International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Harnessing Machine Learning in Biotechnology for Advanced Natural Language Processing

Dr. Ranjith Gopalan[1], Rathi Ramakrishnan[2]

[1]PhD, Principal Consultant, Ranjith.gopalan@gmail.com
[2]MPhil, MSc, Lecturer, arathi.ramakshnan@gmail.com

Abstract
This paper discusses machine learning techniques in biotechnology for analyzing protein structure and extracting features based on Natural language processing techniques in Artificial intelligence. Modern biotechnology is providing breakthroughs in products and technologies to address debilitating and rare diseases, thereby contributing to global healing. The convergence of artificial intelligence and biotechnology has emerged as a transformative force in the life sciences, enabling researchers to leverage vast amounts of data and derive meaningful insights. In the context of environmental biotechnology, real-time data processing facilitated by natural language processing is becoming increasingly crucial.
The paper explains details about Protein structure is a fascinating and complex topic, essential for understanding how proteins function and different AI techniques for extracting key features associated with specific protein structures. This paper also talks about ethical Implications of AI in Biotechnology Data Privacy and Security Issues and Addressing Bias in Machine Learning Models

Keywords: Artificial intelligence, Natural language Processing, Biotechnology, Protein structure, feature engineering, Deep learning

Introduction

The intersection of artificial intelligence (AI) and biotechnology has emerged as a transformative force in the life sciences, enabling researchers to harness vast amounts of data and derive meaningful insights. This synergy is particularly evident in the realm of natural language processing (NLP), where machine learning algorithms are utilized to analyze and interpret complex biological texts.
Artificial intelligence (AI) is increasingly used for the analysis of documents, extraction of information and insights from text, automatic creation of new text etc. Natural language processing (NLP) systems process word embeddings, i.e. words previously converted into numerical vector representations. Efficient models employ contextual embedding with self-attention mechanisms, whereby words of input text are arranged in one-dimensional sequences. Language models play a central role in recent advances in AI. Presently they are also used for solving problems in bioinformatics like function analysis of proteins, generative protein design c. In the language of proteins, amino acid sequences correspond to words and the representation of proteins as numerical vectors allows for the application of ML models and natural language processing techniques.
The integration of artificial intelligence into biotechnology is reshaping the landscape of research and innovation. By leveraging machine learning for advanced natural language processing, automated literature reviews, protein structure prediction, and real-time data processing, researchers can significantly

enhance their productivity and the quality of their work. As these technologies continue to evolve, they promise to unlock new avenues for exploration and discovery, ultimately driving forward the field of biotechnology and its applications in health, environment, and beyond.

Methodology

To recognize key features associated with specific protein structures is very imp in biotechnology and health. State-of-the-art natural language processing models like the Transformer architecture can be applied to protein sequence data, treating amino acid sequences like words in natural language. These models can learn representations of proteins that capture meaningful structural and functional properties, enabling tasks like protein classification, function prediction, and de novo protein design.

Here paper explains Protein structure and AL techniques for analyzing key features associated with specific protein structures.

Protein structure

Protein structure is a fascinating and complex topic, essential for understanding how proteins function. Proteins play critical roles in biological processes, and understanding their structure is essential for drug development and disease understanding. Proteins are biological polymers made up of amino acids, and their structure can be described at four different levels:

Primary Structure: This is the linear sequence of amino acids in a polypeptide chain. The order of amino acids is determined by the gene encoding the protein. Even a small change in this sequence can significantly affect the protein's function

Secondary Structure: This level refers to the local folding of the polypeptide chain into structures such as alpha-helices and beta-pleated sheets. These structures are stabilized by hydrogen bonds between the backbone atoms.

Tertiary Structure: This is the overall three-dimensional shape of a single polypeptide chain. It is formed by the folding of secondary structures into a compact, globular shape. Various interactions, including hydrogen bonds, ionic bonds, and disulfide bridges, stabilize the tertiary structure.

Quaternary Structure: Some proteins consist of multiple polypeptide chains, known as subunits. The quaternary structure describes how these subunits are arranged and interact with each other. Hemoglobin, for example, has a quaternary structure composed of four subunits

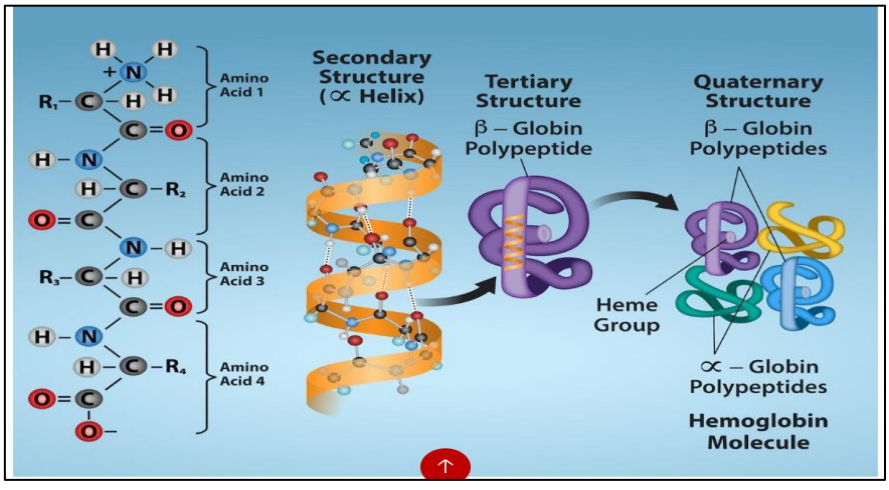

Fig1 – This diagram explains protein structure.

Textual Data Sources for Protein Information

Understanding the types of textual data available is essential. Textual data sources are invaluable for advancing our understanding of proteins and their functions. Here are some key sources and their roles in biotechnology:

Protein Databases: Databases like UniProt and Protein Data Bank (PDB) provide comprehensive information on protein sequences and structures. These databases are essential for training machine learning models and validating predictions. Protein databases are essential resources for storing and accessing information about proteins. Here are some of the most prominent ones:

RCSB Protein Data Bank (PDB): This database provides access to experimentally-determined 3D structures of proteins and nucleic acids. It also includes computed structure models from resources like AlphaFold DB

UniProt: UniProt is a comprehensive resource for protein sequence and functional information. It includes reviewed (Swiss-Prot) and unreviewed (TrEMBL) entries, offering detailed annotations on protein sequences, structures, and functions

Protein Information Resource (PIR): PIR offers integrated databases and tools for protein sequence analysis and functional annotation. It supports various bioinformatics applications, including protein identification and characterization.

Pfam: Pfam is a database of protein families, each represented by multiple sequence alignments and hidden Markov models. It helps in identifying and annotating protein domains and families.

InterPro: InterPro integrates diverse protein signature databases, providing functional analysis of protein sequences by classifying them into families and predicting domains and important sites.

Natural language Processing techniques and models

Natural Language Processing (NLP) can be a powerful tool for extracting key features associated with specific protein structures. Please see the ways NLP is applied in this field:

1. Text Mining and Information Extraction:

NLP techniques can be used to mine scientific literature and databases to extract relevant information about protein structures. Semantic similarity and parse tree analysis can help identify and extract residues and structural features

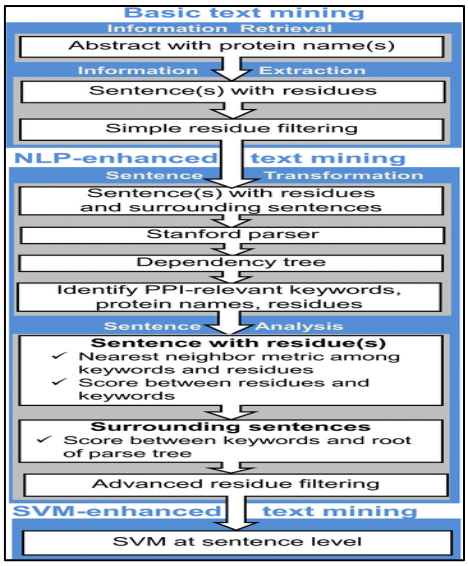

Fig2 — Flowchart of NLP-enhanced text mining system.

2. Protein Language Models (pLMs):

Advanced natural language processing models, such as transformer-based models like ProtT5, can be used to predict protein structures. By treating protein sequences as analogous to words in natural language, these models can learn rich representations of protein structure and function. This enables them to perform tasks like protein classification, structure prediction, and de novo design Protein language models decode the "language of life" by learning to predict masked amino acids based solely on the context provided by the amino acid sequences of millions of proteins. In this way, NLP words/tokens correspond to amino acids, and full-length proteins correspond to sentences. The embeddings extracted by these models capture the information they have learned, allowing them to demonstrate remarkable capabilities in tasks like protein classification, structure prediction, and de novo design, similar to natural language models.

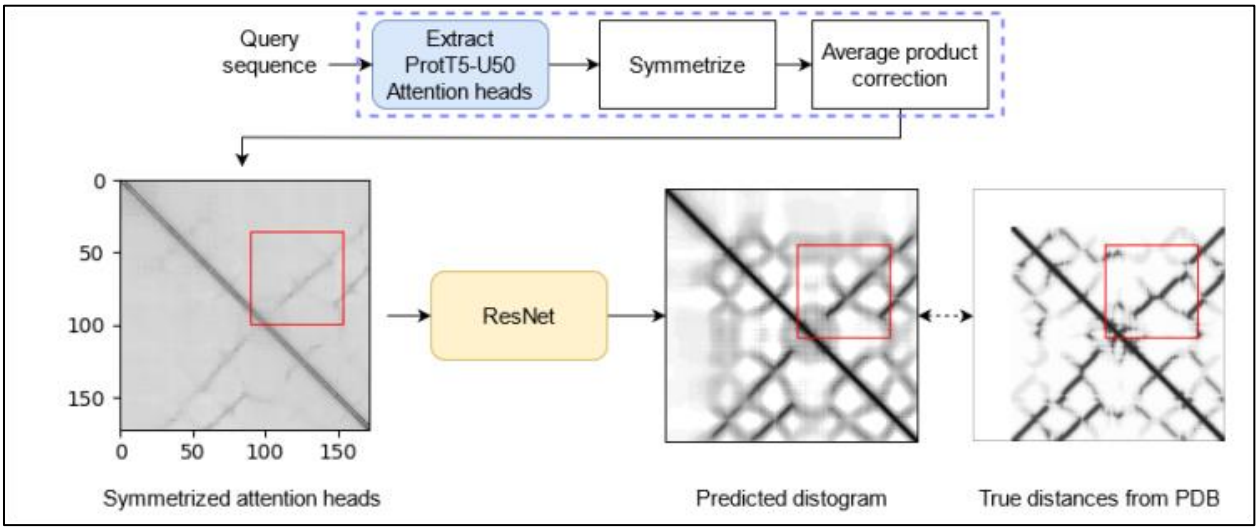

Fig 3 shows Sketch of the approach. The residual CNN (ResNet, yellow orange) is similar across models. Models using 1D protein embeddings from ProtT5 ,ProtBERT, ProtAlbert, or SeqVec adapted their architecture to account for expansion from 1D to 2D (Methods). The red square illustrates that proteins were split into overlapping crops of size 64x64 (64 consecutive residues) as introduced by AlphaFold 1. For each crop the CNN (ResNet) predicted the corresponding patch of the distogram. The example above shows the average over the attention heads after applying symmetry and APC for T1026 (CASP14), suggesting that ProtT5 already learned aspects of protein structure without supervision. Our baseline model trained on information from co-evolution for comparison replaced the modules marked by a dashed blue line by the generation of MSAs and estimation of parameters for the Potts model.

Input: The main input features used in the model were either protein representations derived from our pre-trained language models (pLMs) or, for comparison, the co-evolution signal in the form of Potts model parameters. Additionally, we included normalized residue positions, normalized protein length, and the long-normalized number of effective sequences as input channels. We also masked residues that were not resolved experimentally, both as single amino acid input and as residue pairs during the loss computation.

Training. The model was trained using the Adamax optimizer, with an initial learning rate of 0.01 and a batch size of 75. Early stopping was performed, and the best model checkpoint was saved when the MCC on the validation set did not improve for 10 consecutive iterations.

3. Representation Learning:

NLP techniques can transform the continuous, complex space of protein structures into discrete representations. In the context of protein structures involves transforming the complex, continuous space

of these structures into discrete representations. This is achieved using **vector-quantized autoencoders**. Here's a breakdown of the process:

1. **Vector-Quantized Autoencoders (VQ-VAE)**: These are a type of neural network that compresses data into a discrete latent space. For protein structures, this means converting the continuous 3D coordinates of proteins into a set of discrete tokens.
2. **Tokenization**: By breaking down the protein structures into tokens, it becomes easier to analyze and generate new structures. This is similar to how words in a sentence are tokenized in natural language processing (NLP).

This approach uses vector-quantized autoencoders to tokenize protein structures, making it easier to analyze and generate new protein structures

4. FEGS (Feature Extraction based on Graphical and Statistical features) is a model that leverages NLP techniques to analyze protein sequences. Here's how it works:

1. **Graphical Representation**: FEGS uses the physicochemical properties of amino acids to create a graphical representation of protein sequences. This involves mapping out the interactions and properties of amino acids in a way that can be visually and computationally analyzed.
2. **Statistical Features Extraction**: From these graphical representations, FEGS extracts statistical features. This means converting the complex information about protein sequences into numerical vectors. For instance, FEGS can transform a protein sequence into a 578-dimensional vector, capturing various statistical properties.

By leveraging these NLP techniques, researchers can gain deeper insights into protein structures, predict new structures, and even design novel proteins.

Models like Alpha Fold have demonstrated remarkable accuracy in predicting protein structures, which has significant implications for drug discovery and understanding diseases at a molecular level. These models can process vast amounts of textual data, including scientific literature and protein sequences, to generate accurate structural predictions. Moreover, the accessibility of these technologies allows research students and early-career scientists to engage with complex computational methods, fostering a deeper understanding of molecular biology. As these machine learning techniques continue to evolve, we can expect even more groundbreaking discoveries in biotechnology and medicine. By integrating textual data from scientific literature with structural prediction algorithms, researchers can accelerate the identification of potential drug targets and enhance the development of therapeutic interventions.

Results

The synergy between ML and NLP is particularly promising, as it allows for the extraction of valuable information from vast textual datasets, enriching the landscape of biotechnology research. As machine learning continues to evolve, its applications in structure prediction are expected to expand, paving the way for breakthroughs in our understanding of biological systems. Highly accurate protein structure prediction with Alpha Fold has demonstrated the ability to predict protein structures with atomic accuracy, even in cases where no similar structure is known. This achievement has significant implications for drug discovery, as it enables researchers to better understand the molecular mechanisms underlying diseases and identify potential therapeutic targets.

The transfer of protein models to natural language processing (NLP) can indeed bring about significant advancements. Here are some key points to consider:

Transfer Learning: Leveraging pre-trained models from one domain (e.g., protein sequences) to another

(e.g., NLP) can enhance performance and reduce training time. This approach has been successfully applied in models like ProLLaMA, which adapts large language models (LLMs) for protein language processing

Model Efficiency: Reducing the size of language models is crucial for sustainability. Techniques like model pruning, quantization, and knowledge distillation can help create smaller, more efficient models without significantly compromising performance

Sustainability: By optimizing models to be more efficient, we can reduce their carbon footprint. This is particularly important given the high computational costs associated with training large models

Multi-Task Learning: Models like ProLLaMA demonstrate the potential of handling multiple tasks in protein language processing, such as protein sequence generation and property prediction, which can be extended to NLP tasks

Innovative Applications: Applying protein models to NLP can lead to innovative applications, such as improved text mining for scientific literature, better understanding of protein-related patents, and enhanced clinical trial matching using NLP techniques

Ethical Considerations and Challenges

Here paper talks about ethical considerations and challenges while cconsidring AI technologies for improving productivity in biotechnology research area especially protein structure analysis

1. Ethical Implications of AI in Biotechnology

The integration of artificial intelligence and biotechnology raises ethical concerns that researchers must address. As machine learning becomes more prevalent in biotech research, such as in automated literature reviews, protein structure prediction, and real-time environmental data processing, questions about the ethical use of these technologies are crucial. This paper examines the primary ethical implications of deploying AI in biotechnology, emphasizing the need for responsible research practices.

One key ethical concern is data privacy and security. Biotechnology research often relies on large datasets that may contain sensitive information, like genetic data. The application of machine learning to analyze this data must be accompanied by robust measures to protect privacy. Ethical frameworks should ensure researchers are transparent about how data is collected, used, and shared, particularly in the context of automated literature review systems, where AI could inadvertently expose private information if not properly managed.

Another key ethical concern with using AI in biotechnology is the risk of algorithmic bias. The reliability of machine learning models depends on the data used to train them, and biased datasets can lead to skewed research outcomes. In biotechnology, this could result in certain populations being underrepresented in genetic studies. To mitigate this, researchers must carefully curate diverse and representative datasets, ensuring their findings are applicable across different demographic groups. This responsibility also extends to the development of machine learning models for protein structure prediction, where biased data could hinder scientific progress.

Furthermore, the potential for AI to replace human expertise raises ethical questions about the role of researchers in biotechnology. While AI can enhance efficiency and accuracy in tasks like real-time data processing, it should not be seen as a replacement for human judgment and intuition. Overreliance on AI tools in research could lead to overconfidence in automated outputs, potentially resulting in a lack of critical analysis. It is essential for researchers to maintain a balance, using AI as a supportive tool rather than a decision-maker, to preserve the integrity of the scientific process.

2. Data Privacy and Security Issues

Protecting data privacy and security is crucial when combining biotechnology and artificial intelligence, especially when using machine learning for advanced natural language processing. As students explore these innovative fields, they face a complex landscape involving sensitive biological data, proprietary research, and ethical concerns. While AI integration in biotech offers transformative potential, it raises significant issues around data collection, processing, and storage. As machine learning models rely increasingly on large datasets, understanding the implications of data privacy and security is paramount.

One primary concern in biotechnological research is the handling of personal and sensitive genetic information. When using NLP for automated literature reviews or extracting insights from scientific texts, researchers often encounter datasets containing individuals' identifiable data. The risk of data breaches or unauthorized access to such information can lead to severe consequences, including ethical violations and legal repercussions. Therefore, research students must be well-versed in regulations like the Health Insurance Portability and Accountability Act and the General Data Protection Regulation, which govern the use of personal health information and data privacy standards.

3. Addressing Bias in Machine Learning Models

Addressing bias in machine learning models is a critical aspect of developing reliable and ethical AI systems, especially in fields like biotechnology where decisions can have significant implications. Bias can stem from various sources, including the data used for training models, the algorithms themselves, and the interpretations made by researchers. In the context of advanced natural language processing (NLP) applications, bias can lead to skewed results that may misrepresent biological phenomena or overlook important findings within the vast body of scientific literature. Therefore, it is essential for research students to understand the origins of bias and the strategies available for its mitigation.

Future Trends and Directions

Here paper discuss about advancement of AI in biotechnology As artificial intelligence and machine learning continue to evolve, researchers are exploring innovative ways to integrate these technologies into various aspects of biotechnology to drive scientific discovery and unlock new possibilities.

1. Emerging Technologies in Biotech and AI

The convergence of biotechnology and artificial intelligence has given rise to a revolutionary landscape in research and development. As these fields continue to evolve, innovative applications leveraging machine learning are becoming increasingly prominent, particularly in enhancing natural language processing capabilities. This section will explore the latest advancements in these domains, focusing on how they contribute to more efficient and effective research methodologies within the biotechnology field.

The intersection of biotechnology and AI is ushering in a new era of research capabilities that were previously unimaginable. Automated literature review systems, advanced protein structure prediction models, and real-time data processing technologies exemplify how machine learning can be leveraged to enhance the efficiency and effectiveness of biotechnology research.

2. The Future of NLP in Biotech Research

Natural language processing has profound applications in the realm of biotechnology, particularly in the realm of protein structure prediction. The complex relationship between a protein's sequence and its three-dimensional structure holds significant implications for drug design and the understanding of disease mechanisms. By leveraging NLP techniques to analyze existing research data, machine learning models can be developed that predict protein structures with remarkable accuracy. These predictive models

integrate information from diverse sources, including genomic data and scientific literature, enabling more holistic and insightful predictions. As a result, researchers can expedite the discovery of new therapeutic targets and enhance the design of biology.

Moreover, NLP has the potential to significantly impact environmental biotechnology research through real-time data processing. The ability to analyze and interpret data from various environmental sources, such as sensor readings, field reports, and scientific publications, can facilitate timely interventions and solutions for pressing environmental challenges.

Conclusion

In summary, the paper explores the application of natural language processing techniques in biotechnology, particularly in analyzing protein structures for drug discovery and enabling real-time data processing for environmental biotechnology research. It delves into the various NLP models and methods, such as Protein Language Models, Text Mining, Representation Learning, and FEGS, which are used for fast, accurate, and alignment-free protein structure prediction. These new solutions can generate protein-specific predictions, allowing researchers to distinguish between structural features that differentiate members of the same protein family, which were previously indistinguishable using other top methods.

Moreover, the paper emphasizes the importance of addressing data privacy concerns and mitigating bias in machine learning models to ensure the development of ethical and reliable AI systems in biotechnology research. The future of NLP in biotechnology research is described as holding immense promise, with the continued evolution of these technologies poised to profoundly impact our understanding of biological systems and drive innovative solutions to global challenges.

Reference

1. Ao, C., Xiao, Z., Guan, L., & Yu, L. (2023, January 1). Computational Approaches for Predicting Drug-Disease Associations: A Comprehensive Review. Cornell University. https://doi.org/10.48550/arxiv.2309.06388
2. BaHammam, A S., Trabelsi, K., Pandi-Perumal, S R., & Jahrami, H. (2023, January 1). Adapting to the Impact of AI in Scientific Writing: Balancing Benefits and Drawbacks while Developing Policies and Regulations. Cornell University. https://doi.org/10.48550/arXiv.2306.
3. Bahja, M. (2021, May 19). Natural Language Processing Applications in Business. IntechOpen. https://doi.org/10.5772/intechopen.92203
4. Bala, R. (2022, August 18). Challenges and Ethical Issues in Data Privacy. IGI Global, 12(2), 1-7. https://doi.org/10.4018/ijirr.299938
5. Cabrero-Daniel, B., & Cabrero, Á. (2023, July 11). Perceived Trustworthiness of Natural Language Generators. https://doi.org/10.1145/3597512.3599715
6. Calhoun, B., Kiel, J M., & Morgan, A A. (2018, August 20). Health Insurance Portability and Accountability Act Violations by Physician Assistant Students: Applying Laws to Clinical Vignettes. Lippincott Williams & Wilkins, 29(3), 154-157. https://doi.org/10.1097/jpa.0000000000000215
7. Checco, A., Bracciale, L., Loreti, P., Pinfield, S., & Bianchi, G. (2021, January 25). AI-assisted peer review. Palgrave Macmillan, 8(1). https://doi.org/10.1057/s41599-020-00703-8
8. Choi, B., Dayaram, T., Parikh, N., Wilkins, A D., Nagarajan, M., Novikov, I B., Bachman, B J., Jung, S Y., Haas, P J., Labrie, J L., Pickering, C R., Adikesavan, A K., Regenbogen, S., Kato, L., Lelescu, A., Buchovecky, C M., Zhang, H., Bao, S., Boyer, S., Lichtarge, O. (2018, September 28). Literature-

based automated discovery of tumor suppressor p53 phosphorylation and inhibition by NEK2. National Academy of Sciences, 115(42), 10666-10671. https://doi.org/10.1073/pnas.1806643115

9. Clegg, L E., & Gabhann, F M. (2015, September 1). Molecular mechanism matters: Benefits of mechanistic computational models for drug development. Elsevier BV, 99, 149-154. https://doi.org/10.1016/j.phrs.2015.06.002
10. Filipp, F V. (2019, December 1). Opportunities for Artificial Intelligence in Advancing Precision Medicine. Springer Science+Business Media, 7(4), 208-213. https://doi.org/10.1007/s40142-019-00177-4
11. Hsieh, Y L., Chang, Y., Chang, N., & Hsu, W. (2017, November 1). Identifying Protein-protein Interactions in Biomedical Literature using Recurrent Neural Networks with Long Short-Term Memory., 2, 240-245. https://www.aclweb.org/anthology/I17-2041.pdf
12. Hu, Z., Yu, Q., Guo, Y., Wang, T., King, I., Gao, X., Song, L., & Li, Y. (2023, January 1). Drug Synergistic Combinations Predictions via Large-Scale Pre-Training and Graph Structure Learning. Cornell University. https://doi.org/10.48550/arxiv.2301.05931
13. Jacobsen, A., Dijk, E V., Mouhib, H., Stringer, B., Ivanova, O., Gavaldá-García, J., Hoekstra, L S., Feenstra, K A., & Abeln, S. (2023, January 1). Introduction to Protein Structure. Cornell University. https://doi.org/10.48550/arxiv.2307.02169
14. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S., Ballard, A J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., . . . Hassabis, D. (2021, July 15). Highly accurate protein structure prediction with AlphaFold. Nature Portfolio, 596(7873), 583-589. https://doi.org/10.1038/s41586-021-03819-2
15. Kalkman, S., Mostert, M., Gerlinger, C., Delden, J M V., & Thiel, G J M W V. (2019, March 28). Responsible data sharing in international health research: a systematic review of principles and norms. BioMed Central, 20(1). https://doi.org/10.1186/s12910-019-0359-9
16. Kang, H., Hou, L., Gu, Y., Lu, X T., Li, J., & Li, Q. (2023, May 22). Drug–disease association prediction with literature based multi-feature fusion. Frontiers Media, 14. https://doi.org/10.3389/fphar.2023.1205144
17. Khare, R., Leaman, R., & Lu, Z. (2014, January 1). Accessing Biomedical Literature in the Current Information Landscape. Springer Science+Business Media, 11-31. https://doi.org/10.1007/978-1-4939-0709-0_2
18. Krishna, R., Kelleher, K J., & Stahlberg, E. (2007, April 1). Patient Confidentiality in the Research Use of Clinical Medical Databases. American Public Health Association, 97(4), 654-658. https://doi.org/10.2105/ajph.2006.090902
19. Li, X., Lin, X., Cao, D., Zeng, X., Yu, P S., He, L., Nussinov, R., & Cheng, F. (2022, January 1). Deep learning for drug repurposing: methods, databases, and applications. Cornell University. https://doi.org/10.48550/arxiv.2202.05145
20. Mikołajczyk-Bareła, A., & Grochowski, M. (2023, January 1). A survey on bias in machine learning research. Cornell University. https://doi.org/10.48550/arXiv.2308.
21. Nagasundaram, N., Yapp, E K Y., Le, N Q K., Kamaraj, B., Al-Subaie, A M., & Yeh, H. (2019, November 11). Application of Computational Biology and Artificial Intelligence Technologies in Cancer Precision Drug Discovery. Hindawi Publishing Corporation, 2019, 1-15. https://doi.org/10.1155/2019/8427042

22. Narganes-Carlón, D., Crowther, D., & Pearson, E R. (2023, May 24). A publication-wide association study (PWAS), historical language models to prioritise novel therapeutic drug targets. Nature Portfolio, 13(1). https://doi.org/10.1038/s41598-023-35597-4
23. Ofer, D., Brandes, N., & Linial, M. (2020, December 31). The language of proteins: NLP, machine learning & protein sequences. https://www.sciencedirect.com/science/article/pii/S2001037021000945
24. Petkovic, D. (2022, January 1). It is not "accuracy vs. explainability" -- we need both for trustworthy AI systems. Cornell University. https://doi.org/10.48550/arXiv.2212.
25. Pollock, N W. (2020, March 1). Managing Bias in Research. Elsevier BV, 31(1), 1-2. https://doi.org/10.1016/j.wem.2020.01.001
26. Qureshi, R., Irfan, M., Gondal, T M., Khan, S., Wu, J., Hadi, M U., Heymach, J V., Le, X., Yan, H., & Alam, T. (2023, July 1). AI in drug discovery and its clinical relevance. Elsevier BV, 9(7), e17575-e17575. https://doi.org/10.1016/j.heliyon.2023.e17575
27. Rosenblatt, M., Boutin, M., & Nussbaum, S R. (2016, October 25). Innovation in Medicine and Device Development, Regulatory Review, and Use of Clinical Advances. American Medical Association, 316(16), 1671-1671. https://doi.org/10.1001/jama.2016.12486
28. Rumbold, J., & Pierścionek, B. (2017, April 8). A critique of the regulation of data science in healthcare research in the European Union. BioMed Central, 18(1). https://doi.org/10.1186/s12910-017-0184-y
29. Schwartz, R., Down, L., Jonas, A., & Tabassi, E. (2021, June 22). A Proposal for Identifying and Managing Bias in Artificial Intelligence. https://doi.org/10.6028/nist.sp.1270-draft
30. Seebode, C., Ort, M., Regenbrecht, C., & Peuker, M. (2013, October 1). BIG DATA infrastructures for pharmaceutical research. https://doi.org/10.1109/bigdata.2013.6691759
31. Sica, G T. (2006, March 1). Bias in Research Studies. Radiological Society of North America, 238(3), 780-789. https://doi.org/10.1148/radiol.2383041109
32. Staunton, C., Slokenberga, S., & Mascalzoni, D. (2019, April 17). The GDPR and the research exemption: considerations on the necessary safeguards for research biobanks. Springer Nature, 27(8), 1159-1167. https://doi.org/10.1038/s41431-019-0386-5
33. Taha, K., & Yoo, P D. (2015, August 1). An information extraction system for protein function prediction. https://doi.org/10.1109/cibcb.2015.7300300
34. Tomlinson, B., Torrance, A W., & Black, R W. (2023, January 1). ChatGPT and Works Scholarly: Best Practices and Legal Pitfalls in Writing with AI. Cornell University. https://doi.org/10.48550/arXiv.2305.
35. Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zieliński, M., Žídek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., Velankar, S., Kleywegt, G J., Bateman, A., Evans, R., Pritzel, A., Figurnov, M., Ronneberger, O., Bates, R., Kohl, S., . . . Hassabis, D. (2021, July 22). Highly accurate protein structure prediction for the human proteome. Nature Portfolio, 596(7873), 590-596. https://doi.org/10.1038/s41586-021-03828-1
36. Ye, C., Swiers, R., Bonner, S., & Barrett, I P. (2022, January 1). A Knowledge Graph-Enhanced Tensor Factorisation Model for Discovering Drug Targets. Institute of Electrical and Electronics Engineers, 1-11. https://doi.org/10.1109/tcbb.2022.3197320
37. Yue, T., & Wang, H. (2018, January 1). Deep Learning for Genomics: A Concise Overview. Cornell University. https://doi.org/10.48550/arXiv.1802.