DOI: 10.21088/ijnns.0975.0223.17225.4

International Journal of Neurology and Neurosurgery
Volume 17, Number 2 May-August 2025

REVIEW ARTICLE

Use of AI in Biological Databases

Ayushi Mishra[1], Shrijal Singh[2], Prachi Srivastava[3]

HOW TO CITE THIS ARTICLE:

Ayushi Mishra, Shrijal Singh, Prachi Srivastava. Use of AI in Biological Databases. International Journal of Neurology and Neurosurgery. 2025; 17(2): 88-101.

ABSTRACT

Biological databases are fundamental to modern life sciences, serving as repositories for genomic sequences, protein structures, metabolic pathways, and more. With the increasing complexity and volume of biological data, artificial intelligence (AI) has become indispensable in enhancing data integration, analysis, and interpretation. This review explores the role of AI in transforming biological databases, highlighting key applications such as protein structure prediction, functional annotation, pandemic surveillance, and antibiotic resistance analysis. AI-driven tools like Alpha Fold have revolutionized protein structure determination, while UniProt leverages machine learning for automated protein function annotation. In genomic research, AI enhances text mining in PubMed through PubTator, aiding in extracting relevant biological information. AI also facilitates pathway predictions in KEGG, improves virus tracking in GISAID, and strengthens protein-protein interaction predictions in STRING. Furthermore, AI has advanced CRISPR-Cas9 off-target effect predictions, ensuring safer genome editing, and has been pivotal in analyzing antibiotic resistance through the CARD database. Despite challenges such as data heterogeneity and model interpretability, AI continues to expand the capabilities of biological databases, accelerating research, innovation, and applications in medicine, drug discovery, and personalized healthcare.

KEYWORDS

• Artificial Intelligence • Biological Databases • CRISPR-Cas9 • KEGG • STRING

AUTHOR'S AFFILIATION:

[1]Student, Amity Institute of Biotechnology, Amity University, Lucknow, Uttar Pradesh, India.
[2]Student, Amity Institute of Biotechnology, Amity University, Lucknow, Uttar Pradesh, India.
[3] Associate Professor, Amity Institute of Biotechnology, Amity University, Lucknow, Uttar Pradesh, India.

CORRESPONDING AUTHOR:

Prachi Srivastava, Amity Institute of Biotechnology, Amity University, Lucknow, Uttar Pradesh, India.
E-mail: psrivastava@amity.edu

Received: 13-03-2025 **Accepted:** 08-05-2025

Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution NonCommercial 4.0 License (http://www.creativecommons.org/licenses/by-nc/4.0/) *which permits non-Commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the Red Flower Publication and Open Access pages* (https://www.rfppl.co.in)

INTRODUCTION

Biological databases are very essential tools in modern life sciences, serving as repositories for storing, retrieving, and analyzing vast amounts of biological data. These databases integrate diverse datasets, including genomic sequences, protein structures, gene expressions, metabolic pathways, and more, making them invaluable for research in genomics, proteomics, and systems biology. For instance, GenBank, maintained by the National Center for Biotechnology Information (NCBI), is one of the most widely used nucleotide sequence repositories, enabling researchers to access, submit, and analyze DNA sequence data[1] including whole genome shotgun (WGS. Similarly, the Protein Data Bank (PDB) is a key resource for structural biology, offering three-dimensional structural data of biomolecules critical for understanding molecular function and drug design.[2]

Other specialized databases, such as UniProt, provide curated information about protein sequences and functions, supporting proteomic and functional annotation studies (UniProt Consortium, 2021). KEGG (Kyoto Encyclopedia of Genes and Genomes) offers insights into various metabolic pathways and molecular interactions, bridging the gap between genes and cellular functions Ensembl, a genome browser, allows integrated access to genomic data from various organisms and supports comparative genomics analyses[3] and it also provides deep comprehensive annotation for key species. Genomes representing a greater diversity of species are increasingly being sequenced. In response, we have focussed our recent efforts on expediting the annotation of new assemblies. Here, we report the release of the greatest annual number of newly annotated genomes in the history of Ensembl via our dedicated Ensembl Rapid Release platform (http://rapid.ensembl.orgas high-throughput sequencing technologies such as next-generation sequencing (NGS) continue to generate unprecedented volumes of data, biological databases have adapted by incorporating advanced algorithms, big data analytics, and cloud-based architectures to ensure efficient data storage, retrieval, and sharing.[4] Moreover, the establishment of specialized databases such as GISAID for genomic data related to viruses like SARS-CoV-2 highlights the critical role of these resources in addressing global health challenges.[5]

Artificial intelligence (AI) is revolutionizing biology by enabling the analysis of complex datasets and the discovery of patterns that were previously beyond human reach. Machine learning algorithms, a subset of AI, are widely used in genomics to predict gene functions, identify mutations, and analyze high-throughput sequencing data.[6] In proteomics, tools like AlphaFold leverage AI to predict protein structures with remarkable accuracy, addressing long-standing challenges in structural biology.[7] and understanding their structure can facilitate a mechanistic understanding of their function. Through an enormous experimental effort[1-4], the structures of around 100,000 unique proteins have been determined,[5] but this represents a small fraction of the billions of known protein sequences.[6,7] Structural coverage is bottlenecked by the months to years of painstaking effort required to determine a single protein structure. Accurate computational approaches are needed to address this gap and to enable large-scale structural bioinformatics. Predicting the three-dimensional structure that a protein will adopt based solely on its amino acid sequence-the structure prediction component of the 'protein folding problem'8-has been an important open research problem for more than 50 years[9]. Despite recent progress[10-14], existing methods fall far short of atomic accuracy, especially when no homologous structure is available. Here we provide the first computational method that can regularly predict protein structures with atomic accuracy even in cases in which no similar structure is known. We validated an entirely redesigned version of our neural network-based model, AlphaFold, in the challenging 14th Critical Assessment of protein Structure Prediction (CASP14 AI also plays a pivotal role in drug discovery, accelerating target identification, virtual screening, and optimizing lead compounds.[8]

Artificial intelligence (AI) is transforming the landscape of biological databases by enhancing data integration, analysis, and interpretation. Traditional biological databases house vast amounts of heterogeneous data, including genomic sequences, protein structures, and functional annotations. However, analyzing this growing complexity and volume of data requires sophisticated computational

approaches. AI, particularly machine learning and deep learning, addresses this challenge by enabling automated data curation, prediction, and pattern recognition.

For instance, AI algorithms are increasingly employed to annotate gene and protein functions in databases like UniProt and Ensembl, significantly reducing manual effort[9] the overwhelming majority of protein products can only be annotated computationally. If computational predictions are to be relied upon, it is crucial that the accuracy of these methods be high. Here we report the results from the first large-scale community-based critical assessment of protein function annotation (CAFA. Natural language processing (NLP) techniques are used to extract relevant biological information from scientific literature and integrate it into databases, such as PubMed, for broader accessibility.[10] AI also enhances data quality through anomaly detection, ensuring consistency and reliability in datasets.

Moreover, AI-powered tools such as AlphaFold are integrated into databases like the Protein Data Bank to predict protein structures, expanding the range of data available for structural biology research.

This review aims to discuss the various applications of AI in biological databases, many of such databases are discussed as follows:

AlphaFold and the Protein Data Bank (PDB)

AlphaFold is an artificial intelligence (AI) system developed by Deep Mind that predicts the three-dimensional structures of proteins directly from their amino acid sequences with remarkable accuracy.[11] Protein structure prediction has long been a challenge in biology, as experimental methods like X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and cryo-electron microscopy are resource-intensive and time-consuming.[12] AlphaFold addresses these limitations by employing a computational approach that drastically reduces the time and cost of structure determination, enabling the scientific community to study proteins that were previously inaccessible. Its integration into the Alpha Fold Protein Structure Database and collaborations with the Protein Data Bank (PDB) have significantly expanded the availability of high-accuracy protein models, fueling advancements in structural biology, drug discovery, and biotechnology.[13]

The working principle of AlphaFold is centered on deep learning, particularly the use of attention based neural networks and transformers. It begins by analyzing protein sequences alongside evolutionary information derived from multiple sequence alignments (MSAs). These alignments help the model identify conserved residues critical for protein folding.[14] AlphaFold employs a novel "attention mechanism" that allows the model to focus on relationships between amino acid residues and their spatial configurations. The model predicts inter-residue distance maps and torsion angles, which are key determinants of a protein's 3D structure. By iteratively refining these predictions, AlphaFold outputs atomic-level coordinates of the protein structure.[15]

A key innovation of AlphaFold is its end-to-end learning approach, which allows the system to predict structures directly from raw sequence data without requiring intermediate modeling steps. Additionally, the system uses an energy function to ensure the predicted structure is physically plausible and consistent with known protein folding principles.[16] AlphaFold's predictions were validated during the CASP14 (Critical Assessment of protein Structure Prediction) competition, where it achieved near-experimental accuracy for most targets, significantly outperforming previous methods.

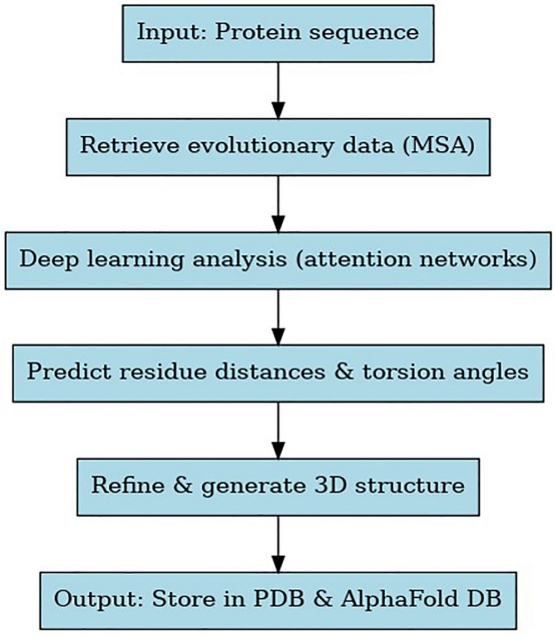


Figure 1: Flowchart depicting brief mechanism of AlphaFold and Protein Data Bank (PDB)

UniProt Annotation Using AI

UniProt is one of the most comprehensive and widely used biological databases, providing curated information about protein sequences and their functional annotations. The challenge in maintaining UniProt lies in the vast amount of biological data generated from sequencing projects, which makes manual curation increasingly unfeasible.[17] To address this, UniProt has integrated artificial intelligence (AI) tools to automate and enhance the annotation process, allowing researchers to derive meaningful insights from raw sequence data. AI in UniProt annotation focuses on automating the prediction of protein functions, interactions, subcellular localizations, and post-translational modifications. These predictions leverage machine learning models trained on extensive datasets of experimentally validated protein information.[18] For functional annotation, AI models analyze sequence features such as conserved motifs, domains, and evolutionary relationships to predict protein functions. Sequence similarity searches using tools like BLAST are augmented with machine learning algorithms to refine functional assignments. Support vector machines (SVMs) and neural networks help distinguish subtle functional differences between homologous proteins.[19] AI-driven approaches are also utilized to predict protein-protein interactions by analyzing co-evolutionary patterns and domain-domain interactions, which helps build comprehensive interaction networks that inform on biological pathway.[20] One of the critical roles of AI in UniProt is annotating hypothetical or uncharacterized proteins. Machine learning models predict potential functions based on homology and structural odelling, bridging the gap for proteins with no experimental data.[21] DeepGO, an AI-based tool, predicts Gene Ontology (GO) terms for proteins by leveraging sequence information and hierarchical relationships in GO terms, significantly improving the accuracy of functional annotations in UniProt.[22] Additionally, UniProt integrates InterPro, a database of protein families and domains, where AI-based algorithms in InterProScan predict functional sites and domain architectures, aiding in annotating novel proteins.[23] Text mining approaches, such as the use of PubTator, are employed to extract relevant annotations from scientific literature. These natural language processing (NLP) tools identify mentions of proteins, their functions, and interactions, which are then integrated into UniProt entries.[18] The AI-driven annotations in UniProt have catalyzed discoveries in several domains, including drug discovery, where functional annotations of target proteins help identify druggable sites and design inhibitors,[24] disease research, where understanding the roles of disease associated proteins aids in developing diagnostics and therapeutics, and synthetic biology, where annotated proteins provide a foundation for designing synthetic enzymes and pathways. Despite the advancements, AI still faces challenges, as it heavily depends on training datasets, and biases or errors in these datasets can propagate into predictions. Furthermore, certain protein functions and interactions require experimental validation to ensure accuracy.[20,21]

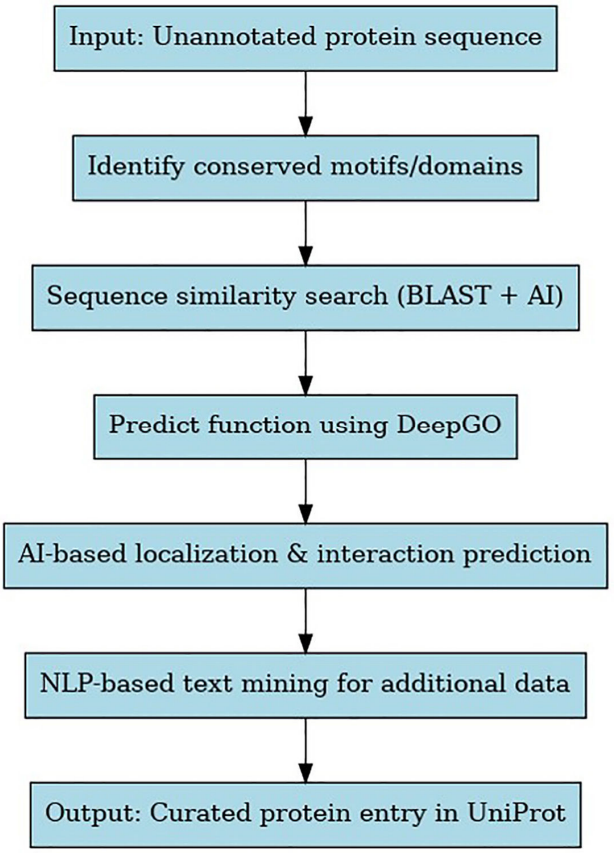


Figure 2: Flowchart depicting the mechanism of UniProt annotation using AI

Text Mining with PubTator in PubMed

Text mining is an essential tool in the biological sciences, helping researchers analyze large volumes of scientific literature efficiently. One prominent application of text mining in

biological research is PubTator, a web-based text mining tool designed to extract relevant biological entities, such as proteins, genes, diseases, and chemical compounds, from scientific literature in PubMed. PubTator enables researchers to rapidly annotate large datasets of articles, identifying and linking biological concepts to resources such as the Gene Ontology (GO), the Protein Data Bank (PDB), and clinical data repositories (21). This is particularly beneficial in fields like drug discovery, genomics, and clinical research, where large-scale data integration and annotation are critical for driving novel insights and accelerating the understanding of biological phenomena.[21]

PubTator works by utilizing a combination of natural language processing (NLP) algorithms and machine learning models to automatically annotate text. Initially, PubTator extracts and identifies biological terms from articles based on specific linguistic patterns or by leveraging pre-built dictionaries of biomedical terms.[25] The system is built on high-performance entity recognition tools that can identify genes, proteins, diseases, and other relevant terms in unstructured text. Once these entities are detected, PubTator then classifies them according to predefined categories such as diseases, drugs, or biological pathways. This classification process involves algorithms that use context and co-occurrence patterns within the text to ensure accurate annotation.[21]

In addition to entity recognition, PubTator supports named entity disambiguation, a crucial feature for distinguishing between entities with the same name but different meanings. For example, the term "EGFR" could refer to the epidermal growth factor receptor, a protein involved in cancer, or it could be used in different contexts in other diseases. PubTator employs advanced disambiguation techniques to associate the correct entity with the appropriate biological concept, drawing upon databases like UniProt, the National Center for Biotechnology Information (NCBI), and the Comparative Toxicogenomics Database (CTD) for resolution.[26]

PubTator has been widely adopted for multiple use cases in biological research. For instance, in the area of drug discovery, PubTator helps researchers identify key drug targets and the relationships between chemicals and diseases by processing large numbers of research papers and linking them to relevant databases.[25] Furthermore, PubTator's utility extends to clinical research, where it has been employed to track the latest findings on disease biomarkers and therapeutic interventions, facilitating faster updates to clinical guidelines and enhancing precision medicine efforts.[26] The tool is also useful in genomics by enabling rapid annotation of gene function, providing insights into gene-disease associations, and supporting the functional annotation of newly discovered genes.[27] Despite these challenges, PubTator continues to evolve, incorporating more sophisticated AI models for entity recognition and disambiguation.[27]

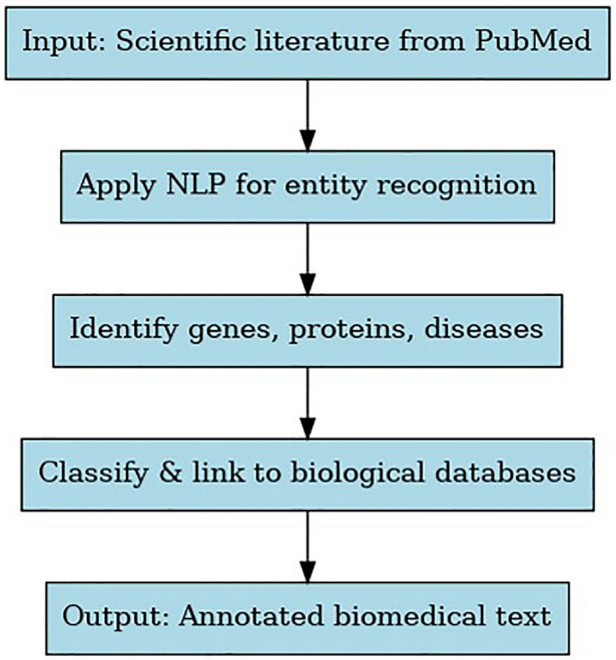


Figure 3: Flowchart depicting the mechanism of Text Mining with PubTator in PubMed.

KEGG Pathway Predictions with AI

KEGG (Kyoto Encyclopedia of Genes and Genomes) is a well-known bioinformatics resource that provides information on the molecular pathways governing cellular functions, as well as how genes and proteins interact within these pathways. KEGG pathways are essential for understanding biological processes, diseases, and drug mechanisms, but the vast number of pathways and their complexity make manual curation and analysis impractical. As a solution, artificial intelligence (AI) techniques have been employed to predict, map, and analyze KEGG pathways more efficiently, enhancing both the accuracy and speed of pathway annotations.

AI techniques, particularly machine learning models, have become increasingly valuable in predicting biological pathways from large-scale omics data, such as genomics, proteomics, and transcriptomics. These AI-driven predictions aim to identify which genes or proteins participate in specific biological pathways and how they interact with each other, ultimately providing insights into cellular processes, disease mechanisms, and potential drug targets. One of the key challenges in pathway prediction is dealing with the complexity of biological systems, where different genes and proteins may be involved in multiple pathways, and pathway components often interact in non-linear and dynamic ways.[28] AI models address this by using large-scale datasets and advanced algorithms that can detect patterns and relationships that may not be immediately obvious through traditional methods.

Machine learning techniques, such as deep learning and random forests, are employed in KEGG pathway prediction to analyze data from multiple sources, such as gene expression profiles, protein-protein interaction data, and gene sequence data. These models learn from existing annotated pathways in KEGG, as well as other biological databases like Reactome and BioCyc, to predict new pathway associations. For example, deep learning algorithms, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have been applied to sequence data to predict pathway membership based on sequence motifs and protein domains.[29] Additionally, supervised learning algorithms can classify genes and proteins according to their involvement in different pathways by training on labeled datasets of known interactions and pathway components.

One prominent AI tool in KEGG pathway predictions is PathPredict, a machine learning based platform that predicts pathway associations for genes and proteins based on their expression profiles and sequence data. By integrating omics data with prior knowledge of biological pathways, PathPredict provides insights into which pathways may be impacted by gene dysregulation in diseases, such as cancer or metabolic disorders.[30] The system also provides probabilistic pathway maps that allow researchers to visualize how individual genes and proteins are connected within the context of larger biological systems.

Another example of AI application in KEGG pathway prediction is the use of network-based algorithms, such as Graph Convolutional Networks (GCNs), to model protein-protein interaction networks.[31] GCNs can predict how proteins interact and participate in various biological pathways by learning from the connectivity and topological features of known protein interaction networks.

AI-based pathway prediction also has significant applications in drug discovery and disease research. For instance, pathway based models can help identify potential drug targets by identifying genes and proteins that are key drivers of disease pathways. Machine learning models can analyze gene expression data from patients with specific diseases, identify disrupted pathways, and suggest therapeutic strategies based on predicted interactions within those pathways. This approach can provide valuable insights for precision medicine and the development of targeted therapies.[28,32]

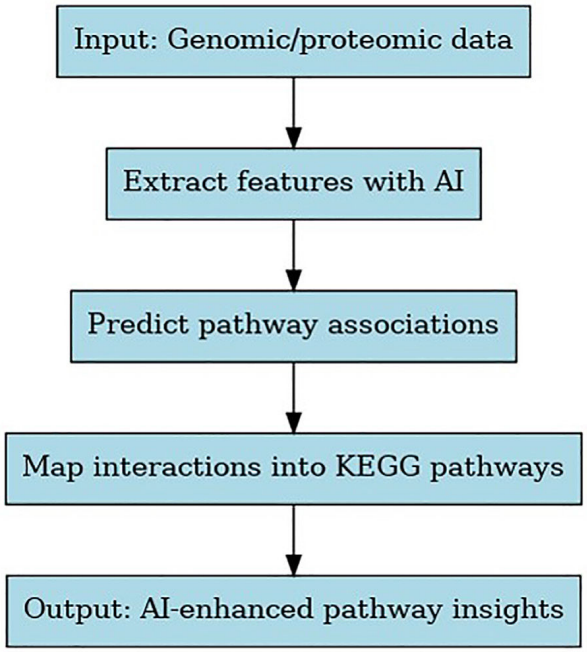


Figure 4: Flowchart depicting the mechanism of KEGG Pathway Predictions with AI

GISAID and AI in Pandemic Surveillance

The Global Initiative on Sharing All Influenza Data (GISAID) is a pivotal database that facilitates the sharing of genomic data on influenza viruses and other pathogens, including SARS-CoV-2. This platform has been instrumental in pandemic surveillance, enabling scientists to track the evolution

of viruses, monitor mutations, and inform public health responses. The integration of artificial intelligence (AI) into GISAID's data analysis workflows has further enhanced its capacity to manage and analyze the vast amounts of genomic data generated during global pandemics. AI-driven approaches help streamline genomic surveillance, detect emerging variants, and predict the implications of viral mutations on disease transmissibility, severity, and vaccine efficacy.

AI plays a crucial role in analyzing the massive influx of viral genome sequences uploaded to GISAID daily. Deep learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), are employed to identify patterns in viral sequences, enabling the rapid classification of variants based on their genetic signatures.[33] For instance, AI models can detect critical mutations in the spike protein of SARS-CoV-2, which are associated with increased transmissibility or immune evasion, thereby identifying variants of concern (VoCs) and variants under monitoring.[34]

Natural language processing (NLP) techniques have been applied to annotate genome sequences with metadata, including information on geographic location, collection date, and associated clinical outcomes. For example, clustering algorithms have been used to identify the geographic spread of SARS-CoV-2 variants and predict how specific variants may propagate under varying public health conditions.[35] These insights are critical for designing targeted interventions, such as lockdowns or vaccination campaigns, in regions at higher risk.

Machine learning models have also been utilized to predict the phenotypic effects of mutations detected in GISAID data. Tools like AlphaFold have been adapted to model structural changes in viral proteins caused by specific mutations, providing insights into how these changes might affect viral infectivity or neutralization by antibodies.[36] Additionally, AI-driven predictive odelling has been used to assess the effectiveness of existing vaccines against emerging variants by simulating interactions between viral antigens and antibodies derived from vaccinated individuals.[37]

AI enhances pandemic preparedness by enabling real-time monitoring of viral evolution. For instance, anomaly detection algorithms are used to flag unusual mutation patterns in viral sequences uploaded to GISAID, which might indicate the emergence of new VoCs. This capability allows for the early identification of potential threats and helps guide public health responses, such as updating diagnostic tests, vaccines, and therapeutics.[38] AI applications in GISAID also extend to the detection of zoonotic spillovers, which are often the source of pandemics. Machine learning models trained on genomic data from animal reservoirs have been used to predict the likelihood of cross-species transmission of viruses, enabling proactive surveillance and mitigation efforts in high-risk regions.[39] Despite these advancements, challenges remain in integrating AI into GISAID workflows. The reliability of AI models depends heavily on the quality and completeness of the input data, and biases or gaps in the data can lead to inaccurate predictions.[40]

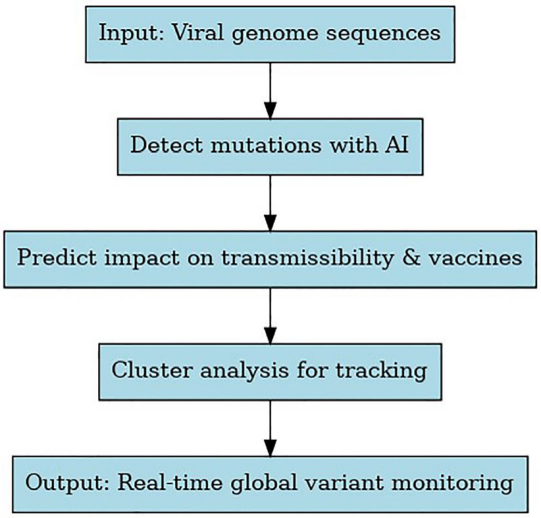


Figure 5: Flowchart depicting the mechanism of GISAID and AI in Pandemic Surveillance

AI in STRING for Protein-Protein Interaction Predictions

The STRING database is a comprehensive resource that consolidates known and predicted protein-protein interactions (PPIs) from various sources, including experimental data, computational predictions, and public text collections.[41] The integration of artificial intelligence (AI) into STRING has significantly enhanced its predictive capabilities, allowing for more accurate and extensive mapping of PPIs.[42] AI methodologies, particularly

machine learning (ML) and deep learning (DL), have been instrumental in advancing PPI predictions. These approaches analyze vast datasets to identify patterns and relationships that may not be evident through traditional methods.[43] For instance, ML-based tools have been developed to predict PPIs by utilizing diverse datasets and input features, employing various architectures to improve prediction accuracy.[44] Similarly, DL techniques have been applied to predict protein functions and interactions, demonstrating significant advancements in the field.[45]

The application of AI in PPI prediction is not limited to STRING alone. Tools like Protein Prompt have been designed to quickly and reliably predict contact propensities based on input sequences, enabling the scanning of large sequence libraries for potential interactions.[46] Additionally, INtERAcT utilizes vector representations of words to infer interaction networks from biomedical literature, showcasing the versatility of AI in extracting meaningful biological relationships.[47] Despite these advancements, the reliability of PPI predictions using databases like STRING has been a topic of discussion among researchers.[48] While AI has improved predictive accuracy, the quality of the input data and the inherent limitations of computational models necessitate cautious interpretation of results.[49] Evaluations of link prediction tools on PPI networks from multiple species have highlighted the importance of continuous refinement and validation of these models.[50,51]

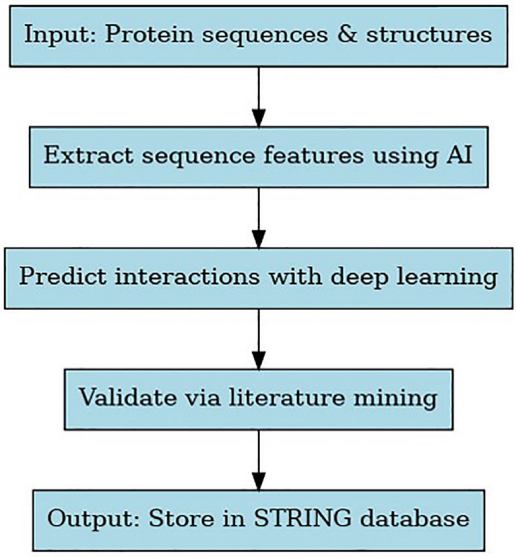


Figure 6: Flowchart depicting the mechanism of AI in STRING for Protein-Protein Interaction Predictions

AI for Predicting CRISPR-Cas9 Off-Target Effects

The use of artificial intelligence (AI) to predict off-target effects in CRISPR-Cas9 genome editing has become an area of growing research, enhancing the overall efficiency and safety of gene editing techniques. CRISPR-Cas9, while revolutionary for its ability to target specific genetic sequences, also faces challenges due to the risk of unintended DNA modifications, which could result in harmful mutations. AI has become an essential tool in predicting these off-target effects, aiming to ensure that CRISPR-based therapies are more accurate and safer, particularly in clinical applications.[52,53] These AI models help researchers understand the interactions between guide RNAs and their target sequences by analyzing vast amounts of genomic data, which would be too complex for traditional computational methods.

Deep learning, in particular, has shown promise in improving the specificity of CRISPR-Cas9 technology by refining the prediction of potential off-target sites. A notable study by Chen *et al*. (2023) introduced an AI-based model that used neural networks to predict the off-target effects of various CRISPR guide RNAs, significantly reducing false positives and enhancing editing accuracy. By training on extensive datasets containing both on-target and off-target sequences, these models can identify patterns in the DNA sequences that are more likely to cause unintended cutting by the Cas9 enzyme.[54] Additionally, machine learning algorithms are increasingly being utilized for feature selection, where important features in genomic data such as DNA sequence context and chromatin accessibility are identified to predict where off-target effects are more likely to occur.[55]

AI has also facilitated the identification of the most effective guide RNAs for CRISPR-Cas9 gene editing. By predicting the binding efficiency of the guide RNA to its intended target, AI models can assist researchers in selecting the most effective guides for their experiments while minimizing off-target risks.[56,57] Despite these advancements, challenges remain in the application of AI to CRISPR-Cas9 genome editing. One significant challenge is the need for large, high-quality datasets to train these AI models. Genomic data is inherently noisy, and high-quality

datasets that include all possible variables are difficult to obtain. Additionally, while AI models can accurately predict off-target effects, the interpretation of these predictions remains a challenge. The models may be able to identify regions of the genome that are likely to be cleaved by Cas9, but understanding the biological implications of these predictions requires further experimental validation.[56]

Furthermore, the integration of AI into CRISPR-based therapeutics faces ethical and regulatory challenges. Ethical considerations, such as the potential for unintended genetic modifications and the equitable distribution of CRISPR-based treatments, must be addressed as AI technologies are further integrated into clinical settings. AI models must also comply with stringent regulatory guidelines to ensure that they provide reliable predictions that will not result in harmful genetic alterations in clinical trials.[58]

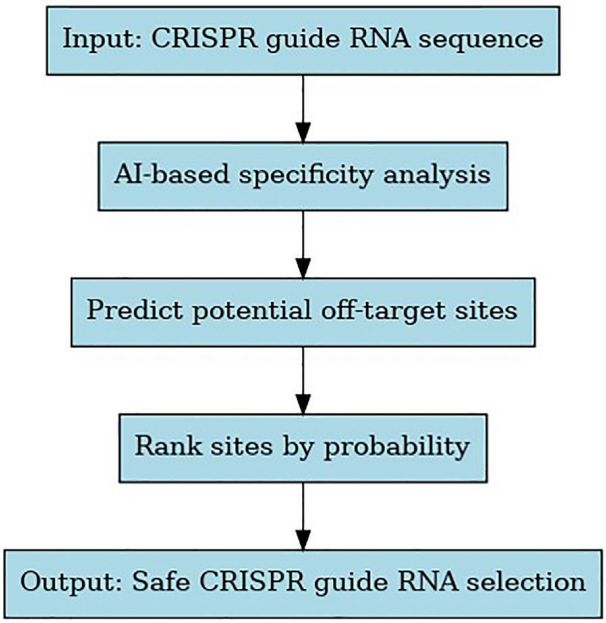


Figure 7: AI for Predicting CRISPR-Cas 9 Off-Target Effects

AI for Antibiotic Resistance Analysis in CARD

The growing issue of antibiotic resistance represents one of the most pressing challenges in modern medicine, and artificial intelligence (AI) has emerged as a powerful tool in combating this global health threat. The Comprehensive Antibiotic Resistance Database (CARD) plays a crucial role in identifying and understanding antibiotic resistance mechanisms, as it collects and curates comprehensive data on resistance genes, mutations, and the interactions of antibiotics with microbial genomes. AI, particularly machine learning (ML) and deep learning (DL), has been applied to the CARD to enhance the detection, prediction, and understanding of antibiotic resistance patterns, ultimately helping to guide treatment strategies and identify potential new antibiotics.

One of the major contributions of AI in antibiotic resistance analysis has been its ability to predict and identify resistance genes from microbial genomes. By analyzing large scale genomic data, AI models can pinpoint the presence of known resistance genes and predict the likelihood of new, unknown resistance genes emerging. A significant breakthrough in this domain was the development of AI-powered algorithms to automatically predict antibiotic resistance by analyzing bacterial genomic sequences.[59] These AI models were trained on vast datasets containing both resistant and non-resistant bacterial strains, enabling them to identify resistance profiles in novel bacterial genomes.[60] The CARD database serves as a critical resource for training these models, offering valuable data that enhances the accuracy and robustness of the predictions.[61]

Deep learning, a subset of AI, has proven particularly useful in analyzing the genetic and functional aspects of antibiotic resistance. Through the use of convolutional neural networks (CNNs) and recurrent neural networks (RNNs), deep learning models have been able to predict the presence of resistance genes in previously sequenced genomes with higher accuracy than traditional methods.[62] These models also offer the advantage of learning complex patterns from genomic data without the need for extensive feature engineering, making them particularly effective in the context of genomic analysis where numerous variables and interactions are at play. AI-driven approaches also allow for the prediction of novel mutations and genetic variations that may contribute to antibiotic resistance, even in strains where no resistance gene has been previously identified.[63]

Furthermore, AI models have been applied to predict the spread and evolution of antibiotic-resistant bacteria. For example, AI techniques like reinforcement learning have been used to model how antibiotic resistance might

propagate across different populations of bacteria under various treatment conditions.[64] By simulating different scenarios, AI can provide insights into the dynamics of resistance development and help design more effective strategies to combat antibiotic resistance.[65]

One of the most significant challenges in antibiotic resistance analysis is the vastness of the data involved. Genomic data from diverse bacterial strains, clinical data on antibiotic resistance profiles, and environmental factors all contribute to the complexity of understanding how antibiotic resistance spreads and evolves. AI can assist in this regard by integrating multiple types of data into cohesive models that provide a more holistic view of resistance patterns. For example, integrating environmental data and clinical data with genomic sequences from CARD has enabled the creation of more comprehensive models that predict resistance patterns in real-world environments.[66]

AI has also contributed to the identification of potential drug candidates capable of overcoming resistance mechanisms. By leveraging AI for drug repurposing, researchers can analyze large compound libraries to identify drugs that could be effective against antibiotic-resistant strains. AI models can predict how certain compounds may interact with bacterial resistance proteins, potentially offering new therapeutic options that have not been considered using traditional methods.[67] Such approaches are crucial as they provide an alternative pathway to developing new antibiotics, which is essential given the limited pipeline for new antibiotic development.[68]

Despite the promising applications of AI in antibiotic resistance analysis, there are several challenges that must be addressed. One of the primary concerns is the availability and quality of data. To train robust AI models, large, high-quality datasets are essential.[69] Additionally, ethical and privacy concerns regarding the use of genomic data in AI applications must be considered, as these models may inadvertently expose sensitive information about microbial strains and their resistance profiles.[70]

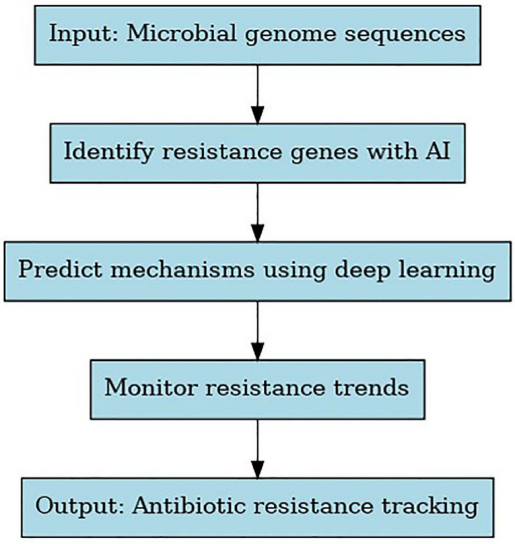


Figure 8: Flowchart depicting the mechanism of AI for Antibiotic Resistance Analysis in CARD

CONCLUSION

The integration of artificial intelligence (AI) into biological databases has revolutionized research by enhancing precision, scalability, and efficiency. AI applications span protein structure prediction, functional annotation, pathogen surveillance, and antibiotic resistance analysis, addressing longstanding challenges in biology. Machine learning, deep learning, and natural language processing (NLP) have enabled databases to not only store vast data but also analyze and interpret it in unprecedented ways.

A key example is AlphaFold's integration with the Protein Data Bank (PDB), which has transformed protein structure prediction with remarkable accuracy. This has accelerated insights into protein functions, interactions, and disease roles. Similarly, AI-driven annotation in UniProt enhances protein function analysis while reducing manual efforts. In text mining, PubTator leverages NLP to extract biomedical information from PubMed, automating the identification of genes, proteins, and diseases, thus expediting biological discoveries. AI also improves pathway prediction in the Kyoto Encyclopedia of Genes and Genomes (KEGG), uncovering novel biochemical interactions that support drug discovery.

AI has been instrumental in pathogen surveillance, as demonstrated in GISAID during the COVID-19 pandemic, where it tracked viral mutations and informed vaccine development. AI models extend this capability to forecasting virus and bacterial evolution, aiding infectious disease management. In CRISPR-Cas9 gene editing, AI enhances off-target effect predictions, ensuring safer applications. The Comprehensive Antibiotic Resistance Database (CARD) employs AI to predict resistance patterns, aiding antimicrobial drug development. In metabolomics, AI in the Human Metabolome Database (HMDB) improves metabolite identification and biomarker discovery, advancing disease research.

Despite challenges like data heterogeneity and the need for continuous algorithm refinement, AI's role in biological databases continues to grow. As AI technology evolves, it will drive innovation in medicine, biotechnology, and environmental science, offering transformative solutions for global health challenges.

Conflict of Interest: Nil

REFERENCES

1. Benson D.A., Cavanaugh M., Clark K., Karsch-Mizrachi I., Lipman D.J., Ostell J., *et al.* GenBank. Nucleic Acids Res. 2013 Jan 1; 41(D1): D36–42.
2. Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H, *et al*. The Protein Data Bank. Nucleic Acids Res. 2000 Jan 1; 28(1): 235–42.
3. Cunningham F., Allen J.E., Allen J., Alvarez-Jarreta J., Amode M.R., Armean I.M., *et al.* Ensembl 2022. Nucleic Acids Res. 2022 Jan 7; 50(D1): D988–95.
4. Stephens Z.D., Lee S.Y., Faghri F., Campbell R.H., Zhai C., Efron MJ, *et al.* Big Data: Astronomical or Genomical? PLoS Biol. 2015 Jul; 13(7): e1002195.
5. Elbe S., Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health. Glob Chall Hoboken NJ. 2017 Jan; 1(1): 33–46.
6. Angermueller C., Pärnamaa T., Parts L., Stegle O. Deep learning for computational biology. Mol Syst Biol. 2016 Jul 29; 12(7): 878.
7. Jumper J., Evans R., Pritzel A., Green T., Figurnov M., Ronneberger O., *et al.* Highly accurate protein structure prediction with AlphaFold. Nature. 2021 Aug; 596(7873): 583–9.
8. Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T. The rise of deep learning in drug discovery. Drug Discov Today. 2018 Jun; 23(6): 1241–50.
9. Radivojac P., Clark W.T., Oron T.R., Schnoes A.M., Wittkop T., Sokolov A., *et al.* A large-scale evaluation of computational protein function prediction. Nat Methods. 2013 Mar; 10(3): 221–7.
10. Lu Z. PubMed and beyond: a survey of web tools for searching biomedical literature. Database J Biol Databases Curation. 2011; 2011: baq036.
11. Senior A.W., Evans R., Jumper J., Kirkpatrick J., Sifre L., Green T., *et al.* Improved protein structure prediction using potentials from deep learning. Nature. 2020 Jan; 577 (7792): 706–10.
12. Janssen B.J.C., Gros P. Structural insights into the central complement component C3. Mol Immunol. 2007 Jan; 44(1–3): 3–10.
13. Tunyasuvunakool K., Adler J., Wu Z, Green T., Zielinski M., Žídek A, *et al.* Highly accurate protein structure prediction for the human proteome. Nature. 2021 Aug; 596(7873): 590–6.
14. 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. Proc Natl Acad Sci U S A. 2021 Apr 13; 118(15): e2016239118.
15. Rao R., Liu J, Verkuil R, Meier J, Canny JF, Abbeel P, *et al*. MSA Transformer [Internet]. bioRxiv; 2021 [cited 2025 Jan 13]. p. 2021.02.12.430858. Available from: https://www.biorxiv.org/content/10.1101/2021.02.12.430858v3
16. 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 Aug 20; 373(6557): 871–6.
17. Zeng M., Li M., Wu FX, Li Y., Pan Y. DeepEP: a deep learning framework for identifying essential proteins. BMC Bioinformatics. 2019 Dec 2; 20(16): 506.
18. Hong J, Luo Y., Zhang Y., Ying J., Xue W., Xie T., *et al*. Protein functional annotation of simultaneously improved stability, accuracy and false discovery rate achieved by a sequence-based deep learning. Brief Bioinform. 2019 Aug 2; 21(4): 1437–47.
19. Yu C.S., Chen Y.C., Lu C.H., Hwang J.K. Prediction of protein subcellular localization.

Proteins. 2006 Aug 15; 64(3): 643–51.

20. Liu S., Liu C., Deng L. Machine Learning Approaches for Protein−Protein Interaction Hot Spot Prediction: Progress and Comparative Assessment. Mol Basel Switz. 2018 Oct 4; 23(10): 2535.

21. Wei C.H., Kao H.Y, Lu Z. PubTator: a web-based text mining tool for assisting biocuration. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W518-522.

22. Hong J., Luo Y., Zhang Y., Ying J., Xue W., Xie T., *et al*. Protein functional annotation of simultaneously improved stability, accuracy and false discovery rate achieved by a sequence-based deep learning. Brief Bioinform. 2019 Aug 2; 21(4): 1437–47.

23. Yue Y., Ye C., Peng P.Y., Zhai H.X., Ahmad I., Xia C., *et al*. A deep learning framework for identifying essential proteins based on multiple biological information. BMC Bioinformatics. 2022 Aug 4; 23(1): 318.

24. Yu C.S., Chen Y.C., Lu C.H., Hwang J.K. Prediction of protein subcellular localization. Proteins. 2006 Aug 15; 64(3): 643–51.

25. Jin Q., Leaman R., Lu Z. PubMed and beyond: biomedical literature search in the age of artificial intelligence. eBioMedicine. 2024 Feb 1; 100: 104988.

26. Wei C.H., Allot A., Leaman R., Lu Z. PubTator central: automated concept annotation for biomedical full text articles. Nucleic Acids Res. 2019 Jul 2; 47(W1): W587–93.

27. Zhao S., Su C., Lu Z., Wang F. Recent advances in biomedical literature mining. Brief Bioinform. 2020 May 18;22(3):bbaa057.

28. Zhang Y., Luo M., Wu P., Wu S., Lee T.Y., Bai C. Application of Computational Biology and Artificial Intelligence in Drug Design. Int J Mol Sci. 2022 Nov 5;23(21):13568.

29. Helmy M., Smith D., Selvarajoo K. Systems biology approaches integrated with artificial intelligence for optimized metabolic engineering. Metab Eng Commun. 2020 Dec 1; 11: e00149.

30. van Hilten A., Kushner SA, Kayser M., Ikram M.A., Adams H.H.H., Klaver C.C.W., *et al*. GenNet framework: interpretable deep learning for predicting phenotypes from genetic data. Commun Biol. 2021 Sep 17;4:1094.

31. Kipf T.N., Welling M. Semi-Supervised Classification with Graph Convolutional Networks [Internet]. arXiv; 2017 [cited 2025 Jan 20]. Available from: http://arxiv.org/abs/1609.02907

32. Janizek J.D., Spiro A., Celik S., Blue B.W., Russell J.C., Lee T.I., *et al*. PAUSE: principled feature attribution for unsupervised gene expression analysis. Genome Biol. 2023 Apr 19;24(1):81.

33. Tosta S., Moreno K., Schuab G., Fonseca V., Segovia F.M.C., Kashima S., *et al*. Global SARS-CoV-2 genomic surveillance: What we have learned (so far). Infect Genet Evol. 2023 Mar 1;108:105405.

34. Singh J., Rahman S.A., Ehtesham N.Z., Hira S, Hasnain S.E. SARS-CoV-2 variants of concern are emerging in India. Nat Med. 2021 Jul; 27(7):1131–3.

35. Dhama K., Nainu F., Frediansyah A., Yatoo MohdI, Mohapatra R.K., Chakraborty S., *et al*. Global emerging Omicron variant of SARS-CoV-2: Impacts, challenges and strategies. J Infect Public Health. 2023 Jan 1; 16(1): 4–14.

36. Gutnik D., Evseev P., Miroshnikov K., Shneider M. Using AlphaFold Predictions in Viral Research. Curr Issues Mol Biol. 2023 Apr 21; 45(4): 3705–32.

37. Greaney A.J., Loes A.N., Crawford K.H.D., Starr T.N., Malone K.D., Chu H.Y., *et al*. Comprehensive mapping of mutations in the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human plasma antibodies. Cell Host Microbe. 2021 Mar 10; 29(3): 463-476.e6.

38. Shu Y., McCauley J. GISAID: Global initiative on sharing all influenza data – from vision to reality. Eurosurveillance. 2017 Mar 30;22(13):30494.

39. Parums D.V. Editorial: Infectious Disease Surveillance Using Artificial Intelligence (AI) and its Role in Epidemic and Pandemic Preparedness. Med Sci Monit Int Med J Exp Clin Res. 2023 Jun 1; 29: e941209.

40. Harishbhai Tilala M., Kumar Chenchala P., Choppadandi A., Kaur J., Naguri S., Saoji R., *et al*. Ethical Considerations in the Use of Artificial Intelligence and Machine Learning in Health Care: A Comprehensive Review. Cureus. 16(6): e62443.

41. Szklarczyk D., Gable A.L., Lyon D., Junge A., Wyder S., Huerta-Cepas J., *et al*. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019 Jan 8; 47(D1): D607–13.

42. Xian L., Wang Y. Advances in Computational Methods for Protein–Protein Interaction Prediction. Electronics. 2024 Jan; 13(6): 1059.

43. Soleymani F., Paquet E., Viktor H., Michalowski W., Spinello D. Protein–protein interaction prediction with deep learning: A comprehensive review. Comput Struct Biotechnol J. 2022 Jan 1; 20: 5316–41.

44. Li Y., Wang Z., You Z.H., Li L.P., Hu X. Predicting Protein-Protein Interactions via Random Ferns with Evolutionary Matrix Representation. Comput Math Methods Med. 2022 Feb 22; 2022: 7191684.

45. Deep learning methods for protein function prediction - Boadu - 2025 - PROTEOMICS - Wiley Online Library [Internet]. [cited 2025 Jan 20]. Available from: https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/pmic.202300471?af=R

46. Canzler S., Fischer M., Ulbricht D., Ristic N., Hildebrand P.W., Staritzbichler R. ProteinPrompt: a webserver for predicting protein-protein interactions. Bioinforma Adv. 2022; 2(1): vbac059.

47. The use of artificial intelligence for automating or semi-automating biomedical literature analyses: A scoping review - ScienceDirect [Internet]. [cited 2025 Jan 20]. Available from: https://www.sciencedirect.com/science/article/pii/S1532046423001107?via%3Dihub

48. Kiouri D.P., Batsis G.C., Chasapis C.T. Structure-Based Approaches for Protein–Protein Interaction Prediction Using Machine Learning and Deep Learning. Biomolecules. 2025 Jan; 15(1): 141.

49. Espina-Romero L., Gutiérrez Hurtado H., Ríos Parra D., Vilchez Pirela R.A., Talavera-Aguirre R., Ochoa-Díaz A. Challenges and Opportunities in the Implementation of AI in Manufacturing: A Bibliometric Analysis. Sci. 2024 Dec; 6(4): 60.

50. Kewalramani N., Emili A., Crovella M. State-of-the-art computational methods to predict protein–protein interactions with high accuracy and coverage. PROTEOMICS. 2023; 23(21–22): 2200292.

51. Casadio R., Martelli P.L., Savojardo C. Machine learning solutions for predicting protein–protein interactions. WIREs Comput Mol Sci. 2022; 12(6): e1618.

52. Zhang Z., Lamson A.R., Shelley M., Troyanskaya O. Interpretable neural architecture search and transfer learning for understanding CRISPR–Cas9 off-target enzymatic reactions. Nat Comput Sci. 2023 Dec; 3(12): 1056–66.

53. Özden F., Minary P. Learning to quantify uncertainty in off-target activity for CRISPR guide RNAs. Nucleic Acids Res. 2024 Oct 14; 52(18): e87.

54. Chen Q., Chuai G., Zhang H., Tang J., Duan L., Guan H., *et al.* Genome-wide CRISPR off-target prediction and optimization using RNA-DNA interaction fingerprints. Nat Commun. 2023 Nov 18; 14(1): 7521.

55. Lee M. Deep learning in CRISPR-Cas systems: a review of recent studies. Front Bioeng Biotechnol [Internet]. 2023 Jul 3 [cited 2025 Jan 20]; 11. Available from: https://www.frontiersin.org/journals/bioengineering-and-biotechnology/articles/10.3389/fbioe.2023.1226182/full

56. Dixit S., Kumar A., Srinivasan K., Vincent PMDR, Ramu Krishnan N. Advancing genome editing with artificial intelligence: opportunities, challenges, and future directions. Front Bioeng Biotechnol [Internet]. 2024 Jan 8 [cited 2025 Jan 20]; 11. Available from: https://www.frontiersin.org/journals/bioengineering-and-biotechnology/articles/10.3389/fbioe.2023.1335901/full

57. Sari O., Liu Z., Pan Y., Shao X. Predicting CRISPR-Cas9 off-target effects in human primary cells using bidirectional LSTM with BERT embedding. Bioinforma Adv. 2025; 5(1): vbae184.

58. Khoshandam M., Soltaninejad H., Mousazadeh M., Hamidieh A.A., Hosseinkhani S. Clinical applications of the CRISPR/Cas9 genome-editing system: Delivery options and challenges in precision medicine. Genes Dis. 2024 Jan 1; 11(1): 268–82.

59. O'Neill J. Tackling drug-resistant infections globally: final report and recommendations [Internet]. Government of the United Kingdom; 2016 May [cited 2025 Jan 20]. Available from: https://apo.org.au/node/63983

60. Popa S.L., Pop C., Dita M.O., Brata V.D., Bolchis R., Czako Z., *et al*. Deep Learning and Antibiotic Resistance. Antibiotics. 2022 Nov 21; 11(11): 1674.

61. Olatunji I., Bardaji D.K.R., Miranda R.R., Savka M.A., Hudson A.O. Artificial intelligence tools for the identification of antibiotic resistance genes. Front Microbiol. 2024 Jul 12; 15: 1437602.

62. Kim J.I., Maguire F., Tsang K.K., Gouliouris

T, Peacock S.J., McAllister TA, *et al*. Machine Learning for Antimicrobial Resistance Prediction: Current Practice, Limitations, and Clinical Perspective. Clin Microbiol Rev. 35(3): e00179-21.

63. Li Y., Cui X., Yang X., Liu G., Zhang J. Artificial intelligence in predicting pathogenic microorganisms' antimicrobial resistance: challenges, progress, and prospects. Front Cell Infect Microbiol. 2024 Nov 1; 14: 1482186.

64. Weaver D.T., King E.S., Maltas J., Scott J.G. Reinforcement learning informs optimal treatment strategies to limit antibiotic resistance. Proc Natl Acad Sci U S A. 121(16): e2303165121.

65. Guo W., Sun F., Liu F., Cao L., Yang J., Chen Y. Antimicrobial resistance surveillance and prediction of Gram-negative bacteria based on antimicrobial consumption in a hospital setting: A 15-year retrospective study. Medicine (Baltimore). 2019 Sep; 98(37): e17157.

66. Ryu B., Jeon W., Kim D. Integrating genomic and molecular data to predict antimicrobial minimum inhibitory concentration in Klebsiella pneumoniae. Sci Rep. 2024 Oct 29; 14(1): 25951.

67. Behling A.H., Wilson B.C., Ho D., Virta M., O'Sullivan J.M., Vatanen T. Addressing antibiotic resistance: computational answers to a biological problem? Curr Opin Microbiol. 2023 Aug 1; 74: 102305.

68. Lv J., Deng S., Zhang L. A review of artificial intelligence applications for antimicrobial resistance. Biosaf Health. 2021 Feb 1; 3(1): 22–31.

69. Sakagianni A., Koufopoulou C., Feretzakis G., Kalles D., Verykios V.S., Myrianthefs P., *et al*. Using Machine Learning to Predict Antimicrobial Resistance—A Literature Review. Antibiotics. 2023 Feb 24; 12(3): 452.

70. Sakagianni A., Koufopoulou C., Feretzakis G, Kalles D., Verykios V.S., Myrianthefs P., *et al.* Using Machine Learning to Predict Antimicrobial Resistance—A Literature Review. Antibiotics. 2023 Feb 24; 12(3): 452.