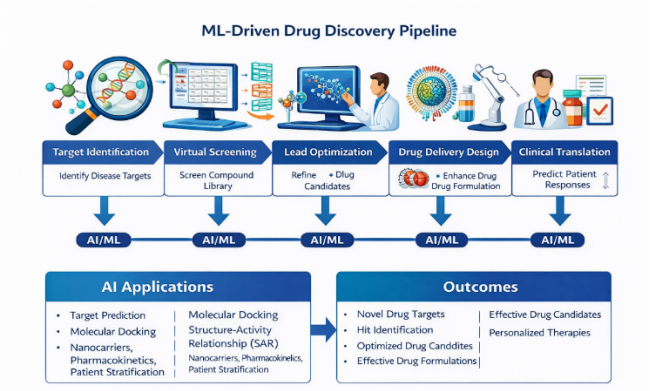


Figure 1: Schematic representation of the integration of machine learning across the drug discovery pipeline

5.2 De Novo Drug Design

Modelling Generative models enable systematized exploration of chemical space by the synthesis of new molecules with desired physicochemical properties. VAEs that learn molecular representations acquire continuous latent manifolds in which chemically similar molecules cluster in space, thus allowing for optimizing properties using gradient-based methods in the latent space [31]. Other forms of generative paradigm are adversarial autoencoders and generative adversarial networks (GANs), where the model is optimized to sample probability distributions that are similar to the ones seen in the training corpus.

Reinforcement-learning (RL) models can be used to view the synthesis of molecules as a sequence of decisions, in which a sequence of decisions is to add atoms or sub-structures to a chemical graph. The policy is guided by curated reward functions that encode physicochemical target criteria. This type of RL-based strategy has provided new chemotypes with improved qualities in various therapeutic areas [32].

Recurrent neural networks (RNNs) and transformer architectures produce molecular representations as linear strings (e.g., SMILES) with the help of methods of natural language processing. The effectiveness of these models is enhanced by massive pre-training on libraries of curated chemicals, and this equips them with insights into syntactic rules and structural motifs. The conditional generation methods also provide explicit control over the target properties and, thus, allow synthesizing entities that satisfy a combination of design characteristics at the same time [33].

5.3 Target Identification and Validation

Machine learning enables the identification of targets since it incorporates various biological data to help predict disease associations and druggability. Network-based methods are an integration of protein-protein interaction networks, gene expression data, and genetic association to rank therapeutic targets. Applications of graph neural networks to biological networks have been effective in drug-target interaction prediction and repurposing opportunities [34].

Virtual screening utilizes machine learning to estimate the binding affinity of a compound and a protein of interest and thus can be used to screen large chemical libraries quickly. Deep learning models that are trained on structural or fingerprint representations of protein-ligand complexes have a high accuracy in buy-backers versus non-buy-backers. These methods considerably decrease the number of compounds that have to be tested in experiments and, hence, can speed up the process of hits [35].

6. Medical and Clinical Applications

6.1 Disease Diagnosis and Classification

The systematic analysis of clinical data, medical imaging, and omics profiles has transformed the field of medical diagnostics profoundly, as machine learning has been applied to ensure the analysis of these types of data. Deep learning models have reached the level of expertise in detecting diabetic retinopathy, neoplastic malignancies, and classifying thoracic radiographs [36]. Convolutional neural networks learn hierarchical visual features directly by examining images, and thus, require no manual feature engineering; they also learn subtle patterns that are suggestive of pathological conditions.

Multi-modal learning is the process of creating more diagnostic accuracy by synthesising heterogeneous data modalities, such as electronic clinical records, imaging studies, and laboratory measurements. The attention processes help to identify the most influential data modalities on predictive outputs and thus provide explainable information on clinical decision-making. This type of methodology has demonstrated significant effectiveness in complex diagnostic settings where data presented by different sources is provided as complementary evidence [37].

6.2 Precision Medicine and Treatment Selection

Precision medicine aims at tailoring interventions to treat individual patients by incorporating genetic, environmental and lifestyle factors. Machine-learning models predict treatment responses by explaining the statistical associations between the features of patients and their clinical outcomes. Molecular tumor profiling has been shown to achieve significant effectiveness in cancer therapy by guiding the choice of therapy, and in pharmacogenomics, where genetic variations are used to determine optimal drug dosages [38].

The idea of survival analysis enhanced by deep-learning methods is an extension of the traditional Cox proportional-hazard models that describe a complex, non-linear correlation between patient characteristics and patient survival outcomes. These sophisticated models combine omics data of high dimension with more traditional clinical variables to put patients into distinct risk groups and give them individualized prognostic estimates. Attention-based model design and architecture promote model interpretation by detecting the most influential features that drive predictive performance [39].

6.3 Electronic Health Records Analysis

Electronic health records (EHRs) are longitudinal datasets in which the history of patients, diagnosis, pharmacologic interventions, and clinical outcomes are systematically recorded. The combination of deep learning approaches with natural language processing allows the extraction of structured data out of unstructured clinical narratives and, thus, achieves comprehensive phenotyping and predictive modelling of patient outcomes. Transformer and recurrent neural network models learn the temporal dynamics of EHR data and make predictions with regard to hospital rehospitalization, disease progression, and adverse events [40].

Privacy-conserving machine learning models, such as federated learning and differential privacy, provide the opportunity to train models in many institutions without necessarily sharing information on a patient-by-patient basis. These plans consider very important issues related to patient confidentiality and building strong models educated on heterogeneous populations. The effectiveness of federated learning in clinical prediction tasks has already been shown, showing the same performance as centralized training but preserving the security of the data [41].

7. Systems Biology and Network Analysis

7.1 Gene Regulatory Networks

Gene regulatory networks are complex networks of interrelationships among transcription factors, genes, and regulatory components. By applying machine learning models to datasets of gene expression, the networks can be inferred, and this can be used to predict regulatory relationships and identify key transcriptional regulators. Bayesian networks provide probabilistic network inferences, that is, they describe conditional relationships among genes with consideration of uncertainty in measurements [42].

Deep learning algorithms have enhanced network inference, modelled the nonlinear regulatory relationship, and integrated heterogeneous data modalities. Graph neural networks use graphs directly, and the information flows through the edges to optimize predictions of the interactions between genes. The approaches have revealed new regulatory networks and master regulators that participate in development and disease [43].

7.2 Metabolic Modelling

Constraint-based metabolic modelling uses mathematical optimization methods to predict the distribution of metabolic fluxes under specified physiological conditions. Machine learning strengthens such models by extracting objective functions directly out of empirical data, and thus, can be used to predict the growth rates and determine the important metabolic genes. Addition of the transcriptomic evidence to the metabolic reconstructions through machine learning further increases the precision of context-specific metabolic predictions [44].

Neural network models have been applied to predict metabolic fluxes with respect to the environmental variables and transcriptomic profiles, bypassing specific processes needed to solve the constraint-based optimization processes. These data-intensive models are skillful in capturing complex genotype-phenotype interactions that can be challenging to replicate in mechanistic models. Mechanistic-based hybrid methodologies that combine machine learning with mechanistic understanding have a high potential for application in the field of metabolic engineering [45].

7.3 Protein-Protein Interaction Networks

Protein-protein interaction (PPI) networks give important information on how cells are organized and map out functional modules. Machine-learning algorithms are used to deduce the interactions on the basis of sequence, structure, and expression information. Thus, the weakness

of experimental methods, missing temporal or condition-specific interactions, is reduced. These PPI-related graph neural networks have demonstrated a high accuracy in the link-prediction tasks, which allowed the discovery of new protein-protein interactions and explained the functions of proteins [46].

Network strategies that employ random walks and graph-embedding are also network-based strategies in which proteins can be encoded as vectors in continuous spaces. These embeddings enable a wide range of downstream analyses, such as functional annotation, prioritizing disease-related genes, and drug repurposing programs. Network embedding approaches that use deep-learning methods can represent more complex connectivity patterns and topological structures of higher order [47].

8. Challenges and Limitations

8.1 Data Quality and Availability

The quality and quantity of training data are vital to machine learning performance. Biological datasets often exhibit systematic biases, batch effects, and small samples compared to feature dimensionality. Noise in the form of experimental errors, technical variability and interlaboratory protocol differences can cause poor performance of the model. In biological settings, the issue of class imbalance can be quite problematic, with such infrequent cases as diseases or functional variants being a small part of the data [48].

Another serious problem is the quality of data annotation. Supervised learning requires precisely labelled training examples; however, biological annotations are usually missing, inconsistent or even outdated. Function annotations of proteins, such as those of protein location, are biased in favor of poorly studied organisms and biological processes. These biases are spread to machine-learning models, which may restrict their usage to under-studied systems [49].

8.2 Model Interpretability and Biological Insight

Despite their higher predictive power, deep learning models often act as black boxes and thus do not reveal much mechanistic understanding. This opaqueness is a significant drawback of biological research, in which predictive utility is not as important as the ability to explain mechanisms. Saliency maps, attention mechanisms, and integrated gradients provide solutions by identifying input features that make predictions, but only a limited understanding of the learned mechanisms can be achieved [50].

The ability to extract biologically relevant information about machine learning models is a point of intense research. Models' distillation, rule extraction or similar methods aim to convert advanced neural networks into a format that people can understand, but it is quite challenging to represent deep neural networks in a human-readable format. A combination of mechanistic and machine-learning models demonstrates potential to have interpretability and use data-driven inference [51].

8.3 Generalization and Transfer Learning

Machine learning models may violate their generalization capabilities when used outside of the training distributions, which becomes particularly acute when using such models in biological contexts where an experiment, a population group, and a measurement technology differ significantly. Domain shift, referring to differences between distributions when training and those in the context of real-world applications, can significantly deteriorate predictive performance. An example is when models trained on known cell lines will tend to have reduced efficacy with primary tissue samples, and models calibrated on one population could be inaccurate when used in other demographic groups [52].

Transfer learning has a possible remedy to such generalization problems because it would exploit information obtained in similar tasks or fields. One simple approach here is to pre-train on large unrestricted corpora and then fine-tune the resultant representations to particular downstream tasks, a strategy that has shown good results in a wide range of biological problems. However, it is not clear under what conditions transfer learning has substantive advantages and how to optimize the use of pre-trained models in new areas. The negative transfer, i.e. the step of pre-training, degrades the performance instead of enhancing it, and may occur in cases when the similarity between the source and target domains is not good [53].

8.4 Computational Resources and Accessibility

Training of state-of-the-art deep-learning models requires a lot of computational resources, including specialized hardware and long training times. These requirements on resources create barriers to entry for many research consortia and can result in the concentration of resources in well-funded institutions. Besides, training large models comes with a carbon footprint that creates environmental apprehensions, but it should be balanced with the future benefits of drug discovery and medical diagnostics [54].

Complexity in models also limits their use in resource-constrained environments, such as point-of-care diagnostic equipment and health-care environments with limited resources. Efforts to produce small, efficient models, through approaches like knowledge distillation, pruning and quantization, are largely intended to democratize access to machine-learning technologies [55].

9. Emerging Trends and Future Directions

9.1 Foundation Models and Self-Supervised Learning

Large-scale pretrained systems that can be trained on big datasets are known as foundation models and can be fine-tuned to a large number of downstream tasks. This

paradigm, which is effective in natural language processing, is increasingly being used in the biological context. Protein language models with a pretrained model, which are trained on millions of sequences, encode evolutionary constraints and structural principles that generalize well to heterogeneous tasks, e.g. can be used to predict structure, annotate functions, and predict variant effects [56] (Table 2).

Self-supervised learning allows models to learn using unlabeled biological data, or by solving pretext tasks: predicting masked elements of sequences and recovering corrupted data. With these strategies, the annotation bottleneck can be reduced and the richness of available sequences, including structural and imaging data stored in biological repositories, can be exploited. Since the size and complexity of foundation models are projected to keep growing, they have the potential to speed up progress in investigation by providing strong, general-purpose descriptions of biological items [57].

9.2 Multi-Modal and Multi-Omics Integration

Biological systems are highly complex in both their scale and modalities; these can be genetic sequences, cellular behavior, or organism-level phenotypes. Synchronization of various data types has become an uphill task with differences in magnitude, dimension, and technical nature. The systems-level organization is being shown as a promise by multi-modal learning methods that concurrently analyze genomics, transcriptomics, proteomics, and imaging data [58] (Figure 2).

The graph neural networks are a natural model to integrate multi-omics data, with the regulation between distinct layers of molecules being represented as nodes. The attention mechanisms allow the models to acquire knowledge about the most informative modality in a particular prediction, but also learn the interactions between modalities at the same time. Such methods have explained the new disease pathways and treatment options by using complementary types of data [59].

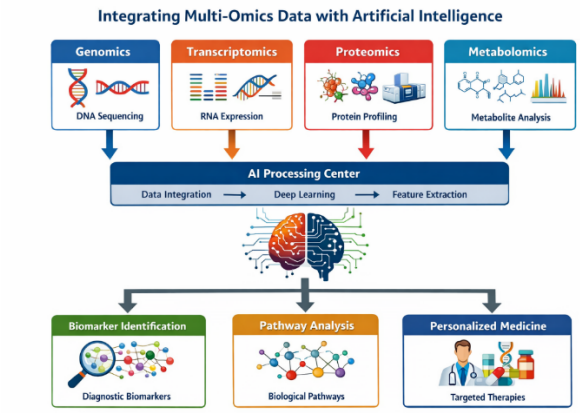


Figure 2: Integrating Multi-Omics Data with Artificial Intelligence

9.3 Causal Inference and Mechanism Discovery

Although machine learning is an effective tool for prediction, the explanation of cause-and-effect relationships is critical in biological studies and drug development. The methods of causal inference attempt to distinguish causation and correlation, and thus determine interventions that can produce desired results. Combining machine learning with causal reasoning is a new area of research that promises to enhance predictive accuracy, as well as mechanistic understanding [60].

The data from experimental perturbation provides important information to learn the causal relationships. CRISPR screens, pharmacological interventions and genetic interventions generate data sets that have causal associations. Machine-learning models trained on data including such perturbation can be calibrated to forecast the impact of interventions as well as to discover the causal links. With the combination of observational and interventional data in causal constructs, it is possible to speed up the target recognition and reduce the failure in the drug development at the late stage [61].

9.4 Active Learning and Experimental Design

Active learning alleviates the issue of data scarcity by making a strategic choice of informative examples to be experimentally validated. Instead of passively learning through the data accessible to it, the active-learning algorithms generate experiments that are expected to optimally improve the performance of the model. Active learning has been demonstrated in the context of protein engineering to be able to identify better variants with fewer experiments than traditional directed evolution methods [62].

Bayesian optimization provides an active learning methodology that balances exploration of uncharted areas and exploitation of clearly characterized promising ones. When extended to the process of drug discovery, the methodologies guide the synthesis and testing of compounds with the most desirable properties. Reinforcement learning offers other models where agents obtain the best experimental strategies by interactively operating within laboratory settings. With laboratory automation, closed-loop systems combining machine learning with robotic experimentation are promising to discover more quickly in a variety of fields in biology [63].

Table 2: Machine Learning Models in Computational Biology

Model	Year	Primary Application	Architecture Type	Performance Metric	Scientific Impact
Alpha Fold2	2021	Protein Struct	Attention-based	Near-experimental	Revolutionized structura

		ure Prediction	neural network	accuracy (GDT >90)	l biology
DeepSEA	2015	Regulatory Genomics	Convolutional Neural Network	AUC 0.90+ for chromatin features	Enabled in silico mutagenesis studies
DeepVariant	2018	Variant Calling	CNN-based classifier	F1 score >0.999 for SNPs	Improved accuracy over traditional methods
scVI	2018	Single-cell Analysis	Variational Autoencoder	Superior batch correction	Standard tool for scRNA-seq analysis
Enformer	2021	Gene Expression Prediction	Transformer-based	Predicts expression from a 200kb context	Captures long-range regulatory interactions
MPNNs	2015-2019	Molecular Property Prediction	Graph Neural Network	State-of-the-art on multiple benchmarks	Accelerated drug discovery pipelines

9.5 Quantum Machine Learning

Quantum computing has the potential to radically change a group of computational activities through utilizing quantum superposition and entanglement. Quantum machine learning is an area of study exploring the use of quantum algorithms in learning tasks, in the hope that they can be useful in optimization, sampling, and pattern recognition. In the computational biology field, quantum computing might be used to complement molecular simulations, drug discovery, and intricate biological network analysis [64].

Even though hardware limits the real-world impact of tangible quantum benefits, quantum-inspired classical algorithms have proven to be promising. These have been developed by replacing classical computational systems with the principles of quantum algorithms and thus could provide computational benefits without requiring quantum hardware. With the development of quantum computing technology, it can also be applied in biological research to help in simulating molecular systems [65].

10. Conclusion

Machine learning has been an essential tool in the field of computational biology, with demonstrated transformative impact on a continuum of research fields. In the protein structure prediction field, as well as drug discovery, in the medical diagnostics field, and in other biomedical uses, machine learning techniques have achieved breakthroughs that were inconceivable ten years ago. The combination of complex algorithms and a growing body of larger biological data promises an ongoing speed-up of discovery as well as translational activity.

However, the application of machine learning in the biological sciences presents several major challenges. Issues that relate to the quality of the data, the model interpretability, generalization ability, and the capability to compute them are also acute and require long-term methodological creativity. The field should be able to balance the need to achieve predictive performance with the production of biological understanding, such that machine learning can be more than a black box predictor, but a sensible tool to understand the biological processes.

Decentralized and fragmented approaches are likely to result in prospective advances of synergistic integration between various methodological strands. General-purpose biological representations, causal inference methods that can identify intervention points, and active learning strategies that can guide empirical validation are all suggested to influence the next generation of computational tools. Also, when mechanistic knowledge is blended into the data-driven frameworks, i.e. through hybrid architectures and physics-inspired learning, there is a possibility of obtaining approaches that will strike the right balance between the flexibility of machine learning and the interpretability of conventional modelling.

Democratizing machine-learning tools is a prerequisite for extensive scientific influence. Availability of open-source software, easy-to-use interfaces, and extensive educational materials enable any researcher to use these formidable methodologies. With the field becoming more mature, reproducibility can be ensured through the introduction of strict guidelines regarding model validation, benchmarking, and reporting, and a robust comparison between competing approaches should become possible.

The field of machine learning in computational biology is at an inflexion point and is poised to answer some of the fundamental questions concerning living systems as well as enable practical uses in medicine and biotechnology. Further partnership of machine-learning researchers and

experts in the domain will be critical towards the application of algorithmic advancements to biological insights and clinical utility. The more developed and sophisticated the methods are, the more machine learning will be influencing the way in which we explore, learn, and eventually manipulate biological systems in the best interest of humanity.

References

[1] Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nat Rev Genet 2015;16:321–32. https://doi.org/10.1038/nrg3920.

[2] Ema RR, Khatun MtA, Adnan MdN, Kabir SkS, Galib SMd, Hossain MdA. Protein Secondary Structure Prediction based on CNN and Machine Learning Algorithms. IJACSA 2022;13. https://doi.org/10.14569/IJACSA.2022.0131108.

[3] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596:583–9. https://doi.org/10.1038/s41586-021-03819-2.

[4] Eraslan G, Avsec Ž, Gagneur J, Theis FJ. Deep learning: new computational modelling techniques for genomics. Nat Rev Genet 2019;20:389–403. https://doi.org/10.1038/s41576-019-0122-6.

[5] Alharbi WS, Rashid M. A review of deep learning applications in human genomics using next-generation sequencing data. Hum Genomics 2022;16:26. https://doi.org/10.1186/s40246-022-00396-x.

[6] Yoosefzadeh-Najafabadi M, Xavier A, Eskandari M, Hesami M. Machine learning after a decade: is it still a missing keystone in genomic-based plant breeding? Artif Intell Rev 2025;58:260. https://doi.org/10.1007/s10462-025-11274-y.

[7] Rube HT, Rastogi C, Feng S, Kribelbauer JF, Li A, Becerra B, et al. Prediction of protein–ligand binding affinity from sequencing data with interpretable machine learning. Nat Biotechnol 2022;40:1520–7. https://doi.org/10.1038/s41587-022-01307-0.

[8] Hu X, Fernie AR, Yan J. Deep learning in regulatory genomics: from identification to design. Current Opinion in Biotechnology 2023;79:102887. https://doi.org/10.1016/j.copbio.2022.102887.

[9] Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proceedings of the National Academy of Sciences 2021;118:e2016239118. https://doi.org/10.1073/pnas.2016239118.

[10] Andrews TS, Kiselev VY, McCarthy D, Hemberg M. Tutorial: guidelines for the computational analysis of single-cell RNA sequencing data. Nat Protoc 2021;16:1–9. https://doi.org/10.1038/s41596-020-00409-w.

[11] Cashman D, Keller M, Jeon H, Kwon BC, Wang Q. A Critical Analysis of the Usage of Dimensionality Reduction in Four Domains. IEEE Transactions on Visualization and Computer Graphics 2025;31:9405–23. https://doi.org/10.1109/TVCG.2025.3567989.

[12] Boyeau P, Regier J, Gayoso A, Jordan MI, Lopez R, Yosef N. An empirical Bayes method for differential expression analysis of single cells with deep generative models. Proceedings of the National Academy of Sciences 2023;120:e2209124120. https://doi.org/10.1073/pnas.2209124120.

[13] Zhang H, Yin Z, Wei Y, Schmidt B, Liu W. DeepFilter: A Deep Learning Based Variant Filter for VarDict. Tsinghua Science and Technology 2023;28:665–72. https://doi.org/10.26599/TST.2022.9010032.

[14] Ghadie M, Sardaar S, Trakadis Y. Disease-Specific Prediction of Missense Variant Pathogenicity with DNA Language Models and Graph Neural Networks. Bioengineering 2025;12:1098. https://doi.org/10.3390/bioengineering12101098.

[15] Lan AY, Corces MR. Deep learning approaches for noncoding variant prioritization in neurodegenerative diseases. Front Aging Neurosci 2022;14. https://doi.org/10.3389/fnagi.2022.1027224.

[16] Liu Q, Yao E, Liu C, Zhou X, Li Y, Xu M. M2GCN: multi-modal graph convolutional network for modeling polypharmacy side effects. Appl Intell 2023;53:6814–25. https://doi.org/10.1007/s10489-022-03839-z.

[17] Pipoli V, Cappelli M, Palladini A, Peluso C, Lovino M, Ficarra E. Predicting gene expression levels from DNA sequences and post-transcriptional information with transformers. Computer Methods and Programs in Biomedicine 2022;225:107035. https://doi.org/10.1016/j.cmpb.2022.107035.

[18] Farahpour M, Safarpour H, Razavi SM, Taghipour-Gorjikolaie M. Hybrid optimized PSO-CatBoost framework for high-accuracy cell-type classification and identification in single-cell RNA-Seq data. Netw Model Anal Health Inform Bioinforma 2025;14:145. https://doi.org/10.1007/s13721-025-00626-x.

[19] Weiler P, Van den Berge K, Street K, Tiberi S. A Guide to Trajectory Inference and RNA Velocity. In: Calogero RA, Benes V, editors. Single Cell Transcriptomics: Methods and Protocols, New York, NY: Springer US; 2023, p. 269–92. https://doi.org/10.1007/978-1-0716-2756-3_14.

[20] Hrovatin K, Moinfar AA, Zappia L, Parikh S, Lapuerta AT, Lengerich B, et al. Integrating single-cell RNA-seq datasets with substantial batch effects. BMC Genomics 2025;26:974. https://doi.org/10.1186/s12864-025-12126-3.

[21] Weissenow K, Rost B. Are protein language models the new universal key? Current Opinion in Structural Biology 2025;91:102997. https://doi.org/10.1016/j.sbi.2025.102997.
[22] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596:583–9. https://doi.org/10.1038/s41586-021-03819-2.
[23] Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science 2021;373:871–6. https://doi.org/10.1126/science.abj8754.
[24] Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proceedings of the National Academy of Sciences 2021;118:e2016239118. https://doi.org/10.1073/pnas.2016239118.
[25] Gligorijević V, Renfrew PD, Kosciolek T, Leman JK, Berenberg D, Vatanen T, et al. Structure-based protein function prediction using graph convolutional networks. Nat Commun 2021;12:3168. https://doi.org/10.1038/s41467-021-23303-9.
[26] Defresne M, Barbe S, Schiex T. Protein Design with Deep Learning. International Journal of Molecular Sciences 2021;22:11741. https://doi.org/10.3390/ijms222111741.
[27] Hie BL, Yang KK. Adaptive machine learning for protein engineering. Current Opinion in Structural Biology 2022;72:145–52. https://doi.org/10.1016/j.sbi.2021.11.002.
[28] Chen J, Schwaller P. Molecular hypergraph neural networks. J Chem Phys 2024;160:144307. https://doi.org/10.1063/5.0193557.
[29] Ding K, Chin M, Zhao Y, Huang W, Mai BK, Wang H, et al. Machine learning-guided co-optimization of fitness and diversity facilitates combinatorial library design in enzyme engineering. Nat Commun 2024;15:6392. https://doi.org/10.1038/s41467-024-50698-y.
[30] Koirala M, Yan L, Mohamed Z, DiPaola M. AI-Integrated QSAR Modeling for Enhanced Drug Discovery: From Classical Approaches to Deep Learning and Structural Insight. Int J Mol Sci 2025;26:9384. https://doi.org/10.3390/ijms26199384.
[31] He C, Zhang C, Bian T, Jiao K, Su W, Wu K-J, et al. A Review on Artificial Intelligence Enabled Design, Synthesis, and Process Optimization of Chemical Products for Industry 4.0. Processes 2023;11:330. https://doi.org/10.3390/pr11020330.
[32] Korshunova M, Huang N, Capuzzi S, Radchenko DS, Savych O, Moroz YS, et al. Generative and reinforcement learning approaches for the automated de novo design of bioactive compounds. Commun Chem 2022;5:129. https://doi.org/10.1038/s42004-022-00733-0.
[33] Zou J, Zhao L, Shi S. Generation of focused drug molecule library using recurrent neural network. J Mol Model 2023;29:361. https://doi.org/10.1007/s00894-023-05772-5.
[34] Dara ON, Mohammed TA, Ibrahim AA. Evaluating the Effectiveness of Graph Convolutional Network for Detection of Healthcare Polypharmacy Side Effects. Intelligent Automation & Soft Computing 2024;39:1007. https://doi.org/10.32604/iasc.2024.058736.
[35] Meli R, Morris GM, Biggin PC. Scoring Functions for Protein-Ligand Binding Affinity Prediction Using Structure-based Deep Learning: A Review. Front Bioinform 2022;2. https://doi.org/10.3389/fbinf.2022.885983.
[36] Nigar N, Umar M, Shahzad MK, Islam S, Abalo D. A Deep Learning Approach Based on Explainable Artificial Intelligence for Skin Lesion Classification. IEEE Access 2022;10:113715–25. https://doi.org/10.1109/ACCESS.2022.3217217.
[37] Zhou Y, Huang S-C, Fries JA, Youssef A, Amrhein TJ, Chang M, et al. RadFusion: Benchmarking Performance and Fairness for Multimodal Pulmonary Embolism Detection from CT and EHR 2021. https://doi.org/10.48550/arXiv.2111.11665.
[38] Malekar S, Malekar S, Chu H. The perils of big data: understanding the usage in precision medicine. International Journal of Management Concepts and Philosophy 2023;16:89–100. https://doi.org/10.1504/IJMCP.2023.130034.
[39] Adam N, Wieder R. AI Survival Prediction Modeling: The Importance of Considering Treatments and Changes in Health Status over Time. Cancers 2024;16:3527. https://doi.org/10.3390/cancers16203527.
[40] Ramesh J, Keeran N, Sagahyroon A, Aloul F. Towards Validating the Effectiveness of Obstructive Sleep Apnea Classification from Electronic Health Records Using Machine Learning. Healthcare 2021;9:1450. https://doi.org/10.3390/healthcare9111450.
[41] Rauniyar A, Hagos DH, Jha D, Håkegård JE, Bagci U, Rawat DB, et al. Federated Learning for Medical Applications: A Taxonomy, Current Trends, Challenges, and Future Research Directions. IEEE Internet of Things Journal 2024;11:7374–98. https://doi.org/10.1109/JIOT.2023.3329061.
[42] Cutello V, Pavone M, Zito F. Inferring a Gene Regulatory Network from Gene Expression Data. An Overview of Best Methods and a Reverse Engineering Approach. In: Cantone D, Pulvirenti A, editors. From Computational Logic to Computational Biology: Essays Dedicated to Alfredo Ferro to Celebrate His Scientific

Career, Cham: Springer Nature Switzerland; 2024, p. 172–85. https://doi.org/10.1007/978-3-031-55248-9_9.

[43] Wang Y, Chen X, Zheng Z, Huang L, Xie W, Wang F, et al. scGREAT: Transformer-based deep-language model for gene regulatory network inference from single-cell transcriptomics. iScience 2024;27. https://doi.org/10.1016/j.isci.2024.109352.

[44] Kim Y, Kim GB, Lee SY. Machine learning applications in genome-scale metabolic modeling. Current Opinion in Systems Biology 2021;25:42–9. https://doi.org/10.1016/j.coisb.2021.03.001.

[45] Kong D, Qian J, Gao C, Wang Y, Shi T, Ye C. Machine Learning Empowering Microbial Cell Factory: A Comprehensive Review. Appl Biochem Biotechnol 2025;197:4897–913. https://doi.org/10.1007/s12010-025-05260-x.

[46] Gligorijević V, Renfrew PD, Kosciolek T, Leman JK, Berenberg D, Vatanen T, et al. Structure-based protein function prediction using graph convolutional networks. Nat Commun 2021;12:3168. https://doi.org/10.1038/s41467-021-23303-9.

[47] Nahid TH, Jui FA, Shill PC. Protein Secondary Structure Prediction using Graph Neural Network. 2021 5th International Conference on Electrical Information and Communication Technology (EICT), 2021, p. 1–6. https://doi.org/10.1109/EICT54103.2021.9733590.

[48] Chugh M, Chugh N. Paving the Way for Healthcare with AI, ML, and DL: Opportunities, Challenges, and Open Issues. Handbook on Augmenting Telehealth Services, CRC Press; 2024.

[49] Rembeza E, Engqvist MKM. Experimental and computational investigation of enzyme functional annotations uncovers misannotation in the EC 1.1.3.15 enzyme class. PLOS Computational Biology 2021;17:e1009446. https://doi.org/10.1371/journal.pcbi.1009446.

[50] Gündüz HA, Mreches R, Moosbauer J, Robertson G, To X-Y, Franzosa EA, et al. Optimized model architectures for deep learning on genomic data. Commun Biol 2024;7:516. https://doi.org/10.1038/s42003-024-06161-1.

[51] Karpatne A, Jia X, Kumar V. Knowledge-guided Machine Learning: Current Trends and Future Prospects 2024. https://doi.org/10.48550/arXiv.2403.15989.

[52] Bansal S, Sindhi V, Singla BS. Exploration of Deep Learning and Transfer Learning Techniques in Bioinformatics. Applying Machine Learning Techniques to Bioinformatics: Few-Shot and Zero-Shot Methods, IGI Global Scientific Publishing; 2024, p. 238–57. https://doi.org/10.4018/979-8-3693-1822-5.ch013.

[53] Niu S, Liu Y, Wang J, Song H. A Decade Survey of Transfer Learning (2010–2020). IEEE Transactions on Artificial Intelligence 2020;1:151–66. https://doi.org/10.1109/TAI.2021.3054609.

[54] Omar R, Muccini H. Greening the AI: Evaluating tokenization methods for Energy-efficient NLP. 2025 11th International Conference on ICT for Sustainability (ICT4S), 2025, p. 133–42. https://doi.org/10.1109/ICT4S68164.2025.00022.

[55] Santos FMP, Gonçalves A, Sousa JMC, Vieira SM. Distilling Knowledge from Deep Neural Networks to Neuro-Fuzzy Inference Systems. 2025 IEEE International Conference on Fuzzy Systems (FUZZ), 2025, p. 1–6. https://doi.org/10.1109/FUZZ62266.2025.11152050.

[56] Bepler T, Berger B. Learning the protein language: Evolution, structure, and function. Cels 2021;12:654-669.e3. https://doi.org/10.1016/j.cels.2021.05.017.

[57] Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 2023;379:1123–30. https://doi.org/10.1126/science.ade2574.

[58] Bansal H, Luthra H, Raghuram SR. A Review on Machine Learning Aided Multi-omics Data Integration Techniques for Healthcare. In: Rivera G, Cruz-Reyes L, Dorronsoro B, Rosete A, editors. Data Analytics and Computational Intelligence: Novel Models, Algorithms and Applications, Cham: Springer Nature Switzerland; 2023, p. 211–39. https://doi.org/10.1007/978-3-031-38325-0_10.

[59] Teoh JR, Dong J, Zuo X, Lai KW, Hasikin K, Wu X. Advancing healthcare through multimodal data fusion: a comprehensive review of techniques and applications. PeerJ Comput Sci 2024;10:e2298. https://doi.org/10.7717/peerj-cs.2298.

[60] Schölkopf B, Locatello F, Bauer S, Ke NR, Kalchbrenner N, Goyal A, et al. Toward Causal Representation Learning. Proceedings of the IEEE 2021;109:612–34. https://doi.org/10.1109/JPROC.2021.3058954.

[61] Kumar D. Integrated Genomics and Multi-OMICS in Precision and Personalised Medicine: Emerging New Paradigm for Clinical Medicine. Apollo Medicine 2025:09760016251324360. https://doi.org/10.1177/09760016251324360.

[62] Wittmann BJ, Johnston KE, Wu Z, Arnold FH. Advances in machine learning for directed evolution. Current Opinion in Structural Biology 2021;69:11–8. https://doi.org/10.1016/j.sbi.2021.01.008.

[63] Tiukova IA, Brunnsåker D, Bjurström EY, Gower AH, Kronström F, Reder GK, et al. Genesis: Towards the Automation of Systems Biology Research 2024. https://doi.org/10.48550/arXiv.2408.10689.

[64] Kumar D. Applications of Quantum Chemistry in Molecular Spectroscopy and Reactivity. Journal of

Pharmaceutical Research and Integrated Medical Sciences 2025;2:31–48.
[65] Winker T, Groppe S, Uotila V, Yan Z, Lu J, Franz M, et al. Quantum Machine Learning: Foundation, New Techniques, and Opportunities for Database Research. Companion of the 2023 International Conference on Management of Data, New York, NY, USA: Association for Computing Machinery; 2023, p. 45–52. https://doi.org/10.1145/3555041.3589404.

Statements & Declarations

Conflict of Interests

The authors declare to have no conflict of interest with the manuscript and its related sources.

Funding

No funding was received for any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing interest

The authors have no relevant financial or non-financial interests to disclose.

Ethics Approval

Not Applicable

Consent to Participate

Not Applicable

Consent for Publication

Not Applicable

Availability of Data and Material

All data generated or analyzed during this study are included in this article.

Code Availability

Not Applicable