Advanced Machine Learning Techniques In Bioinformatics

Sidharth Selvin, Aswin Mohan, Shahanas Naisam

Dated: 30th June 2023
Keywords: *Machine Learning, Regression analysis, Neural networks, Deep learning.*

INTRODUCTION

Machine learning (ML) is a branch of artificial intelligence (AI) that focuses on developing algorithms and models that enable computers to learn and make predictions or decisions without being explicitly programmed. It is based on the idea that machines can learn from and analyze vast amounts of data to identify patterns, extract insights, and make informed decisions or predictions. The field has emerged as a transformative technology with significant implications across various industries. ML in bioinformatics contributed to genomics, proteomics, microarrays, evolution, systems biology, and text mining. Bioinformatics algorithms used to be manually programmed before the advent of machine learning, which possessed challenges for complex tasks like protein structure prediction. The introduction of machine learning techniques, such as deep learning, enables algorithms to autonomously learn features from datasets instead of relying on individual definitions by programmers. It aids in both algorithm and model development for biological data analysis. These algorithms can combine lower-level features into higher-level abstract features, thus enabling them to make complex predictions when properly trained. The underlying principle of bioinformatics-based ML is to acquire knowledge from data through pattern recognition, which can be subsequently employed across diverse applications

such as sequence alignment, gene expression analysis, protein structure prediction, and drug discovery. Hence, accurate predictions, faster data processing, and efficient data analysis on large datasets are possible for research and development (1, 2, 3).

ML has proven to be immensely advantageous in the field of Genomics and sequence analysis, Protein structure prediction, Disease diagnosis and prediction, Drug discovery and development, Personalized medicine, Biological network analysis, etc (4). The process of machine learning involves training large datasets in algorithms that automatically learn and train from experience. These algorithms use statistical techniques to identify patterns and relationships within the data, allowing the machine to make accurate predictions or decisions when processed with new or unknown data. With the availability of large biological datasets and great advancements in computational power, machine learning continues to revolutionize bioinformatics by enabling the analysis of biological data and providing insights into complex biological processes. It helps researchers and practitioners gain a deeper understanding of biological systems, accelerate discoveries, and develop more precise and personalized approaches to healthcare.

Machine learning techniques in Bioinformatics

- **Regression analysis**

Regression analysis is a statistical technique employed to uncover the connection between a dependent variable and one or multiple independent variables. The approach is used in bioinformatics to model and analyze the relationship between variables and make predictions(5,6). Linear regression, Cox regression, and logistic regression are three major types of regression analysis used in bioinformatics. The association between gene expression levels and different clinical parameters like disease progression, survival outcomes, and drug response can be identified via regression analysis. The application of regression analysis in developing a

diagnostic signature for ischemic stroke using miRNA, identifying immune gene markers, and exploring autophagy-related lncRNAs has been cited in various research articles(7,8).

- **Classification**

Classification, categorize objects based on their features into pre-defined classes, i.e. to analyze large datasets, understand features, and extract patterns or relationships that can aid in understanding biological systems or making predictions(9). Features are the measurable characteristics or properties of the data points used to distinguish between different classes and features can be quantitative like numerical values or qualitative like categorical variables. The presence or absence of a disease, protein function identification, and classification of tumor subtypes are some areas where classification is used in bioinformatics. K-nearest neighbors, support vector machines, and decision trees are the main classification algorithms in use(10). The selection of methods is based on the characteristics of data, the available features, and the specific research question.

Machine learning will always be an impact in the domain of bioinformatics as long as the availability of biological data increases and the advancement of computational power. Some approaches to machine learning in bioinformatics include supervised learning, unsupervised learning, semi-supervised learning, deep learning, and bayesian networks

A. Supervised learning

In supervised learning, a labeled dataset is used to train the algorithm where the anticipated result or response is already known in advance. Depending on the input features the algorithm receives, it tends to predict the outcome variable. The model learns the correlation between the input data and the output labels, enabling it to accurately predict the appropriate label for the novel, unseen data. Bioinformatics employs supervised learning in identifying disease

biomarkers, and specific functional classes, predicting protein structural properties, identifying sequence motifs associated with specific functions, and predicting protein function, etc(11). Supervised learning algorithms used in bioinformatics include Neural Networks, Random Forests, and Support Vector Machines.

- **Support vector machines (SVM)**

A simple type of ML technique is employed for regression tasks or classification analysis. This technique is effective for both linearly separable and non-linearly separable datasets. The fundamental concept of SVMs relies on discovering an ideal hyperplane that maximizes the gap between data points belonging to distinct classes(12). Classification of tumors based on gene expression profiles, protein fold, remote homology detection, identification of gene interactions, and prediction of protein function are some areas where support vector machines are used. They can handle high dimensional data and are robust to noise.

- **Random forest**

A versatile supervised ML algorithm intended for classification, regression, and feature selection. RF is used in bioinformatics as it can handle heterogeneous high-dimensional biological data. They are robust to noise and outliers in the data, making them suitable for handling noisy biological datasets(13). The method offers a way to assess the significance of features, aiding in the identification of the most pertinent characteristics within biological datasets which can aid in understanding the underlying biological mechanisms. Bioinformatics uses random forests mainly to classify tumors based on their gene expression profiles, predict drug response, and identify genes associated with the disease(14, 15). Reliable and accurate predictions are provided by random forests by considering multiple decision trees rather than relying on a single model. Missing data are being handled effectively by imputing missing values based on other available

features which is valuable in bioinformatics, where missing data can be prevalent and can significantly impact analysis results.

- **Neural networks**

As the name suggests, this class of machine learning algorithms was developed after the working and structure of the human brain. NNs comprise artificial neurons organized into interconnected layers that learn from data to make predictions and typically follow a feedforward architecture, where data flows to the output layer through hidden layers from the input layer(16). Multiple interconnected neurons constitute each layer and weighted edges represent the connections between neurons. Sigmoid, rectified linear unit (ReLU), and hyperbolic tangent (tanh) are activation functions NNs apply to their inputs, which introduce non-linearities, enabling the network to learn complex mappings(17). Neural networks are mainly used in gene expression analysis, drug discovery, and protein structure prediction. It consists of nodes that are interconnected to each other. They perform computations and these networks are trained using backpropagation for the optimization of parameters.

B. Unsupervised Learning

In supervised learning, a labeled dataset is used to train the algorithm where the anticipated result or response is already known in advance. In unsupervised learning, an unlabeled dataset is used to train the algorithm where the outcome or the response is not known. Similar data points are identified by the algorithm and grouped. The model explores the input data to discover inherent relationships or patterns on its own. In Bioinformatics, unsupervised learning is found to be useful in the field of clustering gene expression data or identifying protein domains. K-means clustering, Principal Component Analysis (PCA), and Hierarchical clustering are some of the common unsupervised learning algorithms used in bioinformatics(18).

- **Clustering**

Clustering, an unsupervised learning method, groups similar data points according to their features. Clustering excels in finding inherent patterns or structures within the data without prior knowledge or explicit guidance. It helps in discovering hidden relationships, identifying natural clusters, or detecting anomalies within a dataset(19). Clustering proteins based on their functional similarities, identifying subgroups of patients with similar gene expression profiles, clustering group genes based on their co-expression patterns and evolutionary analysis are some areas where clustering is employed. Self-organizing maps, hierarchical clustering, Density-Based Spatial Clustering, and k-mean clustering are common clustering algorithms(20).

- **Dimensionality reduction**

Dimensionality reduction reduces the number of variables or features in a dataset while retaining as much information as possible for analyzing and interpreting high-dimensional biological data.

This technique will be highly applicable to bioinformatics, particularly when dealing with large-scale genomic, transcriptomic, proteomic, or metabolomic datasets. Visualizing high-dimensional data, identifying genes that are differentially expressed between groups, and identifying biomarkers that are predictive of disease are some of the major areas where dimensionality reduction is used(21). The most used dimensionality reduction algorithms include PCA, Uniform Manifold Approximation and Projection, Non-negative Matrix Factorization, and t-distributed stochastic neighbor embedding(22).

C. Deep-learning

Deep learning, a subsidiary of ML, inspired by the functioning and structure of the human brain aims to enable computers to learn and make decisions in a more humane pattern. The method

automates the feature extraction (relevant features earlier identified by humans in traditional machine learning) process by using neural networks with multiple layers of interconnected nodes (artificial neurons or units) to learn complex relationships within the data(23). Bioinformatics finds the use of deep learning in predicting protein structure or identifying regulatory elements in the genome. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are two common algorithms of deep learning used in the area of bioinformatics(24).

- **Convolutional Neural Networks (CNN)**

Convolutional Neural Networks are particularly effective in capturing spatial relationships and local patterns in data, which is an apt method for analyzing biological sequences, images, and other structured data. Multiple layers are one of the main characteristics of CNN, comprising convolutional, pooling, and fully connected layers. The former layers use filters or kernels to perform convolution operations, extracting relevant features from the input data followed by downsampling the data (pooling layers) to reduce its dimensionality while preserving important information. Regression or classification tasks based on learning were performed by the fully connected layers. DNA sequence classification and prediction tasks, such as identifying coding regions, detecting gene promoters, and predicting RNA binding sites(25).

- **Recurrent Neural Networks (RNN)**

RNNs process sequential data by utilizing recurrent connections, enabling them to capture long-term dependencies and temporal dependencies in biological sequences. Recurrent units accept input at each time step, note maintains hidden states which allow RNNs to retain information from previous time steps and utilize it in the current prediction or classification task. Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM) are the popular variants of

RNNs. They can be employed in predicting sequence motifs, and secondary structures, identifying functional regions, and interacting between drugs and target proteins. RNNs are highly efficient for machine learning problems involving sequential data, as they can maintain internal memory(26).

D. Semi-supervised learning

This is a machine learning approach combining both supervised and unsupervised learning to improvise the performance of a model. This type is generally used when the labeled dataset is limited but the algorithm is supposed to learn from both labeled and unlabeled data. This approach focuses on leveraging the additional unlabeled data to improve the model's generalization and performance. Bioinformatics, a domain in which acquiring annotated data can be time-consuming and costly, this method can be particularly useful for applications like predicting protein interactions or identifying genes associated with disease(27). Self-Training and Co-Training are the common semi-supervised learning algorithms used in bioinformatics.

E. Bayesian networks

To represent the complex behavior and relationships between a set of variables, probabilistic graphs called Bayesian networks are used. The Bayesian network is considered a valuable technique to predict the effects of perturbations on the biological system and to provide a probabilistic framework for reasoning about uncertainty. This network comprises a directed acyclic graph along with a collection of conditional probability tables. The nodes in the graph represent random variables, and the directed edges between the nodes indicate the dependencies among the variables(28). Bayesian networks are involved in the modeling of gene regulatory networks, prediction of protein-protein interactions, and identification of genetic interactions.

Applications of Machine Learning in Bioinformatics

- **Genomic Sequencing and Variant Analysis**

Machine learning algorithms have greatly improved the accuracy and efficiency of genomic sequencing and variant analysis. These algorithms can detect genetic variations, such as insertions, deletions, single nucleotide polymorphisms (SNPs), and structural variations, aiding in the understanding of their functional implications. Moreover, machine learning models can predict the impact of genetic variants on protein structure and function, guiding personalized medicine and drug discovery efforts. The models can efficiently process the massive volume of genomic data and extract meaningful insights. In preprocessing steps, like read alignment, base calling, and quality control ML models/ algorithms can play a pivotal role in correcting errors, reducing noise, and improving the accuracy of downstream analyses (29). CNNs and RNNs are effective in capturing intricate patterns and relationships within genomic sequences, enabling the identification of genetic variants that may be missed by traditional methods.

- **Protein Structure Prediction and Folding**

Machine learning techniques have been instrumental in predicting protein structures and understanding their folding patterns. By leveraging large datasets of known protein structures, machine learning algorithms can infer the protein 3-dimensional structure from its primary structure (sequence). Deep learning techniques like CNNs, RNNs, and GNNs have shown remarkable movement in the same. CNN excels at capturing local sequence-structure patterns, enabling the identification of patterns and structural motifs significant for protein folding. RNNs are effective in modeling sequential dependencies of protein data, which is crucial for capturing secondary structure elements and long-range interactions. GNN handles protein structures as graphs, utilizing their connectivity information to make accurate predictions. This information is crucial for drug design, as it aids in identifying potential drug targets and predicting protein-ligand

interactions. Hybrid models leverage the strengths of different approaches to improve the accuracy of protein structure prediction and incorporate molecular dynamics simulations to refine predicted structures(30).

- **Functional Annotation and Prediction**

Machine learning algorithms have facilitated the functional annotation of genes and proteins. By analyzing large-scale biological data, such as gene expression profiles, protein-protein interaction networks, and metabolic pathways, these algorithms can predict the functions of uncharacterized genes or proteins. Valuable insights into functional properties are provided by gene expression data. Gene regulatory networks and PPI networks are biological networks that capture the relationships among biomolecules. Graph-based algorithms and network propagation utilize the associated attributes and network topology to infer functional annotations. Gene co-expression networks, classification models, and clustering can extract patterns from gene expression profiles and predict functions based on co-expression patterns. This knowledge helps unravel complex biological processes and identify potential biomarkers for diseases. Motif discovery, sequence similarity, and profile hidden Markov models (HMMs) to annotate functions are leveraged by sequence-based approaches. The accuracy of algorithms is enhanced by learning from annotated sequences and predicting functions for unannotated ones (31).

- **Drug Discovery and Repurposing**

Machine learning has accelerated the process of drug discovery and repurposing by enabling the rapid screening of large compound libraries to identify potential drug candidates. ML models for ligand-based virtual screening predict compound activity and similarity to known active molecules while structure-based virtual screening employs algorithms to predict the interaction between small molecules and target proteins with their binding affinity. By analyzing diverse chemical and biological data, including molecular structures, protein-ligand interactions, and

pharmacological profiles, machine-learning models can identify novel drug candidates with higher precision and efficiency (32). By analyzing large-scale omics data, these models can suggest existing drugs that could be repurposed for new therapeutic indications, hence saving time and resources in the drug development pipeline.

- **Disease Diagnosis and Prognosis**

Machine learning techniques have shown promise in disease diagnosis and prognosis. By integrating clinical data, genetic information, and imaging data, these algorithms can assist in early detection, accurate diagnosis, and prognosis of various diseases. Machine learning models can identify patterns and biomarkers that are indicative of specific conditions, helping healthcare professionals make informed decisions and improve patient outcomes. Analyzing proteomic, genomic, and clinical datasets, ML algorithms can identify patterns or molecular signatures associated with specific diseases which in turn enables the development of diagnostic tests based on biomarker panels for efficient disease detection. Integration of diverse datasets, such as imaging data, clinical records, and genomic data, ML models can identify disease subtypes, assisting us in tailored and effective treatment strategies (33). On the other hand, these models capture disease dynamics and individual patient variability from the above-mentioned datasets, to provide prognostic predictions, and personalized treatment selection, reducing the risk of adverse effects and optimizing therapeutic outcomes.

Challenges and Future Directions

- **Data availability and quality**

Large and complex biological datasets, including genomics, transcriptomics, and proteomics, are the characteristic features of bioinformatics. Accessing well-annotated and comprehensive

datasets can be challenging due to limited data availability, privacy concerns, and data-sharing policies for specific biological contexts.

- **Interpretability and explainability of machine learning models**

Even though ML models have demonstrated impressive predictive performance in various bioinformatics tasks, they often function as black boxes, making it challenging to understand how and why they make certain predictions. For example, in the context of protein structure prediction, interpretability becomes crucial as it helps researchers gain insights into the underlying principles and mechanisms governing protein folding.

Future research efforts need to focus on developing interpretable machine learning models in bioinformatics which involves designing algorithms that provide transparent and understandable explanations for their predictions. Techniques such as rule extraction, model visualization, and feature importance analysis can aid in unraveling the decision-making process of complex ML models. In addition, methods like saliency maps and attention mechanisms can highlight important features or regions in protein structures or sequences that contribute to the model's predictions. In the upcoming times collaborations between domain specialists, machine learning experts, and bioinformatics researchers, should be fostered to develop interpretable and explainable machine learning models specific to bioinformatics applications.

- **Ethical Considerations in Bioinformatics and Machine Learning**

As the increasing amount of biological data is being generated and analyzed, protecting patient privacy and ensuring data security becomes paramount. ML algorithms rely on training data, which can introduce bias if not carefully curated. Biases in data, such as the underrepresentation of certain populations or diseases, can lead to biased predictions or unequal access to

healthcare interventions. ML models operate as black boxes, which makes their predictions vulnerable in critical applications such as disease diagnosis or drug discovery.

As future directions, establishing clear ethical guidelines and best practices specific to machine learning in bioinformatics can provide researchers, practitioners, and policymakers with a framework for responsible conduct. interpretable machine learning models and techniques in bioinformatics should be developed. Emphasizing robust data governance practices, including anonymization techniques, differential privacy, and secure data-sharing protocols, can enhance privacy protection while facilitating data sharing for research purposes.

Conclusion

Machine learning has emerged as a transformative force in the field of bioinformatics and modern science. Researchers have harnessed the power of data processing by altering/implementing existing techniques through various machine learning. to unlock novel possibilities for scientific discovery. By leveraging a wide range of machine learning techniques, such as deep learning, ensemble methods, and feature selection, bioinformatics has witnessed remarkable progress in areas such as genomics, proteomics, and drug discovery. These advancements have not only accelerated the pace of research but have also paved the way for personalized medicine, precision agriculture, and improved understanding of disease mechanisms. As we continue to refine and expand machine learning methods, the potential for further advancements in bioinformatics remains immense, promising a future where data-driven approaches will continue to revolutionize our understanding of the biological world and shape the development of novel solutions for a healthier and more sustainable future.

Reference

1. Diaz-Flores, E., Meyer, T., & Giorkallos, A. (2022). Evolution of Artificial Intelligence-Powered Technologies in Biomedical Research and Healthcare. 23–60. https://doi.org/10.1007/10_2021_189

2. Vemula, D., Jayasurya, P., Sushmitha, V., Kumar, Y. N., & Bhandari, V. (2022). CADD, AI, and ML in Drug Discovery: A Comprehensive Review. European Journal of Pharmaceutical Sciences, 106324. https://doi.org/10.1016/j.ejps.2022.106324

3. Lu, M., Yin, J., Zhu, Q., Lin, G., Mou, M., Liu, F., Pan, Z., You, N., Lian, X., Li, F., Zhang, H., Zheng, L., Zhang, W., Zhang, H., Shen, Z., Gu, Z., Li, H., & Zhu, F. (2023). Artificial Intelligence in Pharmaceutical Sciences. Engineering. https://doi.org/10.1016/j.eng.2023.01.014

4. Wei, Z., Qi, X., Chen, Y., Xia, X., Zheng, B., Sun, X., Zhang, G., Wang, L., Zhang, Q., Xu, C., Jiang, S., Li, X., Xie, B., Liao, X., & Zhu, A. (2020). Bioinformatics method combined with logistic regression analysis reveal potentially important miRNAs in ischemic stroke. Bioscience Reports, 40(8). https://doi.org/10.1042/bsr20201154.

5. Chen, X., Jin, Y., Gong, L., He, D., Cheng, Y., Xiao, M., Zhu, Y., Wang, Z., & Cao, K. (2020). Bioinformatics Analysis Finds Immune Gene Markers Related to the Prognosis of Bladder Cancer. 11. https://doi.org/10.3389/fgene.2020.00607.

6. Li, J., Zhang, H., & Gao, F. (2022). Identification of miRNA biomarkers for breast cancer by combining ensemble regularized multinomial logistic regression and Cox regression. BMC Bioinformatics, 23(1). https://doi.org/10.1186/s12859-022-04982-7

7. He, L., Fan, Y., Zhang, Y., Tu, T., Zhang, Q., Yuan, F., & Cheng, C. (2022). Single-cell transcriptomic analysis reveals circadian rhythm disruption associated with poor prognosis and drug-resistance in lung adenocarcinoma. 73(1). https://doi.org/10.1111/jpi.12803.
8. Stevens, R., Goble, C., Baker, P., & Brass, A. (2001). A classification of tasks in bioinformatics. Bioinformatics, 17(2), 180–188. https://doi.org/10.1093/bioinformatics/17.2.180.
9. Ebru Simsek, Hasan Badem, & Ibrahim Taner Okumus. (2021). Leukemia Sub-Type Classification by Using Machine Learning Techniques on Gene Expression. 629–637. https://doi.org/10.1007/978-981-16-2102-4_56

Kumar, I., Surya Prakash Singh, & None SHIVAM. (2022). Machine learning in bioinformatics. 443–456. https://doi.org/10.1016/b978-0-323-89775-4.00020-1

Li, Y., Meng, K., Yang, G., Liu, B., Li, C., Zhang, J.-Y., & Zhang, X.-M. (2022). Diagnostic genes and immune infiltration analysis of colorectal cancer determined by LASSO and SVM machine learning methods: a bioinformatics analysis. 13(3), 1188–1203. https://doi.org/10.21037/jgo-22-536

Qi, Y. (2012). Random Forest for Bioinformatics. Ensemble Machine Learning, 307–323. https://doi.org/10.1007/978-1-4419-9326-7_11

Zhang, H., Chi, M., Su, D., Xiong, Y., Wei, H., Yu, Y., Zuo, Y., & Yang, L. (2023). A random forest-based metabolic risk model to assess the prognosis and metabolism-related drug targets in ovarian cancer. 153, 106432–106432. https://doi.org/10.1016/j.compbiomed.2022.106432

Zhao, S., Chi, H., Ji, W., He, Q., Lai, G., Peng, G., Zhao, X., & Cheng, C. (2022). A Bioinformatics-Based Analysis of an Anoikis-Related Gene Signature Predicts the Prognosis of Patients with Low-Grade Gliomas. 12(10), 1349–1349. https://doi.org/10.3390/brainsci12101349

Yang, X., Ye, J., & Wang, X. (2022). Factorizing Knowledge in Neural Networks. 73–91. https://doi.org/10.1007/978-3-031-19830-4_5

Goyal, M., Goyal, R., Venkatappa Reddy, P., & Lall, B. (2019). Activation Functions. Deep Learning: Algorithms and Applications, 1–30. https://doi.org/10.1007/978-3-030-31760-7_1

Parasa, N. A., Namgiri, J. V., Mohanty, S. N., & Dash, J. K. (2021). Introduction to Unsupervised Learning in Bioinformatics. Data Analytics in Bioinformatics, 35–49. https://doi.org/10.1002/9781119785620.ch2

Teng, H., Yuan, Y., & Ziv Bar-Joseph. (2021). Clustering spatial transcriptomics data. 38(4), 997–1004. https://doi.org/10.1093/bioinformatics/btab704

Ciortan, M., & Defrance, M. (2021). GNN-based embedding for clustering scRNA-seq data. Bioinformatics, 38(4), 1037–1044. https://doi.org/10.1093/bioinformatics/btab787

He, S., Dou, L., Li, X., & Zhang, Y. (2022). Review of bioinformatics in Azheimer's Disease Research. Computers in Biology and Medicine, 143, 105269. https://doi.org/10.1016/j.compbiomed.2022.105269

Nikolay Oskolkov. (2022). Dimensionality Reduction. 151–167. https://doi.org/10.1007/978-3-030-88389-8_9

Routhier, E., & Mozziconacci, J. (2022). Genomics enters the deep learning era. PeerJ, 10, e13613. https://doi.org/10.7717/peerj.13613

Wang, W., & Gao, X. (2019). Deep learning in bioinformatics. Methods, 166, 1–3. https://doi.org/10.1016/j.ymeth.2019.06.006

Dimitrios Amanatidis, Konstantina Vaitsi, & Dossis, M. (2022). Deep Neural Network Applications for Bioinformatics. https://doi.org/10.1109/seeda-cecnsm57760.2022.9932895

OUP accepted manuscript. (2021). Briefings in Bioinformatics. https://doi.org/10.1093/bib/bbab533

Li, F., Dong, S., Leier, A., Han, M., Guo, X., Xu, J., Wang, X., Pan, S., Jia, C., Zhang, Y., Webb, G. I., Coin, M., Li, C., & Song, J. (2021). Positive-unlabeled learning in bioinformatics and computational biology: a brief review. 23(1). https://doi.org/10.1093/bib/bbab461

Suter, P., Kuipers, J., & Niko Beerenwinkel. (2022). Discovering gene regulatory networks of multiple phenotypic groups using dynamic Bayesian networks. 23(4). https://doi.org/10.1093/bib/bbac219

Alharbi, W. S., & Rashid, M. (2022). A review of deep learning applications in human genomics using next-generation sequencing data. Human Genomics, 16(1). https://doi.org/10.1186/s40246-022-00396-x

Pearce, R., Li, Y., Omenn, G. S., & Zhang, Y. (2022). Fast and accurate Ab Initio Protein structure prediction using deep learning potentials. PLoS Computational Biology, 18(9), e1010539. https://doi.org/10.1371/journal.pcbi.1010539

Enzyme function prediction using contrastive learning

Pan, X., Lin, X., Cao, D., Zeng, X., Yu, P. S., He, L., Nussinov, R., & Cheng, F. (2022). Deep learning for drug repurposing: Methods, databases, and applications. WIREs Computational Molecular Science. https://doi.org/10.1002/wcms.1597

Ghazal, T. M., Rehman, A. U., Saleem, M., Ahmad, M., Ahmad, S., & Mehmood, F. (2022, February 1). Intelligent Model to Predict Early Liver Disease using Machine Learning Technique. IEEE Xplore. https://doi.org/10.1109/ICBATS54253.2022.9758929