Innovative Computing Review (ICR)
Volume 2 Issue 1, Spring 2022
ISSN$_{(P)}$: 2791-0024 ISSN$_{(E)}$: 2791-0032
Homepage: https://journals.umt.edu.pk/index.php/UMT-AIR

Article QR

Protein Structure Prediction with AlphaFold2, How it Works, Limitations and Solution for Less number of Homotypic and Large number of Heterotypic Contacts

Author (s): Muhammad Noman Khalid[1], Hassan Kaleem[2]

Affiliation (s): [1]Allama Iqbal Medical College Lahore,
[2]SQL Consultancy LTD 9 Frances Street, Crewe, England

DOI: https://doi.org/10.32350.icr.21.02

History: Received: April 10, 2022, Revised: May 25, 2022, Accepted: June 13, 2022

Citation: M. N. Khalid and H. Kaleem, "Protein Structure Prediction with AlphaFold2, How it Works, Limitations and Solution for Less number of Homotypic and Large number of Heterotypic Contacts," *UMT Artif. Intell. Rev.*, vol. 2, no. 1, pp. 00-00, 2022, doi: https://doi.org/10.32350.icr.21.02

Conflict of Interest: Author(s) declared no conflict of interest

A publication of
School of Systems and Technology
University of Management and Technology, Lahore, Pakistan

Protein Structure Prediction with AlphaFold2, How it Works, Limitations and Solution for Less number of Homotypic and Large number of Heterotypic Contacts

Muhammad Noman Khalid[*,1], Hassan Kaleem[2]

[1]Allama Iqbal Medical College Lahore, Pakistan
[2]SQL Consultency LTD 9 Frances Street, Crewe, England

***Abstract*-Knowing the protein structure helps us to investigate diseases in human beings related to abnormal or impaired folded proteins. This research provides a solution for how to identify the misbalance of homotypic and heterotypic contacts on the sequential stage. There are two methods of protein structure prediction, template based and Ab-initio models. Template based model matches the given sequence with the original sequence. Whereas, Ab-initio calculates the weight of the given sequence and identifies whether it is balanced or not. If the sequence is not in balance, it can be labeled as on the initial stage by calculating its weight. In this research, future directions to researchers are provided as how to achieve maximum accuracy in protein structure prediction.**

***Index Terms*-Ab-initio modeling, Alphafold2, hetrotypic, homoptypic, limitations, misbalance, protein structure prediction**

I.Introduction

Knowledge and prediction of protein structures help us to understand their working, that is, how these chemicals help human beings in their daily life. The word 'protein' comes from the Greek word 'proteios' which means 'highest importance'. Individual proteins are categorized based on their functions which describe the tasks they do. Protein structure recognition is used by the immune system, which is in charge of our body's defense. Knowing the protein structure helps us to know diseases in human beings related to abnormal or impaired, folded proteins. Proteins can be categorized on the basis of functions they perform. For example, structural proteins help to determine cell shape and integrity. These proteins also play a vital role in the mitosis and meiosis of the cell's reproduction and also in the immune system of our body (by

[*] Corresponding author: Noman.Khalid.1122.mnk@gmail.com

structural recognition of immunoglobulins). Thus, knowing and modifying these protein structures can revolutionize the medical field.

Linus Paul [1] first predicted the spiral structure of proteins in 1936. Afterwards, with the help of technological advancements in biology, scientists discovered 4 different levels of protein structure which are primary, secondary, tertiary, and quaternary levels. Primary structure comprises the sequence of amino acids in its polypeptide chain. Secondary structure constitutes polypeptide's backbone which is the main chain is its local spatial arrangement. Tertiary structure forms the three-dimensional structure of the entire chain of polypeptides. Lastly, quaternary structure comprises the three-dimensional arrangement of subnet in multi-subnet protein.

There are two methods used to determine protein structure including X-Ray Crystallography and Proton Nuclear Magnetic Resonance (PNMR). These methods help us to visualize the different layers of protein structure. However, the main problem is the cost of determining the protein structure which remains very high. According to the X-Ray Crystallography Facility (XRCF), the average cost of these tests is around 450$ [2] per sample. The Ab-initio [3] model was developed to predict the secondary structure of proteins. Two models were developed, namely PaleAle 4.0 with the accuracy of 80.0% and Porter 4.0 with the accuracy of 82.2%. Bidirectional Recurrent Neural Network (BRNN) [4] was used to predict the secondary structure. DeepCNF [5] was developed based on machine learning to predict the protein structure with the accuracy of 82.3%. DeepCNF is also used to predict the IDRs (intrinsically disordered regions) of proteins. However, with the training of AUC [6], the model achieved the accuracy of 84.5%. Spider 3 [7] was developed to predict the secondary structure of proteins by using Long Short Term Memory (LSTM)[8] and Bidirectional Recurrent Neural Network (BRNNs) [9] It achieved the accuracy of 83.9%. MUFOLD-SS [10] was developed to predict the secondary structure of proteins. It achieved the accuracy of 88.20% in easy cases and 83.37% in hard cases. Easy cases are those in which the hit value or e-value is <=0.5, while hard cases are those where hit value or e-value is >0.5. Ab-initio [11] model was developed with the updated version of Porter 4.0 model. Porter 5.0 achieved the accuracy of 84.19% in protein structure

prediction. SPOT-1D [12] predicted the protein structure with the accuracy of 86.18%. SPOT-1D uses Deep Neural Network (DNN) architecture based on recurrent and convolutional methods. NetSurfP-2.0 [13] was developed to predict the secondary structure of proteins from their primary sequence. All these models were developed and used to predict the one-dimensional secondary structure of proteins. The accuracy of these models is based on three (3) class labels in the current study. The tertiary and quaternary 3D structures were found to be problematic. The challenge is how to visualize the tertiary structure and then combine all the visualized forms to create the quaternary structure. Several methods have been developed to predict the structure of proteins. DeepMind [14] have developed a model to predict protein structure known as AlphaFold 1 [15] with CASP 13 [16]. AlphaFold 1 uses concurrent neural network architecture to predict protein structure, while AlphaFold 2 [17] with CASP 14 [18] uses the transformer. Transformer adopts the self-attention mechanism which takes sequential input data. Still, its prediction is very low in case of homotypic and heterotypic contacts. This research provides a solution for predicting misbalanced homotypic and heterotypic contacts.

II.Related Work

Many models have been developed to predict the protein structure covered in these reviews [19] [20] [21] [22] [23]. Despite applying neural network architecture for prediction [19] [22] [23], the improved structure prediction of protein [15] [24] [25] [26]. These approaches follow the improvement of computer vision systems [27]. They attempt to fold the tertiary structure of proteins to make the quaternary structure [28] [29] [30], which ultimately creates the 3D structure of proteins. Few models have been developed to predict the protein structure, directly [31] [32] [33] [34]. However, these approaches fail to match the previous structure prediction pipelines [35]. Still, the success of transformers, which are self-attention based model for language processing [36] and more recently, of computer vision based models [37] [38], has diverted the attention of researchers to adopt the self-attention based approaches [39] [40] [41].

III. Research Methodology

The goal of this research is to provide a review of existing problems and their solutions for

protein structure prediction. We followed the methodology of a survey that was designed by various researchers. The research objectives of this paper are as follows:

1. Workings of AlphaFold2
2. Limitations of AlphaFold2
3. Solutions for a small number of homotypic and a large number of heterotypic contacts
4. Feature extraction of protein

A. CASP (Critical Assessment of Protein Structure Prediction)

CASP is a community experiment conducted every two years since 1993 on a large scale. Experimentally determined information is passed on to the predicator of protein structure. When predictions are made, neither the predictor nor the organizer and accessor know about them. These predictions are then solved by X-Ray Crystallography and PNMR. Afterwards, these entries are kept in hold by PDB (Protein Data Bank).

IV. Results

A. How Alpha Fold 2 Works

AlphaFold2 [17] achieved the median score of 92.4 GDT (Global Distance Test). It indicates that even with the hardest protein targets, it can predict protein structure comparable to the width of an atom. The model was trained on CASP 14 with 170000 known protein structures. Although, for a problem like protein structure prediction, this is a very small number. They have taken a much larger dataset from unknown structures of protein sequences. They have learned to extract information from unlabeled data, for example, unsupervised learning which enables a lot of AI breakthroughs. GPT 3 [54] (Generative Pre-trained Transformer) was trained on a huge amount of data collected from the web. Then, it was given a slice of sentence and it had to predict which words were likely to come in the next sentence. In another example, a slice of an image was given to the model and the model was asked to predict the remaining part of the image.

B. Limitations of AlphaFold 2

AlphaFold2 [17] uses transformer, a deep learning model based on self-attention mechanism. However, this model slows down when the sequence size of protein is increased [55]. Another limitation highlighted by the AlphaFold2 [17] team is that it's prediction is much weaker for those proteins who have a small number of homotypic contacts.

Table I

Version_Human	Year of Publishing	No of Entries	No. of Sequences	Ref
CASP1	1994	229	1	[42]
CASP2	1996	212	2	[43]
CASP3	1998	235	2	[44]
CASP4	2000	203	1	[45]
CASP5	2002	191	3	[46]
CASP6	2004	217	2	[47]
CASP7	2006	217	1	[48]
CASP8	2008	246	1	[49]
CASP9	2010	229	3	[50]
CASP10	2012	221	3	[51]
CASP11	2014	117	1	[52]
CASP12	2016	140	3	[53]
CASP13	2018	150	1	[16]
CASP14	2020	190	2	[18]

Multiple experiments were conducted for proteins with a large number of heterotypic contacts [56] on a recent PDB dataset [57]. Homotypic contacts are defined by the attachment of one cell to another cell and these cells have to be identical. Whereas, in heterotypic contact protein's physical interaction have different primary structure. In protein 3D structure prediction, there is a sequence. Firstly, primary structure is predicted based on the numbers of amino acids in the polypeptide chain. Then, in the secondary structure, sequence from the primary structure is classified into different parts, namely Alpha Hilux, Beta strand, and random coil. Afterwards, in the tertiary structure, Alpha Hilux, Beta strand, and random coil are visualized separately. Finally, in the quaternary structure, all of these visualized forms are folded together to create a 3D structure of the protein.

C. How to Calculate Homotypic and Heterotypic Contacts

In advanced metastasis tumors, due to the lack of tumor suppressor genes different cells types (heterotypic) grow in between normal cells types, for example, if normal alignment consists of epithelial cells in epithelial cell types, then, in advance metastasis tumors there will be some other type of cells in epithelial cells, such as connective tissue cells [58].

In a template-based model, for example, we have a protein whose original sequence is:

Gly-Ala-Pro-Leu-Val-Met-Val-Pro-Ala-Cys-Gly-Ala-Pro-Leu-Val-Met-Val-Pro-Ala-Cys-Gly-Ala-Pro-Leu-Val-Met-Val-Pro-Ala-Cys-Gly-Ala-Pro-Leu-Val-Met-Val-Pro-Ala-Cys-Gly-Ala-Pro-Leu-Val-Met-Val-Pro-Ala-Cys

The sequence obtained from the user is:

Gly-Trp-Pro-Leu-Val-Met-Val-Pro-Ala-Cys-Gly-Ala-Pro-Leu-Val-Met-Val-Pro-Ala-Cys-Gly-Ala-Pro-Leu-Val-Met-Val-Pro-Ala-Cys-Gly-Ala-Pro-Leu-Val-Met-Val-Pro-Ala-Cys-Gly-Ala-Pro-Leu-Val-Met-Val-Pro-Ala-Cys

So, there is the original sequence and also the original weight of this sequence. The user sequence is matched with the original sequence. After matching the sequence with the original sequence, it was found that alanine is replaced with tryptophan in the user sequence. Tryptophan is the heaviest amino acid among all the essential amino acids. So, it automatically changes the weight of the user sequence, thus making the misbalance of homotypic and

heterotypic contacts in the user sequence easily identifiable. This problem can be solved by using machine learning algorithms. Ten machine learning algorithms were developed and implemented [59], each with K-Fold cross-validation testing. A feature extraction method for protein [60] was created on a live server. This app requires protein sequence in FASTA format and it automatically creates a csv file to be used in machine learning algorithms. The problem is that creating a dataset which has all the original sequences of protein is not possible. In case of human proteins all the genes have been identified. So, this approach can resolve the problem.

The second method is the Ab-initio model which is the computational matrix of quantum chemistry. In this model, the weight of protein sequence is calculated to identify if it is in balance or not.

This fact was revealed when the user sequence was matched with the protein's real sequence [61]. Weight calculation showed that the user sequence has a higher weight because it has tryptophan in the amino acid chain. While, the real sequence has alanine at the position of the R. Since tryptophan is the heaviest among all, thus the result can be assessed by matching it with the real sequences. After calculations, it was found that the real weight of the original amino acid is lower than the user's amino acid sequence. It is due to the fact that the weight of alanine is lower than tryptophan, which is present in the user's amino acid sequence. So, the weight of the amino acid chains can be calculated by matching them with their original sequences and the errors can be shown numerically.

D. Feature Extraction for Protien

For the extraction of protein features, the first step is to compute the matrix formation of the input protein query. The protein sequence length is used to build the following:

1. Position Relative Incidence Matrix (PRIM)
2. Reverse Position Relative Incidence Matrix (RPRIM)
3. Accumulative Absolute Position Incidence Vector (AAPIV)
4. Reverse Accumulative Absolute Position Incidence Vector (RAAPIV)
5. Frequency Vector (FV)

Live server is created for the feature extraction of proteins [60] based on Chou's 5-step rule [62]. The server accepts only FASTA format.

V. Discussion and Future Research Work

The current paper briefly discussed protein structure prediction keeping in view the previous research conducted in this field. The primary sequence was predicted based on the numbers of amino acids in the polypeptide chain. Secondary structure prediction included the identification of Alpha Helix, Beta strand, and random coil. Tertiary structure included the visualization of these classes and in the quaternary structure, these visualized forms were merged together to create the final 3D structure of the target protein. AlphaFold2 with CASP 14 achieved the maximum frequency of 92 GDT. Although, it was also found that there are limitations to AlphaFold2 algorithms. If there are a small number of homotypic and a large number of heterotypic contacts, their prediction GDT is very low. This issue can be resolved by homology based modeling and Ab-initio modeling. In homology based modeling, a protein feature extraction technique is developed on a live server to test and create 10 machine learning algorithms via K-Fold cross-validation testing. However, it was found that the best method to predict the protein structure is Ab-initio. A solution with one real-time example was proposed regarding how to calculate the weight of protein sequences and identify those not in balance. On the basis of this technique, protein folding and repairing technique can also be applied to achieve the maximum GDT in protein structure prediction. Since Ab-initio is based on calculating rather than predicting structure (based on CASP), it is the best method to predict the protein structure.

References

[1] "On the Structure of Native, Denatured, and Coagulated Proteins." https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1076802/ (accessed Jan. 09, 2022).

[2] "Fees," *X-Ray Crystallography Facility (XRCF)*. http://xrcf.caltech.edu/xrcf/fees (accessed Jan. 09, 2022).

[3] C. Mirabello and G. Pollastri, "Porter, PaleAle 4.0: high-accuracy prediction of protein secondary structure and relative

solvent accessibility," *Bioinformatics*, vol. 29, no. 16, pp. 2056–2058, Aug. 2013, doi: 10.1093/bioinformatics/btt344.

[4] P. Baldi, S. Brunak, P. Frasconi, G. Soda, and G. Pollastri, "Exploiting the past and the future in protein secondary structure prediction," *Bioinformatics*, vol. 15, no. 11, pp. 937–946, Nov. 1999, doi: 10.1093/bioinformatics/15.11.937.

[5] S. Wang, J. Ma, and J. Xu, "AUCpreD: proteome-level protein disorder prediction by AUC-maximized deep convolutional neural fields," *Bioinformatics*, vol. 32, no. 17, pp. i672–i679, Sep. 2016, doi: 10.1093/bioinformatics/btw446

[6] "AUC: a misleading measure of the performance of predictive distribution models - Lobo - 2008 - Global Ecology and Biogeography - Wiley Online Library." https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1466-8238.2007.00358.x (accessed Jan. 19, 2022).

[7] R. Heffernan, Y. Yang, K. Paliwal, and Y. Zhou, "Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility," *Bioinformatics*, vol. 33, no. 18, pp. 2842–2849, Sep. 2017, doi: 10.1093/bioinformatics/btx218.

[8] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[9] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997, doi: 10.1109/78.650093.

[10] C. Fang, Y. Shang, and D. Xu, "MUFOLD-SS: New deep inception-inside-inception networks for protein secondary structure prediction," *Proteins Struct. Funct. Bioinforma.*, vol. 86, no. 5, pp. 592–598, 2018, doi: 10.1002/prot.25487.

[11] M. Torrisi, M. Kaleel, and G. Pollastri, "Porter 5: fast, state-of-the-art ab initio prediction of protein secondary

structure in 3 and 8 classes," Oct. 2018. doi: 10.1101/289033.

[12] Hanson, Jack, "Protein Structure Prediction by Recurrent and Convolutional Deep Neural Network Architectures," Nov. 2018, doi: 10.25904/1912/3830.

[13] M. S. Klausen *et al.*, "NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning," *Proteins Struct. Funct. Bioinforma.*, vol. 87, no. 6, pp. 520–527, 2019, doi: 10.1002/prot.25674.

[14] "DeepMind - What if solving one problem could unlock solutions to thousands more?," *Deepmind*. https://deepmind.com/ (accessed Feb. 01, 2022).

[15] A. W. Senior *et al.*, "Improved protein structure prediction using potentials from deep learning," *Nature*, vol. 577, no. 7792, Art. no. 7792, Jan. 2020, doi: 10.1038/s41586-019-1923-7.

[16] "CASP 13." [Online]. Available: https://www.uniprot.org/uniprot/O75601

[17] "Highly accurate protein structure prediction with AlphaFold | Nature." https://www.nature.com/articles/s41586-021-03819-2 (accessed Jan. 09, 2022).

[18] "CASP 14." [Online]. Available: https://www.uniprot.org/uniprot/P31944

[19] R. Pearce and Y. Zhang, "Deep learning techniques have significantly impacted protein structure prediction and protein design," *Curr. Opin. Struct. Biol.*, vol. 68, pp. 194–207, Jun. 2021, doi: 10.1016/j.sbi.2021.01.007.

[20] "Advances in protein structure prediction and design | Nature Reviews Molecular Cell Biology." https://www.nature.com/articles/s41580-019-0163-x (accessed Feb. 02, 2022).

[21] D. S. Marks, T. A. Hopf, and C. Sander, "Protein structure prediction from sequence variation," *Nat. Biotechnol.*, vol. 30, no. 11, Art.

no. 11, Nov. 2012, doi: 10.1038/nbt.2419.

[22] N. Qian and T. J. Sejnowski, "Predicting the secondary structure of globular proteins using neural network models," *J. Mol. Biol.*, vol. 202, no. 4, pp. 865–884, Aug. 1988, doi: 10.1016/0022-2836(88)90564-5.

[23] "Prediction of contact maps with neural networks and correlated mutations | Protein Engineering, Design and Selection | Oxford Academic." https://academic.oup.com/peds/article/14/11/835/1608425?login=true (accessed Feb. 02, 2022).

[24] S. Wang, S. Sun, Z. Li, R. Zhang, and J. Xu, "Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model," *PLOS Comput. Biol.*, vol. 13, no. 1, p. e1005324, Jan. 2017, doi: 10.1371/journal.pcbi.1005324.

[25] J. Yang, I. Anishchenko, H. Park, Z. Peng, S. Ovchinnikov, and D. Baker, "Improved protein structure prediction using predicted interresidue orientations," *Proc. Natl. Acad. Sci.*, vol. 117, no. 3, pp. 1496–1503, Jan. 2020, doi: 10.1073/pnas.1914677117.

[26] Y. Li *et al.*, "Deducing high-accuracy protein contact-maps from a triplet of coevolutionary matrices through deep residual convolutional networks," *PLOS Comput. Biol.*, vol. 17, no. 3, p. e1008865, Mar. 2021, doi: 10.1371/journal.pcbi.1008865.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," 2016, pp. 770–778. Accessed: Feb. 02, 2022. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html

[28] "Identification of direct residue contacts in protein–protein interaction by message passing | PNAS." https://www.pnas.org/content/106/1/67.short (accessed Feb. 02, 2022).

[29] D. S. Marks *et al.*, "Protein 3D Structure Computed from Evolutionary Sequence Variation," *PLOS ONE*, vol. 6, no. 12, p. e28766, Dec. 2011, doi: 10.1371/journal.pone.0028766.

[30] "PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments | Bioinformatics | Oxford Academic." https://academic.oup.com/bioinformatics/article/28/2/184/198108?login=true (accessed Feb. 02, 2022).

[31] "End-to-End Differentiable Learning of Protein Structure - ScienceDirect." https://www.sciencedirect.com/science/article/pii/S2405471219300766 (accessed Feb. 02, 2022).

[32] "Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13) - Senior - 2019 - Proteins: Structure, Function, and Bioinformatics - Wiley Online Library." https://onlinelibrary.wiley.com/doi/full/10.1002/prot.25834 (accessed Feb. 02, 2022).

[33] J. Ingraham, A. Riesselman, C. Sander, and D. Marks, "Learning Protein Structure with a Differentiable Simulator," presented at the International Conference on Learning Representations, Sep. 2018. Accessed: Feb. 02, 2022. [Online]. Available: https://openreview.net/forum?id=Byg3y3C9Km

[34] J. Li, "Universal Transforming Geometric Network," *ArXiv190800723 Cs Q-Bio*, Aug. 2019, Accessed: Feb. 02, 2022. [Online]. Available: http://arxiv.org/abs/1908.00723

[35] J. Xu, M. McPartlon, and J. Li, "Improved protein structure prediction by deep learning irrespective of co-evolution information," *Nat. Mach. Intell.*, vol. 3, no. 7, Art. no. 7, Jul. 2021, doi: 10.1038/s42256-021-00348-5.

[36] A. Vaswani *et al.*, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, 2017, vol. 30. Accessed: Feb. 02, 2022. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[37] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-Cross

Attention for Semantic Segmentation," 2019, pp. 603–612. Accessed: Feb. 02, 2022. [Online]. Available: https://openaccess.thecvf.com/content_ICCV_2019/html/Huang_CCNet_Criss-Cross_Attention_for_Semantic_Segmentation_ICCV_2019_paper.html

[38] "Axial-DeepLab: Stand-Alone Axial-Attention for Panoptic Segmentation | SpringerLink." https://link.springer.com/chapter/10.1007/978-3-030-58548-8_7 (accessed Feb. 02, 2022).

[39] E. C. Alley, G. Khimulya, S. Biswas, M. AlQuraishi, and G. M. Church, "Unified rational protein engineering with sequence-based deep representation learning," *Nat. Methods*, vol. 16, no. 12, Art. no. 12, Dec. 2019, doi: 10.1038/s41592-019-0598-1.

[40] M. Heinzinger *et al.*, "Modeling aspects of the language of life through transfer-learning protein sequences," *BMC Bioinformatics*, vol. 20, no. 1, p. 723, Dec. 2019, doi: 10.1186/s12859-019-3220-8.

[41] "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences | PNAS." https://www.pnas.org/content/118/15/e2016239118.short (accessed Feb. 02, 2022).

[42] "CASP 1." [Online]. Available: https://www.uniprot.org/uniprot/P31944

[43] "CASP 2." [Online]. Available: https://www.uniprot.org/uniprot/P31944

[44] "CASP 3." [Online]. Available: https://www.uniprot.org/uniprot/P31944

[45] "CASP 4." [Online]. Available: https://www.uniprot.org/uniprot/P31944

[46] "CASP 5." [Online]. Available: https://www.uniprot.org/uniprot/P51878

[47] "CASP 6." [Online]. Available: https://www.uniprot.org/uniprot/P55212

[48] "CASP 7." [Online]. Available: https://www.uniprot.org/uniprot/P55210

[49] "CASP 8." [Online]. Available: https://www.uniprot.org/uniprot/Q14790

[50] "CASP 9." [Online]. Available: https://www.uniprot.org/uniprot/P55211

[51] "CASP 10." [Online]. Available: https://www.uniprot.org/uniprot/Q92851

[52] "CASP 11." [Online]. Available: https://www.uniprot.org/uniprot/Q91XW7

[53] "CASP 12." [Online]. Available: https://www.uniprot.org/uniprot/Q6UXS9

[54] L. Floridi and M. Chiriatti, "GPT-3: Its Nature, Scope, Limits, and Consequences," *Minds Mach.*, vol. 30, no. 4, pp. 681–694, Dec. 2020, doi: 10.1007/s11023-020-09548-1.

[55] S. Gao *et al.*, "Limitations of Transformers on Clinical Text Classification," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 9, pp. 3596–3607, Sep. 2021, doi: 10.1109/JBHI.2021.3062322.

[56] K. Tunyasuvunakool *et al.*, "Highly accurate protein structure prediction for the human proteome," *Nature*, vol. 596, no. 7873, Art. no. 7873, Aug. 2021, doi: 10.1038/s41586-021-03828-1.

[57] wwPDB consortium, "Protein Data Bank: the single global archive for 3D macromolecular structure data," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D520–D528, Jan. 2019, doi: 10.1093/nar/gky949.

[58] D. Rusciano, D. R. Welch, and M. M. Burger, Eds., "Homotypic and heterotypic cell adhesion in metastasis," in *Laboratory Techniques in Biochemistry and Molecular Biology*, vol. 29, Elsevier, 2000, pp. 9–64. doi: 10.1016/S0075-7535(00)29003-7.

[59] RaoHassanKaleem, *RaoHassanKaleem/Diebetes-Detection-using-Machine-Learning-Algorithms*. 2022.

Accessed: Feb. 14, 2022. [Online]. Available: https://github.com/RaoHassanKaleem/Diebetes-Detection-using-Machine-Learning-Algorithms

[60] "Feature Extraction App - Proteins · Streamlit." https://share.streamlit.io/raohassankaleem/fetprotextract/main/app.py (accessed Apr. 13, 2022).

[61] "Aminoacids-peptides-primary-structure_0.pdf." Accessed: Feb. 15, 2022. [Online]. Available: https://biochimia.usmf.md/sites/default/files/inline-files/Aminoacids-peptides-primary-structure_0.pdf

[62] S. J. Malebary, M. S. ur Rehman, and Y. D. Khan, "iCrotoK-PseAAC: Identify lysine crotonylation sites by blending position relative statistical features according to the Chou's 5-step rule," *PLOS ONE*, vol. 14, no. 11, p. e0223993, Nov. 2019, doi: 10.1371/journal.pone.0223993.