https://doi.org/10.17048/AM.2023.134

Noha E. El-Attar, Dr. Yehia. A. El-Mashad: Artificial intelligence models for genomics analysis: review article

Noha E. El-Attar, Dr. Yehia. A. El-Mashad
Faculty of Computers and AI, Benha University
Faculty of Engineering, Delta University for Science and Technology

Abstract: Artificial intelligence (AI) including machine learning (ML), and deep learning (DL) models have become powerful tools for analyzing genomics data in recent years. These models can process large amounts of data and identify complex patterns that may not be apparent through traditional statistical methods. ML and DL models have been used for a wide range of genomics applications, including gene expression analysis, variant detection, and drug discovery.

One popular approach for using ML and DL models in genomics is to train these models on large datasets of genomic information. These datasets may include information on gene expression levels, DNA sequences, and epigenetic modifications. By training these models on large datasets, researchers can identify patterns and correlations that may be used to predict disease risk, identify potential drug targets, and develop personalized treatments.

Generally, the use of different AI models in genomics has the potential to transform the field by enabling more accurate and personalized medical treatments. As these models continue to evolve and improve, researchers will be able to extract even more information from genomic data and accelerate the pace of discovery in genomics.

1. Introduction

Artificial intelligence (AI) is a rapidly growing area of research in bioinformatics, which is the application of computational methods to biological data. AI techniques can be used to analyze and interpret large and complex biological datasets, such as genomic and proteomic data, to gain insights into biological processes and disease mechanisms [1].

One common application of AI in bioinformatics is in the field of protein structure prediction. There are many AI-based methods for predicting protein structure, including deep learning approaches such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs). These methods use large amounts of protein sequence and structural data to train models that can predict the 3D structure of a protein from its amino acid sequence [2].

Another area where AI is being used in bioinformatics is in drug discovery. AI techniques can be used to predict the interactions between molecules, which can help identify potential drug candidates. For example, machine learning algorithms can be trained on large databases of known drug-protein interactions to predict the activity of new compounds against specific targets [3].

AI is also being used to analyze large-scale genomic data, such as gene expression data and DNA sequence data. For example, AI techniques can be used to identify patterns in gene expression data that are associated with specific diseases or physiological conditions. This can help researchers identify new drug targets or biomarkers for disease diagnosis and treatment [4][5].

Here are some additional examples of how AI is being used in genomics:

1. Genome assembly: AI can be used to help assemble genomes from DNA sequencing data. Genome assembly involves piecing together short fragments of DNA into longer sequences to create a complete genome. This can be a computationally intensive process, but AI techniques such as deep learning can help speed up the process and improve accuracy [6].

2. Gene editing: AI can be used to help design and optimize gene editing tools such as CRISPR, which are used to make precise changes to the DNA sequence of organisms. By analyzing large-scale genomic datasets, AI can help identify the most effective targets for gene editing and optimize the design of the gene editing tools themselves [7].
3. Personalized medicine: AI can be used to help develop personalized medicine approaches based on an individual's genetic profile. By analyzing an individual's genetic data, AI can identify genetic variations that may be associated with an increased risk of certain diseases or a poor response to certain drugs. This information can be used to develop personalized treatment plans that are tailored to an individual's unique genetic makeup [6].
4. Drug discovery: AI can be used to help identify novel drug targets and design new drugs. By analyzing large-scale genomic datasets, AI can help identify genetic mutations that are associated with specific diseases and develop new drugs that target these mutations.
5. Genome annotation: AI can be used to help annotate genomes, which involves identifying the location and function of genes within the genome. By analyzing large-scale genomic datasets, AI can help identify new genes and predict their function, which can provide insights into the genetic basis of diseases [7].
6. Cancer genomics: AI can be used to analyze genomic data from cancer cells to identify genetic mutations that are driving the growth of tumors. By identifying these mutations, researchers can develop targeted therapies that are tailored to the specific genetic makeup of the tumor [4].
7. Metagenomics: AI can be used to analyze the genetic material from entire microbial communities, such as those found in the human gut. By analyzing this data, researchers can gain insights into the microbiome and its role in human health and disease [15].
8. Epigenetics: AI can be used to analyze epigenetic data, which involves changes to the DNA molecule that do not alter the underlying genetic sequence. By analyzing epigenetic data, researchers can gain insights into how genes are regulated and how changes in gene expression can contribute to disease [3].
9. Functional genomics: AI can be used to analyze the function of genes within the context of the entire genome. By analyzing large-scale genomic datasets, researchers can identify the interactions between genes and how they contribute to biological processes [14].
10. Clinical genomics: AI can be used to analyze genomic data from patients to help diagnose genetic diseases and develop personalized treatment plans. By analyzing an individual's genetic data, AI can help identify genetic mutations that are associated with specific diseases and develop treatment strategies that are tailored to the individual's unique genetic makeup [5].
11. Single-cell genomics: AI can be used to analyze the genetic material from individual cells, allowing researchers to study cellular diversity and identify rare cell types. By analyzing single-cell genomic data, researchers can gain insights into how individual cells contribute to biological processes and disease [6].
12. Multi-omics integration: AI can be used to integrate data from multiple "omics" technologies, such as genomics, proteomics, and metabolomics. By combining data from these different technologies, researchers can gain a more comprehensive understanding of biological processes and diseases.
13. Evolutionary genomics: AI can be used to analyze the evolution of genomes over time, helping researchers understand how genetic variation contributes to species diversity and adaptation [8].
14. Synthetic biology: AI can be used to design and optimize synthetic biological systems, such as engineered cells or organisms. By leveraging AI techniques, researchers can design biological systems that are more efficient, robust, and effective [9].
15. Data sharing and collaboration: AI can be used to facilitate data sharing and collaboration within the genomics community. By developing tools that can analyze and integrate data from multiple sources,

AI can help researchers work together more effectively and accelerate the pace of scientific discovery [10].

16. Genomic data privacy: AI can be used to protect the privacy of genomic data by developing tools that can analyze genomic data without revealing sensitive information about individuals. By using AI to develop privacy-preserving data analysis methods, researchers can ensure that genomic data remains secure and confidential [16].
17. Quality control: AI can be used to identify errors and inconsistencies in genomic data, helping to ensure that data is accurate and reliable. By developing quality control methods that leverage AI techniques, researchers can improve the quality of genomic data and reduce the risk of false findings.
18. Natural language processing: AI can be used to analyze scientific literature and extract information about genes, proteins, and other biological entities. By developing natural language processing tools that can analyze large volumes of scientific literature, researchers can gain insights into the function and regulation of genes and proteins [15].
19. Genomic medicine: AI can be used to develop new diagnostic and therapeutic approaches based on an individual's genomic data. By analyzing an individual's genetic makeup, AI can help identify the underlying causes of disease and develop personalized treatment plans that are tailored to the individual's unique genetic profile [11].
20. Education and outreach: AI can be used to develop educational resources and outreach programs that help to promote genomics literacy and engage the public in scientific research. By leveraging AI to develop interactive learning tools and engaging outreach programs, researchers can help to bridge the gap between the scientific community and the general public [12].

Overall, the use of AI in bioinformatics is a rapidly evolving field that holds great promise for advancing our understanding of biological systems and developing new treatments for diseases.

2. A Primer on Genomics Data and AI Applications

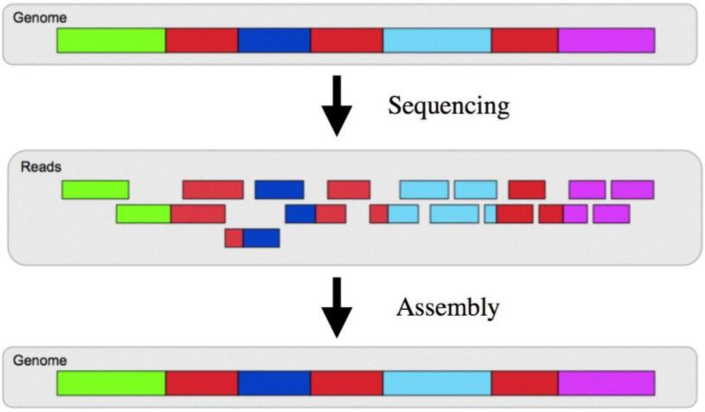
2.1 Genome Assembly

Genome assembly is the process of reconstructing the complete DNA sequence of an organism from a large number of short DNA fragments. This process is necessary because the DNA sequence of most organisms is too large to be sequenced in a single read. Instead, the genome is broken up into many small fragments, which are sequenced separately using high-throughput sequencing technologies.

The process of genome assembly involves several steps, including sequencing, quality control, read trimming, and genome assembly algorithms. The first step in genome assembly is to generate a large number of DNA sequences, known as reads, using high-throughput sequencing technologies such as Illumina or PacBio. These reads are then subjected to quality control to remove any low-quality reads or contaminants.

The next step is to trim the reads to remove any regions with poor sequencing quality or adapter sequences. This is followed by genome assembly algorithms, which use computational methods to align the reads and assemble them into contigs, which are longer contiguous stretches of DNA sequence. The contigs are then further assembled into larger scaffolds, which are ordered and oriented to produce the final genome assembly. This process is summarized in Figure (1).

The quality of the genome assembly depends on the quality and quantity of the reads, the accuracy of the genome assembly algorithms, and the size and complexity of the genome being assembled. Genome assembly is an important step in genomics research, as it provides a complete sequence of an organism's DNA and can help identify genetic variations that are associated with diseases or other phenotypic traits [15][16].



AI roles in genome assembly:

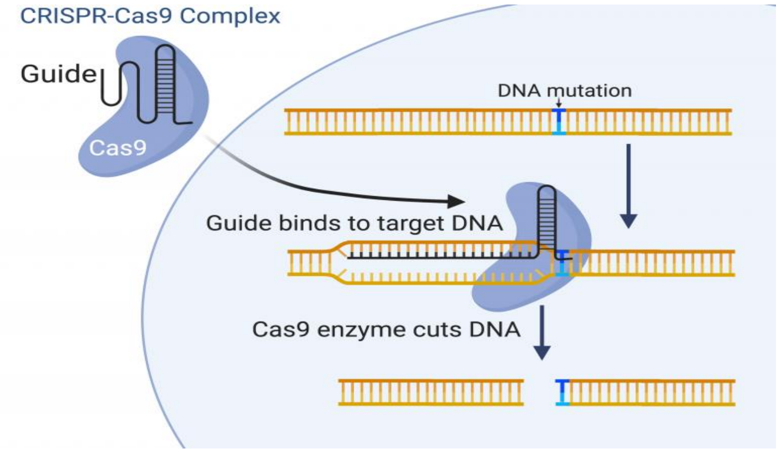
Overall, the use of AI in genome assembly can help improve the accuracy and efficiency of the process, allowing researchers to assemble more complete and accurate genomes. This can lead to new insights into the genetic basis of diseases and the development of personalized medicine. AI can help in genome assembly in several ways:

1. Error correction: One of the main challenges in genome assembly is correcting errors that can arise during the sequencing process. AI can be used to identify and correct these errors in sequencing data, which can improve the accuracy of the final genome assembly.
2. De novo assembly: AI can be used to assist with de novo assembly, which involves assembling a genome from scratch without the use of a reference genome. AI techniques such as deep learning can help speed up the process and improve the accuracy of de novo assembly.
3. Hybrid assembly: AI can be used to assist with hybrid assembly, which involves combining data from multiple sequencing technologies to assemble a genome. By analyzing large-scale genomic datasets, AI can help identify the most effective combination of sequencing technologies for a given genome assembly project.
4. Genome annotation: AI can be used to assist with genome annotation, which involves identifying the location and function of genes within the genome. By analyzing large-scale genomic datasets, AI can help identify new genes and predict their function, which can provide insights into the genetic basis of diseases [17][18].

2.2 Gene editing

Gene editing is a process by which DNA sequences can be precisely modified, added, or removed from the genome of an organism. This is accomplished by using molecular tools that can target specific DNA sequences and make precise cuts in the DNA, which can then be repaired by the cell's natural DNA repair mechanisms.

One of the most commonly used gene editing tools is CRISPR-Cas9, which is a system that can be programmed to target specific DNA sequences using guide RNAs. The Cas9 enzyme then cuts the DNA at the targeted site, which can be repaired by the cell's natural DNA repair mechanisms as shown in Figure (2). This process can be used to modify specific genes or regulatory regions of the genome, which can have a variety of applications in research and medicine [19].



Gene editing can be used to study the function of specific genes by creating mutations that disrupt their function. This can help researchers understand the role of specific genes in biological processes and disease.

AI roles in Gene Editing:

The use of AI in gene editing is an exciting area of research that holds great promise for improving the efficiency, accuracy, and safety of gene editing techniques. AI can play several roles in gene editing, including:

1. Designing guide RNAs: One of the key steps in gene editing is designing guide RNAs that target specific DNA sequences. AI can be used to design more efficient and specific guide RNAs by analyzing large-scale genomic datasets to identify optimal target sites.

2. Predicting off-target effects: One of the potential risks of gene editing is off-target effects, where the Cas9 enzyme cuts DNA at unintended locations. AI can be used to predict the likelihood of off-target effects by analyzing large-scale genomic datasets and simulating the effects of Cas9 cuts on the genome.

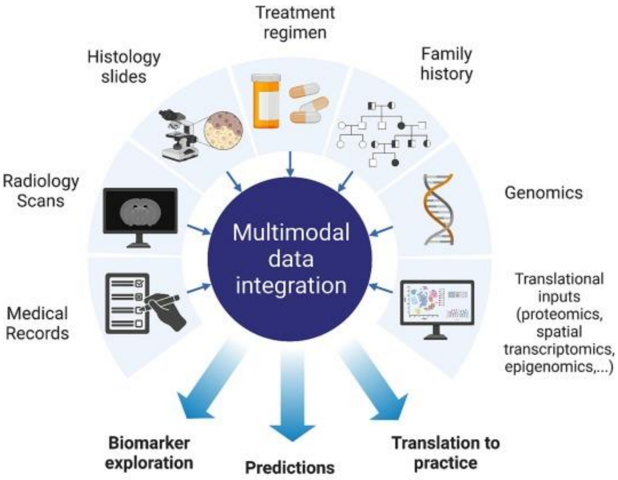
3. Optimizing delivery methods: Gene editing tools such as CRISPR-Cas9 need to be delivered to the target cells in order to be effective. AI can be used to optimize delivery methods by analyzing large-scale genomic datasets to identify the most effective delivery methods for specific cell types.

4. Identifying potential disease targets: AI can be used to analyze large-scale genomic datasets to identify potential disease targets for gene editing. By identifying genetic mutations that are associated with specific diseases, AI can help identify genes or regulatory regions that may be suitable targets for gene editing.

5. Developing new gene editing tools: AI can be used to develop new gene editing tools by simulating the effects of different molecular structures and testing their efficacy in silico. This can help researchers design more efficient and specific gene editing tools [20][21].

2.3 Personalized Medicine

Personalized medicine is an approach to medical treatment that takes into account an individual's unique genetic, environmental, and lifestyle factors to develop more targeted and effective treatment plans. Rather than relying on a "one-size-fits-all" approach to medical treatment, personalized medicine seeks to tailor treatments to the individual needs of each patient, this scenario is called multi model data integration as shown in Figure (3) [22].



One of the key drivers of personalized medicine is advances in genomics research, which has led to a better understanding of the genetic basis of many diseases. By analyzing an individual's genetic information, clinicians can identify genetic variations that are associated with disease risk or treatment response. This information can be used to develop more personalized treatment plans that take into account an individual's genetic makeup [23].

Other factors that can be taken into account in personalized medicine include an individual's environment, lifestyle, and medical history. For example, a personalized medicine approach to cancer treatment may involve analyzing a patient's tumor DNA to identify mutations that are driving the growth of cancer, as well as taking into account factors such as the patient's age, overall health, and treatment preferences [24].

AI roles in Personalized Medicine:

AI is playing an increasingly important role in personalized medicine, an approach to medical treatment that considers an individual's unique genetic, environmental, and lifestyle factors to develop more targeted and effective treatment plans.

AI can play several essential roles in personalized medicine, including:

1. Prediction and diagnosis: AI algorithms can be used to analyze large-scale datasets of patient data, including genomics, medical imaging, and electronic health records, to predict disease risk and diagnose diseases at an earlier stage. By analyzing patterns in the data, AI can help identify patients who are most at risk of developing certain diseases or who have undiagnosed conditions.

2. Treatment planning and decision-making: Once a diagnosis has been made, AI can be used to develop personalized treatment plans that consider an individual's unique characteristics. By analyzing large-scale datasets of patient data, AI can help identify the most effective treatments for specific patient populations based on factors such as genetic makeup, medical history, and lifestyle.

3. Drug discovery and development: AI can be used to accelerate the drug discovery and development process by analyzing large-scale datasets of biological and chemical data to identify new drug targets and predict the efficacy of new drug candidates. This can help reduce the time and cost required to develop new treatments [25].

4. Precision drug delivery: AI can be used to optimize drug delivery by analyzing patient data to identify the most effective dosages and delivery methods for specific patient populations. This can help improve treatment outcomes and reduce side effects.

5. Monitoring and follow-up: AI can be used to monitor patient outcomes and adjust treatment plans in real-time based on changes in patient data. This can help ensure that patients receive the most effective treatments, and can also help reduce healthcare costs by avoiding unnecessary interventions [26][27].

2.4 Drug discovery

Drug discovery is the process of identifying new compounds or molecules that can potentially treat or cure diseases. The goal of drug discovery is to identify compounds that can selectively target disease-causing molecules or pathways while minimizing side effects and toxicity to healthy cells [28].

The drug discovery process typically involves several stages, including:

1. Target identification: The first step in drug discovery is to identify a specific molecular target that is involved in the disease process. This could be a protein, enzyme, or other molecule that is essential for the disease to develop or progress.

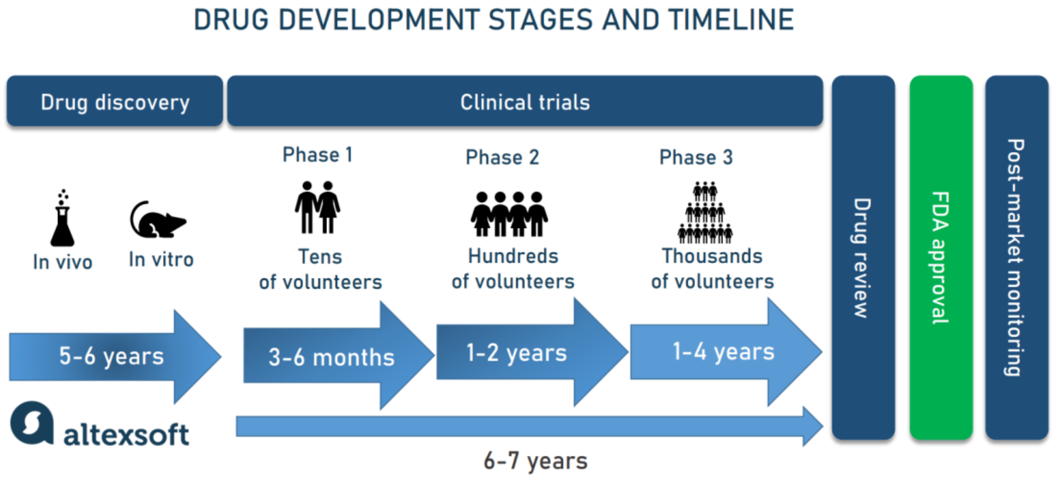
2. Lead discovery: Once a molecular target has been identified, the next step is to search for compounds or molecules that can interact with the target. This involves screening large libraries of compounds to identify those that have the potential to selectively bind to the target molecule.

3. Lead optimization: After identifying lead compounds that have the potential to interact with the target molecule, the next step is to optimize these compounds to improve their potency, selectivity, and safety profile. This involves chemical modifications to the lead compounds to improve their drug-like properties.

4. Preclinical testing: Once lead compounds have been optimized, they undergo preclinical testing to evaluate their efficacy and safety in animal models. This involves testing the compounds for their ability to treat the disease in animal models, as well as assessing their toxicity and potential side effects.

5. Clinical testing: If a lead compound shows promise in preclinical testing, it can move on to clinical testing in humans. Clinical testing involves several phases of trials to evaluate the safety and efficacy of the drug in humans.

6. Regulatory approval: Once a drug has successfully completed clinical testing and has been shown to be safe and effective, it can be submitted for regulatory approval. Regulatory agencies such as the FDA evaluate the safety and efficacy of the drug before approving it for use in the general population. Figure (4) displays the timeline and the stages for drug development, beginning with discovery through clinical trials and ending with post-market monitoring [29][30].

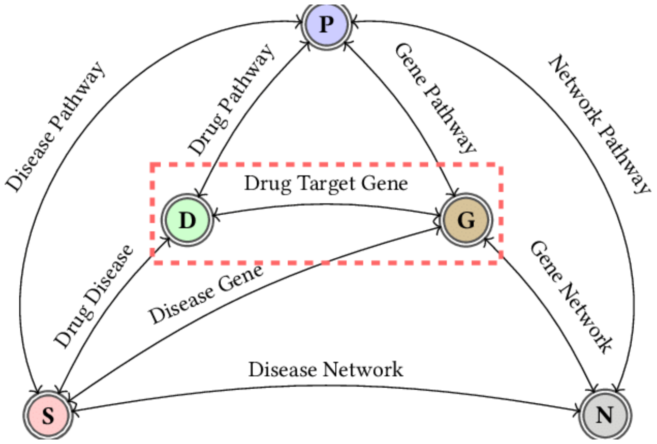


Overall, drug discovery is a complex and time-consuming process that can take many years and cost billions of dollars. However, it is essential to develop new treatments for a wide range of diseases, and it has the potential to improve patient outcomes and quality of life.

AI roles in Drug Discovery:

AI is playing an increasingly important role in drug discovery, which is the process of identifying new compounds or molecules that have the potential to treat or cure diseases. AI can be applied to various stages of the drug discovery process to help accelerate the development of new treatments and reduce the time and cost required to bring them to market. Here are some of the key roles of AI in drug discovery:

1. Target identification: AI algorithms can be used to analyze large-scale datasets of biological and chemical data to identify molecules or pathways that are involved in specific disease processes. This can help accelerate the drug discovery process by identifying new targets that may be suitable for drug development. Figure (5) presents an example of a biomedical knowledge graph.
2. Lead discovery and optimization: AI algorithms can be used to analyze large-scale datasets of chemical and biological data to identify compounds or molecules that have the potential to interact with a specific target. By analyzing the chemical structures of these compounds, AI can also help optimize them to improve their potency, selectivity, and safety profile.
3. Prediction of efficacy and safety: AI algorithms can be used to simulate the effects of drugs on biological systems, and to predict their efficacy and potential side effects. This can help identify promising drug candidates and reduce the time and cost required for preclinical and clinical testing.
4. Clinical trial optimization: AI algorithms can be used to analyze large-scale datasets of patient data to identify patient populations that are most likely to respond to a particular treatment, and to identify potential side effects or adverse events. This can help optimize trial design and reduce the time and cost required for clinical testing.
5. Drug repurposing: AI can be used to analyze existing drugs to identify new applications and potential new therapeutic uses. By analyzing large-scale datasets of biological and chemical data, AI can help identify drugs that may be effective for treating diseases that they were not originally developed for [31][32].



2.5 Genome annotation

Genome annotation is the process of identifying the functional elements and features within a genome sequence, such as genes, regulatory regions, and non-coding regions. The genome annotation process involves analyzing the DNA sequence to identify features such as open reading frames (ORFs), introns, exons, promoter regions, and regulatory regions [33].

The process of genome annotation typically involves several steps, including:

1. Gene prediction: This involves using computational algorithms to identify potential genes within the genome sequence. These algorithms may look for features such as ORFs, splice sites, and codon usage bias to identify potential coding regions.

2. Functional annotation: This involves assigning functions to the predicted genes based on similarity to known genes or functional domains. This can be done using databases such as Gene Ontology or by comparing the predicted proteins to known proteins in other organisms.

3. Regulatory annotation: This involves identifying regulatory regions within the genome sequence, such as promoter regions and enhancer regions. This can be done using computational algorithms that look for features such as transcription factor binding sites and histone modifications.

4. Structural annotation: This involves identifying non-coding regions within the genome sequence, such as introns and intergenic regions. This can be done using computational algorithms that look for features such as repetitive sequences and transposable elements [34][35].

The genome annotation process is an important step in understanding the function and organization of a genome, and can provide insights into the genetic basis of diseases and other biological phenomena. Genome annotation is an ongoing process, as new genomic data becomes available and new computational methods are developed.

AI roles in Genome annotation:

AI plays an important role in genome annotation, which is the process of identifying the functional elements and features within a genome sequence, such as genes, regulatory regions, and non-coding regions. AI can be applied to various stages of the genome annotation process to help accelerate the analysis of genomic data and improve the accuracy of annotations. Here are some of the key roles of AI in genome annotation [36][37]:

1. Gene prediction: AI algorithms can be used to analyze genomic data and identify potential coding regions, such as open reading frames (ORFs) and splice sites. Machine learning algorithms can be trained on large datasets of annotated genes to improve the accuracy of gene predictions.

2. Functional annotation: AI algorithms can be used to predict the functions of predicted genes based on similarity to known genes or functional domains. Machine learning algorithms can be trained on large datasets of annotated genes to improve the accuracy of functional annotations.

3. Regulatory annotation: AI algorithms can be used to identify regulatory regions within the genome sequence, such as promoter regions and enhancer regions. Machine learning algorithms can be trained on large datasets of known regulatory regions to improve the accuracy of regulatory annotations.

4. Structural annotation: AI algorithms can be used to identify non-coding regions within the genome sequence, such as introns and intergenic regions. Machine learning algorithms can be trained on large datasets of annotated genomic data to improve the accuracy of structural annotations.

2. Variant annotation: AI can be used to predict the functional impact of genetic variants, such as single nucleotide polymorphisms (SNPs), and identify variants that are associated with diseases or other traits.

3. Integration of multi-omics data: AI can be used to integrate data from different types of genomic experiments, such as transcriptomics, proteomics, and epigenomics, to provide a more comprehensive view of gene expression and regulation.

4. Quality control: AI algorithms can be used to identify and remove low-quality genomic data, such as sequencing errors or regions of low coverage, which can improve the accuracy of genome annotation.

5. Comparative genomics: AI can be used to compare the genomes of different species and identify conserved regions that are likely to be functional, such as regulatory elements or protein-coding genes.

2.6 Cancer Genomics

Cancer genomics is the field of study that focuses on the genomic alterations that drive the development and progression of cancer. Cancer is a genetic disease that arises from alterations in the DNA sequence of cells, and cancer genomics seeks to identify these alterations and understand their functional impact. One of the critical goals of cancer genomics is to identify the genomic alterations that are specific to different types of cancer, as well as the genomic alterations that are shared across different types of cancer. This can help identify potential drug targets and develop more personalized treatment options for cancer patients [38].

Cancer genomics also plays an important role in developing biomarkers, which are measurable indicators of disease that can be used to predict patient outcomes or response to treatment. By analyzing genomic data from cancer patients, researchers can identify genomic alterations associated with specific clinical outcomes, such as response to treatment or survival. Another critical area of cancer genomics is the study of cancer evolution, which refers to the process by which cancer cells acquire additional genomic alterations over time. By analyzing genomic data from different stages of cancer progression, researchers can identify the genomic alterations responsible for disease progression and develop more effective treatment strategies [39].

AI roles in Cancer Genomics:

AI is helping to accelerate cancer genomics research and improve cancer diagnosis, treatment, and prevention by analyzing large amounts of genomic and clinical data and identifying patterns and associations that would be difficult or impossible to identify using traditional methods. AI’s role in cancer genomics can be summarized in the following areas [40][41]:

1. Diagnosis and prognosis: AI algorithms can be trained to analyze patient data, such as genomic and clinical data, to predict the risk of developing cancer, the likelihood of cancer recurrence, and the response to different treatments.

2. Drug discovery: AI can be used to identify new drug targets and design more effective drugs based on genomic data. For example, AI can be used to identify genomic alterations that are specific to certain types of cancer and design drugs to target those alterations.

3. Genomic profiling: AI can be used to analyze large-scale genomic data from cancer patients to identify patterns and associations between genomic alterations and clinical outcomes. This can help identify biomarkers and personalized treatment options.

4. Image analysis: AI can be used to analyze medical images, such as CT scans and MRIs, to identify features that are characteristic of different types of cancer. This can help improve cancer diagnosis and treatment planning.

5. Clinical trial design: AI can identify patients most likely to benefit from a particular treatment and design clinical trials with more targeted patient populations [42].

3. Genomics Databases

Genomics data refers to the large amounts of data generated by analyzing genomic material, such as DNA or RNA. The specific name of genomics data can vary depending on the type of analysis being performed, but some common types of genomics data include [43]:

Whole genome sequencing data refers to the data generated by sequencing an individual's entire genome.

Transcriptome data refers to the data generated through the sequencing of an individual's RNA, which can provide information about gene expression and regulation.

Epigenomic data: This refers to the data generated through the analysis of modifications to DNA, such as DNA methylation or histone modifications, which can affect gene expression and regulation.

Metagenomic data refers to data generated through the sequencing of microbial communities, such as those found in the gut or soil.

Proteomic data refers to data generated through the analysis of an individual's proteins, which can provide information about protein function and interactions.

Genomics data represents a vast and complex set of information that requires specialized tools and expertise to analyze and interpret. Genomic databases are collections of genomic data that are organized and stored in a structured format for easy access and analysis. These databases are essential resources for researchers in genomics, bioinformatics, and related fields, as they provide a wealth of information about genes, genetic variations, and other genomic features. Some of the most commonly used genomic databases include GenBank, Ensembl, dbSNP, The Cancer Genome Atlas (TCGA), the Human Genome Variation Database (HGVD), and the Exome Aggregation Consortium (ExAC) [44]. Table (1) summarizes some of the popular genome databases.

Database Name	Description	Species Covered	Website
GenBank	A public database of DNA and RNA sequences	Wide range of organisms	https://www.ncbi.nlm.nih.gov/genbank/
Ensembl	A genome browser and annotation database	Wide range of organisms	https://www.ensembl.org/
dbSNP	A database of genetic variations, including SNPs	Wide range of organisms	https://www.ncbi.nlm.nih.gov/snp/
The Cancer Genome Atlas (TCGA)	A database of genomic, transcriptomic, and epigenomic data from cancer patients	Human	https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga
The Human Genome Variation Database (HGVD)	A database of genetic variations found in the Japanese population	Human (Japanese population)	http://www.genome.med.kyoto-u.ac.jp/SnpDB/
ExAC	A database of genetic variations identified in over 60,000 exomes from diverse populations	Wide range of organisms	http://exac.broadinstitute.org/
NCBI Gene	A database of genomic location, function, and sequence of genes	Wide range of organisms	https://www.ncbi.nlm.nih.gov/gene/
UCSC Genome Browser	A genome browser with access to a wide range of genomic data	Wide range of organisms	https://genome.ucsc.edu/
GTEx Portal	A database of transcriptomic data from multiple human tissues	Human	https://gtexportal.org/home/
ENCODE	A project to identify and annotate all functional elements in the human genome	Human	https://www.encodeproject.org/
dbGaP	A database of genotypic and phenotypic data from human studies	Human	https://www.ncbi.nlm.nih.gov/gap/
FlyBase	A database of genetic and genomic data for Drosophila melanogaster	Fruit fly (Drosophila melanogaster)	https://flybase.org/

4. Machine learning and deep learning models for genomics applications

Machine learning (ML) and deep learning (DL) are two powerful approaches in artificial intelligence that have been widely used in genomics applications to analyze and interpret large-scale genomic data. Here are some examples of ML and DL models for genomics applications [45-52]:

Random Forest: Random Forest is an ML algorithm that can be used for classification and regression tasks. It has been used to classify cancer subtypes based on gene expression data and to predict the functional effects of genetic variants.

Support Vector Machines (SVMs): SVMs are another ML algorithm that can be used for classification tasks. They have been used to predict the impact of genetic variants on protein function and to classify patients based on gene expression data.

Convolutional Neural Networks (CNNs): CNNs are a type of DL algorithm that has been used for image and sequence analysis. In genomics, CNNs have been used to predict DNA-protein binding sites and to classify DNA sequences based on their function.

Recurrent Neural Networks (RNNs): RNNs are a type of DL algorithm that can model sequential data, such as gene expression time series data. They have been used to predict gene expression levels and to identify novel regulatory elements.

Generative Adversarial Networks (GANs): GANs are a type of DL algorithm that can generate realistic synthetic data. In genomics, GANs have been used to generate synthetic DNA sequences and predict genetic variants' effects on protein structure.

Autoencoders: Autoencoders are a type of deep learning algorithm that can be used for unsupervised learning. They work by compressing input data into a lower-dimensional representation and then reconstructing the original data from the compressed representation. In genomics, autoencoders have been used to identify patterns in gene expression data, to cluster genes based on their expression profiles, and to predict gene expression levels.

Transfer Learning: Transfer learning is a technique that involves training a deep learning model on one task and then using the learned features to solve a different but related task. In genomics, transfer learning has been used to improve the performance of gene expression classification tasks by pre-training the model on related tasks, such as predicting protein-protein interactions.

Deep Reinforcement Learning: Deep reinforcement learning is a type of deep learning algorithm that can learn to make decisions based on rewards or penalties. In genomics, deep reinforcement learning has been used to design DNA sequences that have desired properties, such as high gene expression levels or low off-target effects.

Graph Neural Networks: Graph neural networks are a type of deep learning algorithm that can operate on graph data, such as protein-protein interaction networks or gene co-expression networks. They work by propagating information between nodes in the graph and updating node features based on that information. In genomics, graph neural networks have been used to predict gene functions, identify disease-associated genes, and classify cancer subtypes based on gene expression networks.

Bayesian Networks: Bayesian networks are a type of probabilistic graphical model that can be used to represent and reason about uncertainty in complex systems. In genomics, Bayesian networks have been used to model gene regulatory networks, identify disease-associated genes, and predict the effects of genetic variants on gene expression.

Literature review on applying machine and deep learning models for genomics applications

Several research scholars have applied machine and deep learning algorithms in the field of genomics applications. For example, a study published in Nature in 2018 used random forest to classify breast cancer subtypes based on DNA methylation data. The model accurately distinguished between different subtypes and identified novel subtype-specific biomarkers. Another example is a study published in Nature in 2015 that used a CNN to predict the DNA-binding specificities of transcription factors. The model was trained on a large dataset of DNA sequences and their corresponding transcription factor binding affinities and was able to predict the binding specificity of new transcription factors accurately. This approach has the potential to improve our understanding of gene regulation greatly and to aid in the development of new therapies. In 2016, a study published in Cell used an RNN to predict the expression levels of genes in response to different stimuli. The model was trained on a dataset of time-series gene expression data and could accurately predict the expression levels of genes in response to new stimuli. GNN is another machine learning algorithm used in a study published in Nature Genetics in 2019 to predict the effects of genetic variants on gene expression. The model was trained on a dataset of genetic variants and their impact on gene expression levels and could accurately predict the effects of new variants on gene expression. In 2018, a study that used an LSTM to predict the binding of transcription factors to DNA sequences was published in Genome Research. The model was trained on a dataset of DNA sequences and their corresponding transcription factor binding affinities and was able to accurately predict the binding of new transcription factors to DNA sequences. Variational Autoencoders (VAEs): VAEs are a type of generative model that can be used for unsupervised learning and dimensionality reduction. It has been used to analyze single-cell RNA sequencing data and to identify rare cell types in the mouse brain. The model was able to identify previously unknown cell types and to generate synthetic data that could be used to improve the accuracy of cell type classification. This study was published in Nature Communications in 2020. These are just a few more examples of the many machine learning and deep learning models applied in genomics research. As the field continues to evolve, new models and algorithms will likely be developed to address the unique challenges and opportunities of genomic data analysis. Table (2) summarizes some of the machine learning and deep learning models applied in genomics research.

Model	Type	Application	Year
Convolutional Neural Networks (CNNs)	Deep Learning	Prediction of gene expression levels, identification of regulatory regions, detection of genetic variants	2015
Recurrent Neural Networks (RNNs)	Deep Learning	Prediction of gene expression levels, identification of cis-regulatory elements, analysis of epigenetic data	2016
Autoencoders	Deep Learning	Unsupervised learning, dimensionality reduction, identification of patterns and features in the data	2018
Graph Neural Networks (GNNs)	Deep Learning	Prediction of the effects of genetic variants on gene expression, analysis of gene regulatory networks and protein-protein interaction networks	2019
Long Short-Term Memory (LSTM) Networks	Deep Learning	Prediction of gene expression levels, identification of cis-regulatory elements, analysis of epigenetic data	2018
Generative Adversarial Networks (GANs)	Deep Learning	Generative modeling, unsupervised learning, generation of synthetic data for training machine learning models	2018
Attention Mechanisms	Deep Learning	Prediction of gene expression levels, identification of splicing events, analysis of epigenetic data	2019
Variational Autoencoders (VAEs)	Deep Learning	Unsupervised learning, dimensionality reduction, generation of synthetic data for training machine learning models	2020
Graph Convolutional Networks (GCNs)	Deep Learning	Prediction of the effects of genetic variants on gene expression, analysis of gene regulatory networks and protein-protein interaction networks	2019

Model	Type	Application	Year
Transformer Networks	Deep Learning	Prediction of gene expression levels, identification of regulatory regions, analysis of epigenetic data, prediction of protein structure from amino acid sequences	2021
Capsule Networks	Deep Learning	Analysis of single-cell RNA sequencing data, identification of rare cell types in the human brain	2018
Deep Belief Networks (DBNs)	Deep Learning	Analysis of gene expression data, identification of subtypes of breast cancer with different clinical outcomes	2014
Siamese Networks	Deep Learning	Prediction of the effects of genetic variants on protein function	2019

5. Challenges in analyzing and interpreting genomics data using machine learning and deep learning

Despite the many successes of machine learning and deep learning in genomics, there are still some challenges in analyzing and interpreting genomics data using these approaches. Here are a few:

1. Data Quality: Genomics data is often noisy, incomplete, and subject to batch effects, which can make it challenging to analyze and interpret. Machine learning and deep learning models are sensitive to data quality, so it is crucial to carefully preprocess and clean the data before training the models.
2. Interpretability: Many machine learning and deep learning models are black boxes, meaning it can be challenging to understand how they make predictions or which features are most important for the predictions. This can make interpreting the results challenging and developing biological insights.
3. Sample Size: Some genomics datasets, such as single-cell RNA sequencing data, may have a small sample size, which can make it challenging to train and validate machine learning and deep learning models. Small sample sizes can also increase the risk of overfitting and reduce the generalizability of the results.
4. Missing Data: Genomics data may have missing values, which can complicate the analysis and interpretation. Imputation techniques can be used to fill in missing values, but these techniques may introduce bias or reduce the accuracy of the models.
5. Generalizability: Machine learning and deep learning models that are trained on one dataset may not generalize well to new datasets or populations. This is particularly challenging in genomics, where genetic and environmental factors vary widely across populations.
6. Dimensionality: Genomics data is high-dimensional, meaning that it has many features or variables. This can make it challenging to train machine learning and deep learning models, as the number of features may exceed the number of samples. Feature selection techniques can be used to reduce the dimensionality of the data, but these techniques may also introduce bias or reduce the accuracy of the models.
7. Class Imbalance: In genomics, some classes of samples or features may be much rarer than others, creating a class imbalance in the data. Machine learning and deep learning models may struggle to accurately predict the minority classes, as they may be underrepresented in the training data. Class balancing techniques, such as oversampling or undersampling, can be used to address this issue, but they may also introduce bias or reduce the accuracy of the models.

8. Reproducibility: Machine learning and deep learning models are highly dependent on the choice of hyperparameters, model architecture, and training data. This can make it difficult to reproduce the results or compare different models' performance. Standardized protocols and benchmarks can help to improve reproducibility and facilitate comparisons between models.
9. Validation: Validating machine learning and deep learning models in genomics can be challenging, as it may be difficult to obtain independent validation datasets or to perform functional experiments to confirm the results. Cross-validation and permutation testing can be used to estimate the performance of the models, but they may not always accurately reflect the real-world performance.
10. Integration: Genomics data is often integrated with other types of data, such as clinical data, imaging data, or environmental data. Integrating multiple data types can be challenging, as different types of data may have different scales, units, or distributions. Machine learning and deep learning models that can handle multiple data types, such as multi-modal neural networks, may be needed to effectively integrate these data types.
11. Ethical Concerns: There are also ethical concerns around the use of machine learning and deep learning in genomics, such as privacy concerns around the sharing of genomic data and the potential for algorithmic bias in the analysis and interpretation of the data.

These are just a few more examples of the challenges of using machine learning and deep learning in genomics. Addressing these challenges will require ongoing research and development of new methods, algorithms, and tools that can effectively analyze and interpret complex and heterogeneous genomics data.

References

Alipanahi, B., Delong, A., Weirauch, M. T., Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology* 33(8), 831-838.

Angermueller, C., Pärnamaa, T., Parts, L., Stegle, O. (2016). Deep learning for computational biology. *Molecular systems biology* 12(7), 878.

Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., Shameer, K. (2018). Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface* 15(141), 20170387.

Gao, J., Liang, F., Zhang, J., Chen, Y., Wang, Y., Zhu, F., Zhou, Y. (2021). Recent progress in deep learning in protein structure prediction. *Briefings in Bioinformatics* 22(2), 1958-1973.

Min, S., Lee, B., Yoon, S. (2017). Deep learning in bioinformatics. *Briefings in Bioinformatics* 18(5), 851-869.

Wang, L., Tang, Y. (2019). Applications of artificial intelligence in drug discovery and development. *Journal of Hematology & Oncology* 12(1), 54.

Zhang, C., Chen, H., Zhou, Y. (2019). Deep learning for genomics: A concise overview. AI Matters, 5(2), 11-23.

Ainscough, B. J., Rahman, F. Z., Glen, E., Lise, S., Ramsay, M. (2017). A roadmap for cost-efficient, high-quality genomic studies using FFPE samples. *Briefings in Bioinformatics* 19(2), 229-237.

Chen, H., Ding, S., Zhou, Y. (2019). High-throughput sequencing-based immune repertoire study: A high-resolution approach for immunological diseases. *Journal of Immunology Research* 2019.

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alfoldi, J., Wang, Q., Genome Aggregation Database Consortium. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581(7809), 434-443.

Khoury, M. J., Gwinn, M., Ioannidis, J. P., Little, J. (2014). Omics data integration for epidemiologic research: opportunities and challenges. *American journal of epidemiology* 180(11), 1148-1154.

Lee, Y., Kim, J. H. (2018). Machine learning in genomics: tools, resources, and progress. *Briefings in Bioinformatics* 19(4), 737-747.

Li, Y., Liu, L. (2018). The application of artificial intelligence in cancer immunotherapy: recent advances and future prospects. *Theranostics* 8(20), 5519.

Wang,Y., Huang, H. (2020). Artificial intelligence in drug discovery: present status and future prospects. *Frontiers in Chemistry* 8, 604.

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Dai, J. (2018). ClusterProfiler 3.0: a versatile and comprehensive R package for enrichment analysis of gene and gene clusters. *Bioinformatics* 34(11), 2021-2023.

Xu, Y., Mo, S., Feng, Q. (2021). Recent advances of artificial intelligence in single-cell omics research. *Briefings in Bioinformatics* 22(5), 2090-2101.

Berlin, K., Koren, S., Chin, C. S., Drake, J. P., Landolin, J. M., Phillippy, A. M. (2015). Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nature Biotechnology* 33(6), 623-630.

Chin, C. S., Peluso, P., Sedlazeck, F. J., Nattestad, M., Concepcion, G. T., Clum, A., Hall, R. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods* 13(12), 1050-1054.

Zhang, F., Wen, Y. (2019). Genome editing using CRISPR-Cas9: from basic research to translational medicine. *Journal of dental research* 98(7), 751-760.

Jha, P., Biswas, R. (2020). AI-assisted gene editing: current progress, challenges and future prospects. *Briefings in Functional Genomics* 20(5), 295-310.

Li, J., Zhao, H. (2018). Next-generation sequencing and CRISPR/Cas genomes editing: technologies and applications for food microbiology research. *Frontiers in microbiology* 9, 1958.

Collins, F. S., Varmus, H. (2015). A new initiative on precision medicine. *New England Journal of Medicine* 372(9), 793-795.

Han, L., Ma, Q., Li, C., Liu, Y., Zhao, B., Li, Y. (2021). Personalized Medicine in the Era of Big Data: A Review. *Journal of Healthcare Engineering* 2021.

Lu, J., Getz, G. (2018). Merging multi-omics data for cancer prognosis and therapeutics. *Drug discovery today* 23(3), 692-700.

Relling, M. V., Evans, W. E. (2015). Pharmacogenomics in the clinic. *Nature* 526(7573), 343-350.

Saria, S., Butte, A. J. (2015). Making big data useful for health care: a summary of the inaugural MIT critical data conference. *Journal of General Internal Medicine* 30(S3), 604-609.

Zeng, X., Zhang, X., Zou, Q. (2019). Integrative approaches for predicting microRNA function and prioritizing disease-related microRNA using biological interaction networks. *Briefings in bioinformatics* 20(4), 1284-1298.

Eder, J., Sedrani, R., Wiesmann, C. (2014). The discovery of first-in-class drugs: origins and evolution. *Nature Reviews Drug Discovery* 13(8), 577-587.

Hughes, J. P., Rees, S., Kalindjian, S. B., Philpott, K. L. (2011). Principles of early drug discovery. *British journal of pharmacology* 162(6), 1239-1249.

Langer, R., Tirrell, D. A. (2004). Designing materials for biology and medicine. *Nature* 428(6982), 487-492.

Ma, D. L., Chan, D. S. (2014). Recent developments in drug discovery at the Chinese University of Hong Kong. Expert opinion on drug discovery, 9(7), 775-787.

Paul, S. M., Mytelka, D. S., Dunwiddie, C. T., Persinger, C. C., Munos, B. H., Lindborg, S. R., Schacht, A. L. (2010). How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery* 9(3), 203-214.

Alqahtani, F., Gao, J. (2019). Recent advances in genome annotation. *Briefings in functional genomics* 18(6), 393-401.

Chawla, K., Kuiper, M. (2021). Machine learning in genomics: a review. Current opinion in genetics & development, 66, 1-9.

Edsall, L., Agrawal, S. (2019). Genome annotation: from sequence to biology. *Nature education knowledge* 10(2), 1-8.

Lee, H., Kang, C. (2019). Genome-wide annotation of human lncRNA stability with RNA-binding protein motifs. *Scientific reports* 9(1), 1-11.

Li, Y., Zhang, Y. (2020). Artificial intelligence in genomics: a review. *Briefings in bioinformatics* 21(4), 1612-1627.

Ma, J., Yu, M. K. (2021). A review of recent advances in the application of machine learning and artificial intelligence in genomics. *Frontiers in genetics* 12, 470.

Bin Abdulrahman, A. K., Alkhateeb, A. (2020). Artificial intelligence in cancer genomics: A review. *Cancer genomics & proteomics* 17(6), 641-655.

Chen, C. J., Li, H. (2020). Role of artificial intelligence in cancer diagnosis. *World journal of clinical cases* 8(11), 2155-2170.

Gao, X., Li, J., Zhang, Y., Wei, G. W. (2021). AI in cancer diagnosis and prognosis: challenges and perspectives. *Frontiers in oncology* 11, 744.

Li, H., Chen, C. J. (2021). Artificial intelligence in cancer genomics and precision oncology. *Cancer letters* 500, 199-210.

Yates, A., Akanni, W., Amode, M. R., Barrell, D., Billis, K., Carvalho-Silva, D., Zerbino, D. R. (2016). Ensembl 2016. Nucleic Acids Research, 44(D1), D710–D716.

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., Exome Aggregation Consortium. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536(7616), 285–291.

LeCun, Y., Bengio, Y., Hinton, G. (2015). Deep learning. *Nature* 521(7553), p.p 436-444.

Alipanahi, B., Delong, A., Weirauch, M. T., Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology* 33(8), 831-838. https://doi.org/10.1038/nbt.3300

Lanchantin, J., Singh, R., Wang, B. (2016). Deep motif dashboard: Visualizing and understanding genomic sequences using deep neural networks. *Nucleic Acids Research* 44(W1), W239-W245. https://doi.org/10.1093/nar/gkw377

Hochreiter, S., Schmidhuber, J. (1997). Long short-term memory. *Neural Computation* 9(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Bengio, Y. (2014). Generative adversarial nets. In Advances in Neural Information Processing Systems 2672-2680. https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494a2d4d47e8ac-Paper.pdf

Kingma, D. P., Welling, M. (2013). Auto-encoding variational Bayes. In Proceedings of the International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1312.6114

Pan, X., Shen, H. B. (2017). Deep learning for miRNA target prediction: An overview. Wiley Interdisciplinary Reviews: *Data Mining and Knowledge Discovery* 7(5), e1213. https://doi.org/10.1002/widm.1213

Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., Telenti, A. (2019). A primer on deep learning in genomics. *Nature Genetics* 51(1), 12-18. https://doi.org/10.1038/s41588-018-0295-5