Comparison of 3D Structures Generated by AlphaFold2 to Experimental Structures in Oncogenic Proteins

Simar Rajpal[1] and Simar Rajpal[#]

[1]John P. Stevens High School
[#]Advisor

ABSTRACT

AlphaFold2 is a machine-learning algorithm that can predict the 3D structure of proteins. 3D protein structures are essential for understanding the function of oncogenic proteins, which can potentially cause cancer. In this study, we compared the structure of 26 oncogenic proteins found experimentally and computationally using AlphaFold2. We used RMSD values to measure how well the AlphaFold2 model fit the experimentally derived protein structures. RMSD values for the oncogenic proteins ranged from 0.204 Å to 1.980 Å with an average of 0.633 Å, showing that AlphaFold2 was a promising tool for predicting the 3D structure of oncogenic proteins. However, we noted that AlphaFold2 has limitations in predicting the structure of highly disordered proteins, proteins with multiple conformations, mutated and artificial proteins, and proteins that are not well-studied experimentally. Our study suggests that AlphaFold2 could be used to identify new targets for cancer treatment and design drugs that can fit into the binding sites of oncogenic proteins. We also hypothesize that AlphaFold2 could be improved by increasing the amount of data available for training, improving the resolution of current data, and using it in conjunction with other protein structure models. We believe that AlphaFold2 is a powerful tool for predicting the structure of oncogenic proteins. Despite the current limitations, we are optimistic that future research will improve AlphaFold2, finding use in cancer research.

Introduction

Proteins and Their Importance of Structural Biology

Proteins are the building blocks of life: every cell in the human body contains proteins. Proteins serve a wide variety of purposes, including catalyzing chemical reactions, regulating the immune system to protect the body from bacteria and viruses, and forming muscle [1]. Proteins play a vital role in almost all biological processes, with around 20,000 protein-coding genes generating nearly 200 million unique protein-protein interactions controlling various function of human body [2]. Proteins are composed of amino acids bonded together in chains, with their side chains determining their overall 3D structure. In order to understand the function of a protein, it is necessary to understand its 3D structure, as every protein obtains its function from its unique 3D structure [3]. The structure of proteins is crucial for many reasons: assisting in drug design and development, understanding diseases and mutations, engineering proteins with desired properties, and understanding how proteins interact with other biomolecules [4]. For example, by knowing the 3D structure of a target protein, scientists can design drugs that fit precisely into its binding site to modulate its activity [5]. As a result, structural biology is essential in helping combat many of the world's most complex biological problems.

Application of Structural Biology: COVID-19

The importance of structural biology is shown in a real-world scenario through the coronavirus, or COVID-19, which caused a worldwide pandemic [6], resulting in more than 287 million confirmed cases globally and more than 5 million deaths [7]. Scientists at companies including Pfizer and Moderna could develop COVID-19 vaccines after solving the 3D structure of proteins in the SARS-CoV-2 virus, particularly the spike protein on its surface [8].

The spike protein plays an important role in the virus's ability to enter human cells by binding to specific receptors on the surface of host cells. By using advanced techniques such as cryo-electron microscopy (cryo-EM), researchers were able to determine the precise atomic-level structure of the spike protein [9]. This detailed structure helped them identify specific regions or domains on the spike protein involved in receptor binding and membrane fusion during the virus's entry into human cells [10]. By targeting these regions, vaccines helped induce a strong immune response against the spike protein, helping the immune system recognize and neutralize the virus. The spike protein's structure played a key role in making the vaccine and helped scientists understand potential variations or mutations in the virus [11]. It was crucial for assessing the vaccine's effectiveness against emerging virus variants, such as Omicron and Delta.

The Protein Folding Problem

However, determining the 3D structure of proteins, known as the Protein Folding Problem, is challenging. This process can be time-consuming because of many reasons. One reason is that proteins have a vast conformational space: a protein can fold and arrange its 3D structure in many possible ways. Despite their vast conformational space, proteins can fold into their native states very quickly, making it practically impossible to go through every conformation in a reasonable amount of time [12]. Larger proteins will also have even more possible conformations, making it more challenging to explore all the conformations efficiently. On the other hand, protein folding is also partially determined by the complex interactions between its amino acids, which rely on many different forces, such as hydrogen bonding and van der Waals interactions [3].

Proteins adopt their energy minimum or native structure in nature through protein folding. Many principles and theories support this process. Anfinsen’s Principle, the most fundamental concept underlying protein folding, states that the amino acid sequence (primary structure) and the interactions between the amino acid residues determine the native structure of a protein [3]. This principle explains that the information for protein folding is hidden within its sequence [13]. Furthermore, thermodynamics supports the protein folding process because the native structure represents the most stable conformation with the lowest Gibbs free energy, a measure of the total energy for a system that can be used to do work [14]. The protein will naturally fold into the native structure because it is energetically favorable, and it is the global energy minimum of the protein’s conformation landscape. Though restricting its possible conformations decreases entropy, interactions such as hydrogen bonds and van der Waals interactions help outweigh the decrease and make the protein folding process favorable: entropy-enthalpy compensation [15]. There are also theories, especially for larger proteins, about the conformational landscape, having many local energy minima and kinetic traps [16] that might hinder or slow down the protein folding process, which makes it essential for molecular chaperones to guide the folding process and ensure that proteins reach their energy minimum state efficiently.

Transition from Experimental Methods to Machine Learning Methods

Over the decades, biologists have determined protein structures experimentally through many different methods, such as X-ray crystallography, nuclear magnetic resonance spectroscopy, and cryo-electron microscopy [17].

X-ray crystallography is a technique that determines the 3D structure of proteins by analyzing the diffraction patterns produced when X-rays pass through a crystallized protein sample [18]. NMR spectroscopy is a method used to study the structure and dynamics of proteins in solution by analyzing the interactions between nuclear spins in a magnetic field [19]. Cryo-electron microscopy uses frozen samples and electron microscopy to obtain high-resolution 3D structures of proteins without the need for crystallization [20]. Although these methods are accurate, they are also expensive and time-consuming. Furthermore, many of these methods pose unique challenges. For example, X-ray crystallography requires the crystallization of the protein and the growth of high-quality crystals. However, not all proteins crystallize, and it is a very time-consuming and unpredictable process, especially for large and complex proteins or hydrophobic proteins [21].

To address the limitations of experimental methods, researchers have applied deep learning methods to predict the 3D structure of proteins. Machine learning methods offer a faster and more cost-effective approach to protein structure prediction by leveraging the power of AI and deep learning algorithms to analyze large datasets of known protein structures and make predictions about the structure of new proteins [22]. Using a neural network called Pocket Miner, a team of researchers from the University of Pennsylvania have found that up to 50% of cancer-signaling proteins once believed to be immune to drug treatment, due to a lack of targetable protein regions, actually may be treatable. PocketMiner was trained on a dataset of over 100,000 proteins and could identify hidden pockets in 50% of the cancer-signaling proteins. These hidden pockets could help identify new drug targets for cancer and other diseases, showing the promise of machine-learning methods [23].

How Does AlphaFold2 Work?

AlphaFold2 is a machine learning method for protein structure prediction using deep learning developed by Google's DeepMind [24] [25]. It utilizes a convolutional neural network architecture (CNN) to predict the 3D structure of proteins from their amino acid sequences. The neural network was trained on a large dataset of known protein structures. This dataset contains pairs of protein sequences and their corresponding experimentally determined structures. The neural network learned to recognize patterns in the amino acid sequences indicative of specific structural features. It also learned to predict the distances between pairs of amino acids and angles between adjacent amino acids, which are essential for finding the 3D structure of the protein. Once the neural network was trained, it can be used for protein structure prediction. Given the amino acid sequence of a protein, AlphaFold2 predicts the distances and angles between the amino acids to generate a 3D model of the protein [26].

The key innovations of AlphaFold2 are the use of attention mechanisms, symmetry principles, and end-to-end differentiability [27]. Attention mechanisms allow the network to focus on different parts of the protein sequence when making predictions, considering the long-range dependencies between amino acids. Furthermore, symmetry principles help reduce the complexity of protein structure predictions, and end-to-end differentiability helps AlphaFold2 learn the relationships between the amino acid sequence and protein structure more naturally.

AlphaFold2 has demonstrated accuracy through its performance in the Critical Assessment of Protein Structure Prediction (CASP) Competition. In the CASP 14 Competition, the AlphaFold2 algorithm created high GDT score models for almost two-thirds of the target protein structures at an accuracy comparable to experimental methods [28]. AlphaFold2 is notable when compared to other computational models because of its accuracy in the CASP competition, its use of attention-based neural networks to capture important features of protein folding, its training on a large dataset such as the PDB, and its speed in finding the three-dimensional structure for individual protein structures.

Oncogenic Proteins and Cancer

Cancer is a group of diseases characterized by uncontrollable cell growth and the ability of cells to invade tissues or spread to parts of the body through metastasis [29]. There are many types of cancer, such as breast, lung, and colon. According to the NIH, in 2023, approximately 1.96 million new cancer cases and 610,000 cancer deaths are projected to occur in the United States [30]. As a result, billions of dollars are spent annually in the form of research grants to find potential drugs and therapies to mitigate the effect of this deadly disease [31].

Oncogenic proteins are proteins that have the potential to cause cancer. They are often mutated or overexpressed in cancer cells, leading to abnormal cell growth and proliferation as well as the formation of tumors [32]. Oncogenic proteins can disrupt normal cellular processes such as cell cycle regulation, DNA repair, and cell signaling pathways. Typically, the growth and division of cells are regulated by many checkpoints and signaling pathways within the cell. However, in cancer cells, oncogenic proteins can interfere with these regulatory mechanisms, leading to uncontrolled cell division and growth, thus causing the accumulation of cells and the formation of tumors [33]. Mutations in oncogenic proteins can promote growth factors and the inactivation of inhibitors, disrupting cell cycle regulation. Furthermore, oncogenic proteins can interfere with DNA repair processes, causing genetic mutations that activate other oncogenes or the inactivation of tumor suppressor genes. Finally, oncogenic proteins can also take control of cell signaling pathways, leading to the activation of pro-growth signals and the inhibition of anti-growth signals [34].

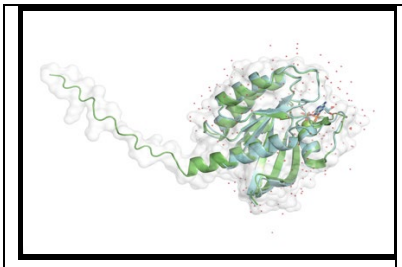
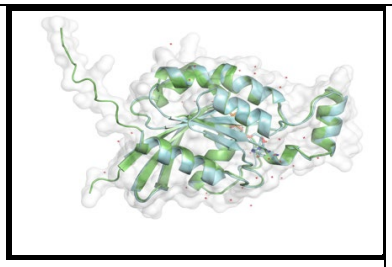
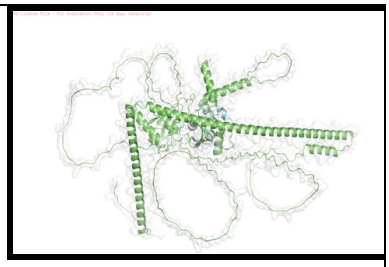
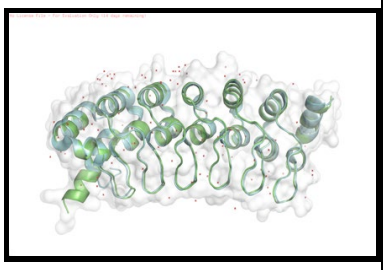
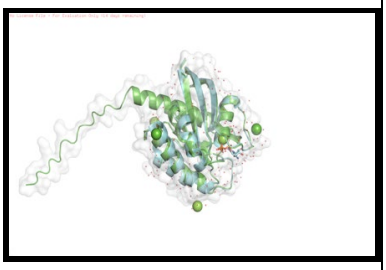
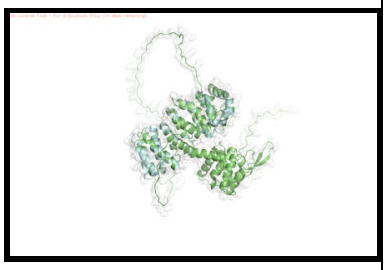
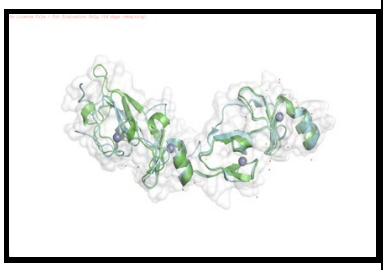
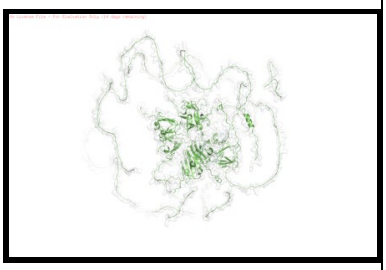
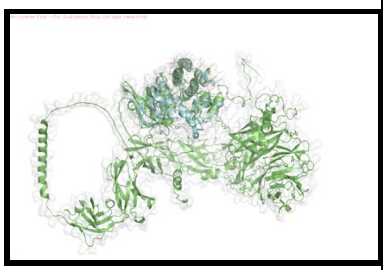
Several types of oncogenic proteins exist, including oncogenes and tumor suppressor genes [35]. Oncogenes are genes that, when mutated or activated, promote cell growth and division. Some examples are the HER2 gene in breast cancer and the KRAS gene in colorectal cancer [36]. Alternatively, tumor suppressor genes are genes that normally regulate cell growth and prevent the formation of tumors. Mutations or inactivations of tumor suppressor genes can lead to uncontrolled cell growth and cancer. Examples of tumor suppressor genes include the TP53 gene and the BRCA1 and BRCA2 themes in breast and ovarian cancer [37].

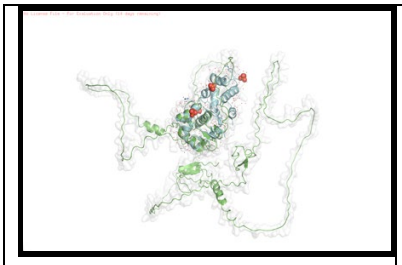
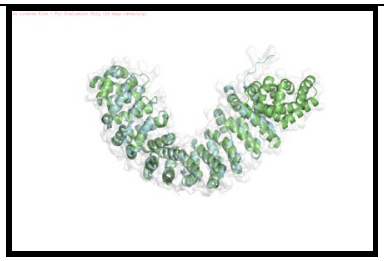
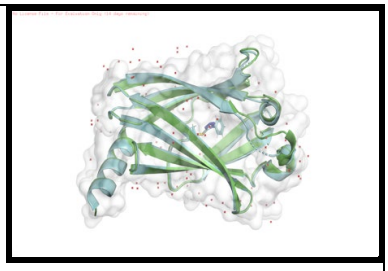
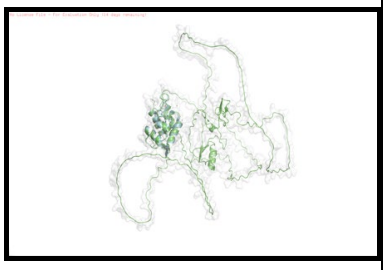
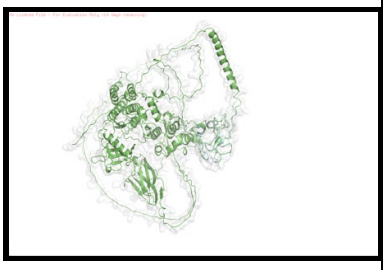
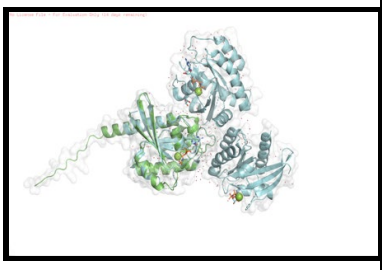
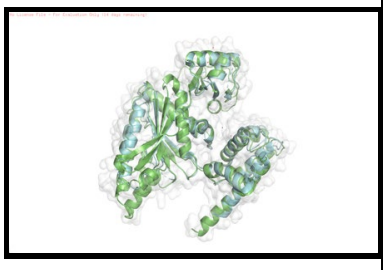
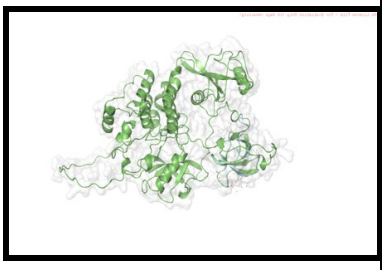
Understanding the role of oncogenic proteins is crucial for developing targeted therapies and personalized medicine approaches for cancer treatment. By identifying specific oncogenic proteins involved in a particular type of cancer, researchers can develop drugs targeting these proteins, inhibiting their activity and preventing cancer growth [38]. One example is the development of tyrosine kinase inhibitors (TKIs) for treating chronic myeloid leukemia, CML. Understanding the structure of the BCR-ABL kinase, an oncogenic protein in CML, led to the development of TKIs like imatinib. These TKIs targeted the abnormal protein and blocked its activity, halting the uncontrolled cell growth in CML. Imatinib and other TKIs reduced symptoms and signs of cancers and prolonged survival [39].

Data

All of the proteins in this study except one were solved experimentally using X-ray diffraction. Protein 5Z55's structure was found using NMR spectroscopy [40].

PDB Name: 121P Pharos Number/UniProt ID: P011112 RMSD Value (Å): 0.391 Picture of Model (PyMOL)	PDB Name: 1A2B Pharos Number/UniProt ID: P61586 RMSD Value (Å): 0.466 Picture of Model (PyMOL)	PDB Name: ISBX Pharos Number/UniProt ID: P12755 RMSD Value (Å): 0.290 Picture of Model (PyMOL)

		
(1a)	(1b)	(1c)
PDB Name: 1UOH Pharos Number/UniProt ID: O75832 RMSD Value (Å): 0.450 Picture of Model (PyMOL)	PDB Name: 1ZW6 Pharos Number/UniProt ID: P01112 RMSD Value (Å): 0.501 Picture of Model (PyMOL)	PDB Name: 2ILR Pharos Number/UniProt ID: Q9HB96 RMSD Value (Å): 0.428 Picture of Model (PyMOL)
		
(1d)	(1e)	(1f)
PDB Name: 2XJY Pharos Number/UniProt ID: P25791 RMSD Value (Å): 1.889 Picture of Model (PyMOL)	PDB Name: 3CW3 Pharos Number/UniProt ID: Q9Y4K1 RMSD Value (Å): 0.621 Picture of Model (PyMOL)	PDB Name: 3DKC Pharos Number/UniProt ID: Q9HB96 RMSD Value (Å): 1.980 Picture of Model (PyMOL)
		
(1g)	(1h)	(1i)
PDB Name: 3EQY Pharos Number/UniProt ID: O15151 RMSD Value (Å): 0.729 Picture of Model (PyMOL)	PDB Name: 5XGC Pharos Number/UniProt ID: P52306 RMSD Value (Å): 0.486 Picture of Model (PyMOL)	PDB Name: 5YAV Pharos Number/UniProt ID: O43924 RMSD Value (Å): 0.511 Picture of Model (PyMOL)

		
(1j)	(1k)	(1l)
PDB Name: 5Z02 Pharos Number/UniProt ID: Q00987 RMSD Value (Å): 0.767 Picture of Model (PyMOL)	PDB Name: 5Z55 Pharos Number/UniProt ID: Q01973 RMSD Value (Å): 0.790 Picture of Model (PyMOL)	PDB Name: 6CU6 Pharos Number/UniProt ID: P01116 RMSD Value (Å): 0.363 Picture of Model (PyMOL)
		
(1m)	(1n)	(1o)
PDB Name: 6LK0 Pharos Number/UniProt ID: Q15645 RMSD Value (Å): 1.011 Picture of Model (PyMOL)	PDB Name: 6N2J Pharos Number/UniProt ID: P01116 RMSD Value (Å): 0.320 Picture of Model (PyMOL)	PDB Name: 6NMW Pharos Number/UniProt ID: P07948 RMSD Value (Å): 0.295 Picture of Model (PyMOL)
		
(1p)	(1q)	(1r)
PDB Name: 6VO7 Pharos Number/UniProt ID: P42336 RMSD Value (Å): 0.497 Picture of Model (PyMOL)	PDB Name: 6VRV Pharos Number/UniProt ID: P11309 RMSD Value (Å): 0.277 Picture of Model (PyMOL)	PDB Name: 7BJ1 Pharos Number/UniProt ID: Q9H7B4 RMSD Value (Å): 0.204 Picture of Model (PyMOL)

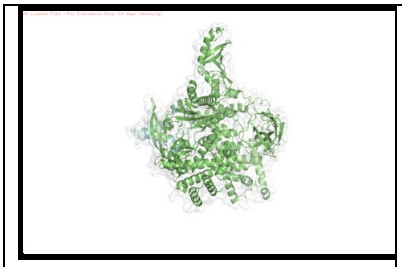
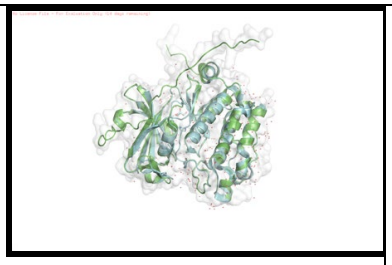
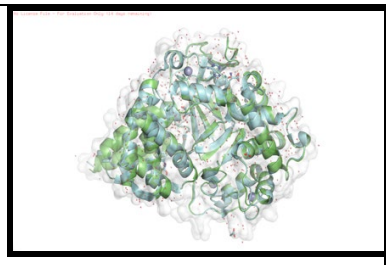
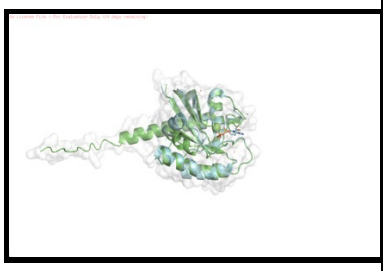
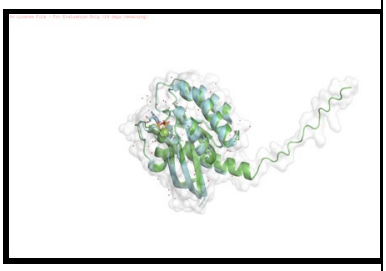
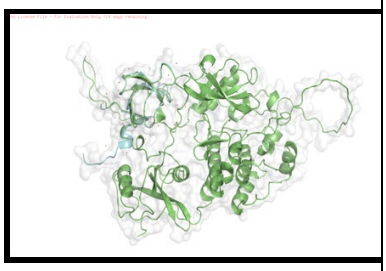
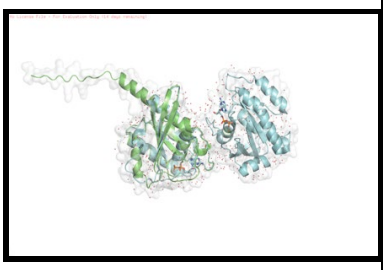
 (1s)	 (1t)	 (1u)
PDB Name: 7C40 Pharos Number/UniProt ID: P01116 RMSD Value (Å): 0.467 Picture of Model (PyMOL)  (1v)	PDB Name: 7JIG Pharos Number/UniProt ID: P01112 RMSD Value (Å): 0.513 Picture of Model (PyMOL)  (1w)	PDB Name: 7JT9 Pharos Number/UniProt ID: P09769 RMSD Value (Å): 0.269 Picture of Model (PyMOL)  (1x)
PDB Name: 7JU2 Pharos Number/UniProt ID: P41212 RMSD Value (Å): 1.479 Picture of Model (PyMOL)  (1y)	PDB Name: 7TLK Pharos Number/UniProt ID: P01116 RMSD Value (Å): 0.461 Picture of Model (PyMOL)  (1z)	

Figure 1. Structures of 26 key oncogenic proteins used in the study. Figures illustrate a comparison of the experimentally derived structure (shown in blue) with AlphaFold2's predicted structure (shown in green), aligned and generated with PyMol.

Methods

Benchmarking the Model

In the experimental study, we used the visualization and analysis program PyMOL to determine the RMSD values between the oncogenic proteins from the PDB and AlphaFold2 [40]. We first benchmarked the model by testing the model to make sure it was accurate. We used the 121P protein, the first protein in the data table, and compared it to itself to confirm that the RMSD for identical proteins is 0 Å. In addition, we compared two proteins with very different structures to confirm a high RMSD [42].

Experiment

After benchmarking the model, we identified oncogenic proteins to use in the study. Specifically, we used the "ONCOPROTEIN" and "ONCOGENE PROTEIN" filters in PDB to focus the study on proteins originating from an initial query of more than 300 proteins. We shortened the list by using a variety of factors. We did not choose proteins from the PDB that were in complex with another protein, we narrowed the search results to proteins from homo sapiens as the organism, we focused on lyases as they have a simpler 3D structure, and proteins up to 61 kDa. Finally, we confirmed that the structures of the proteins in the study were also solved in AlphaFold2's database of proteins and compared their amino acid sequence to make sure they were identical. We used the align command in PyMOL to find the RMSD: align molecule1, molecule2 [41]. In PyMOL, the align command is used to superimpose and align two molecules based on their atomic coordinates. This command helps compare and visualize the structural similarities and differences between different proteins. PyMOL uses a two-step approach for aligning structures: first, it does a sequence alignment, and then it minimizes the RMSD between the aligned sequence residues. The align command can even be used to align specific chains or residues using the sele command to specify the chains or residues of interest.

Image Representation

For the images of the 26 proteins involved in the experiment, we included the structural model of the two proteins in the background. In order to provide a general sense of the volume of the protein while being able to visualize the cartoon representation, we included a surface but reduced the transparency to 80%, leaving the cartoon visible with the surface in the background (Figure 1). We used the cartoon representation because it simplifies the protein structure by showing its backbone as ribbons and helices. It also helps visualize the secondary structure of the protein, its alpha helices and beta sheets, which helps us understand the protein's overall folding pattern [43].

The stick representation can make models appear cluttered and difficult to interpret. It is often used when focusing on specific molecular interactions or to look closely at a protein's active site. The space-filling representation is more helpful in understanding the protein volume and steric clashes, but it also can become cluttered and ineffective for visualizing the overall structure.

Results

We calculated the RMSD between the experimentally determined and AlphaFold2-predicted proteins. The RMSD stands for Root Mean Square Deviation and is used to assess the similarity between two three-dimensional structures of biomolecules. In this instance, it is being used to assess the accuracy of computational models from AlphaFold2 by comparing them to experimentally determined structures obtained primarily from

X-ray diffraction [44]. Once the proteins are aligned, the RMSD value is found by measuring the average deviation between their corresponding atom positions. RMSD has been used to assess the quality and accuracy of the computationally found structures of proteins [42]. Lower RMSD values indicate a higher degree of similarity between the experimentally determined structures and the computational structures, which helps determine the model's reliability.

The lowest RMSD value was 0.204 Å. The highest RMSD value was 1.980 Å. The range was 1.776 Å. The mean RMSD value of the oncogenic proteins in the study is approximately 0.633 Å, and the distribution of average RMSDs is shown in Figure 2. Of the 26 oncogenic proteins, 22 are less than 1.0 Å RMSD, and 19 are less than the mean of 0.633 Å. There are four outliers with RMSD above 1.00 Å. The RMSD values only compared the same regions of both proteins in PyMOL's calculation. As a result, the larger proteins do not necessarily have a higher RMSD value, as shown by proteins such as 1SBX, 3CW3, and 2ILR.

Furthermore, the study also involved looking at AlphaFold2's confidence levels for different sections of the proteins. Regarding the oncogenic proteins involved in the study, AlphaFold2 showed a higher confidence, blue color, in the core and middle regions of the proteins but showed lower confidence, yellow and orange color, in the ends of the proteins [24] [25]. Furthermore, for some of the oncogenic proteins involved in the study, AlphaFold2's prediction had a much larger molecule, with the core region being the same as the experimental model but with other branches. These branches were in general yellow and orange, which indicates a lower confidence level.

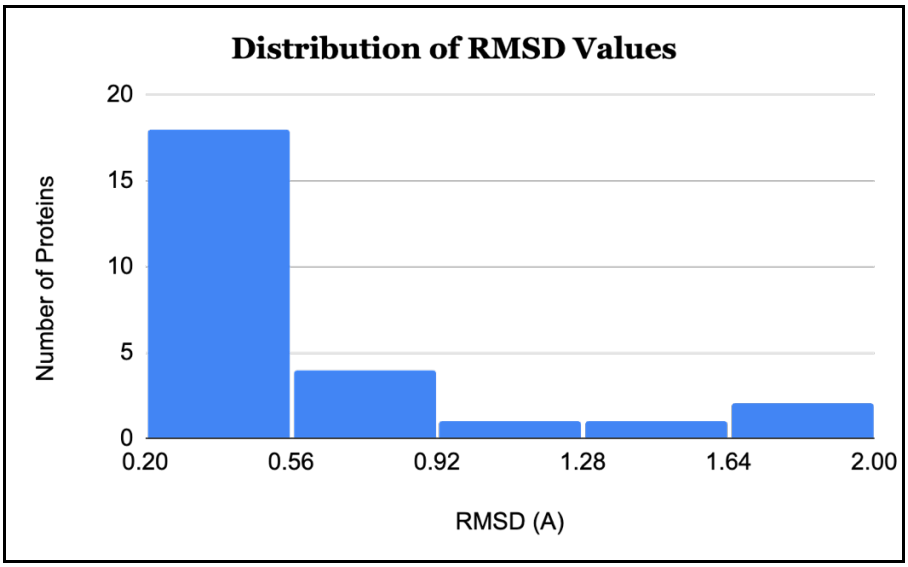


Figure 2. Distribution of RMSD Values for Key Oncogenic Proteins

Analysis

Overview

Oncogenic proteins have features that can make rationalizing their structure prediction difficult. Oncogenic proteins often have mutations that change their amino acid sequence, which could alter the folding pattern of the protein and may cause the protein to unfold completely [45]. Furthermore, oncogenic proteins may be unstable because they are missing or have mutations in nucleotides that code for stabilizing regions, enabling the protein to constantly change shape. Oncogenic proteins also are often involved in complex interactions with

other biomolecules, which play an essential role in their function, making their structure more challenging to predict.

RMSD Value and Applications

In general, an RMSD value below 2 Å is considered good. All the proteins in the experiment had an RMSD value below 2 Å, which shows the promise of AlphaFold2 in predicting the structure of oncogenic proteins. However, the error acceptability of a model depends on the application. In some cases, such as structure-based drug design, the error must be as close to zero as possible. Using rational drug design to design a pharmaceutical therapy targeting an active site whose structure is modeled incorrectly could lead to wasted effort in research and development. If a model is used to predict the protein structure in a patient the error must be small because even larger error could waste research and development resources. In industrial processes, designing enzymes to catalyze biological reactions also requires a precise understanding of the protein's structure.

Hypothesis and Trends

Based on the data, we observed a few trends relating proteins with a higher RMSD to those with a lower RMSD. We hypothesize that proteins that have complex interactions with other biomolecules would have a higher RMSD: proteins in complexes of multiple subunits and proteins binding to molecules such as ligands and co-factors [46]. We further hypothesize that large and complex protein structures would have a higher RMSD: proteins with disordered regions, post-translational modifications, and multiple conformations.

Proteins 2XJY, 3DKC, 7JU2, and 6LK0 all have an RMSD greater than 1.00 Å. Protein 2XJY has a high RMSD of 1.889 Å. We analyzed the protein on AlphaFold2. It has disordered regions and can adopt multiple conformations, supporting our hypothesis [24][25]. Meanwhile, Protein 3DKC also had the highest RMSD of 1.980 Å. 3DKC is the MET receptor tyrosine kinase in complex with ATP. We believe that the flexible and dynamic complex made it difficult for AlphaFold2 to predict its structure due to the complexity of its structure. Furthermore, Protein 7JU2 has an RMSD of 1.479 Å. We determined that Protein 7JU2 is a homodimer [47]: two copies of the protein that bind to each other, making it difficult for AlphaFold2 to predict its structure. Protein 6LK0, which has an RMSD of 1.011 Å, is a large, disordered protein known to have complex interactions with other receptor proteins.

On the other hand, proteins with a lower RMSD likely have simpler structures. Furthermore, they are often well-studied and have more accurate experimental structures because AlphaFold2 uses data from the PDB to make its own prediction. Proteins 121P, 5Z55, and 6VRV, which have RMSD values of 0.391 Å, 0.790 Å, and 0.277 Å, respectively, are small and well-defined.

Conclusion About AlphaFold2's Efficacy

Overall, the results of the analysis of the 26 oncogenic proteins suggest that AlphaFold2 can generally predict the structure of oncogenic proteins with good accuracy. The results suggest that AlphaFold2 is a promising tool for predicting the structure of oncogenic proteins for cancer research. Despite the accuracy of AlphaFold2 in determining the structures, future research is needed to determine the trends and features of proteins that AlphaFold2 needs to improve upon to improve its modeling of oncogenic proteins.

Discussion

Contribution to Current Cancer Research

The results of this analysis have many implications for current cancer research. First, the ability to predict the structure of oncogenic proteins with high accuracy could help researchers identify new targets for cancer treatment. By understanding the structure of oncogenic proteins, researchers can determine the role these proteins play in cancer development, becoming new targets for potential drugs. AlphaFold2 also plays a crucial role in drug design as AlphaFold2 can predict the binding sites of proteins, areas of the protein that interact with other biomolecules [5]. By understanding more about the structure of oncogenic proteins, researchers can design drugs that fit into their binding sites and adjust their function.

Furthermore, understanding the structure of oncogenic proteins helps to predict their interactions between other oncogenic proteins and biomolecules, as the structure of a protein is intricately linked to its function. As a result, the structure of these oncogenic proteins can give researchers insight into the role these proteins play in cancer progression. This knowledge could help researchers understand mechanisms leading to cancer or design new cancer therapies.

How to Improve the Model

Despite the accuracy of AlphaFold2, some steps could be taken to improve the model's accuracy when finding the structures of oncogenic proteins. One method could be increasing the amount of training data since AlphaFold2 is a machine-learning model trained on a dataset of protein sequences and structures. The more available data, the better AlphaFold2 will likely be at understanding and learning patterns. AlphaFold2 can use the data to learn the relationships between different amino acids and how they interact, which is essential for understanding how proteins fold. In the case of oncogenic proteins, having more data is especially important because these proteins are often complex and have complex interactions with other biomolecules. In addition, improving the accuracy of the training data is helpful because it will help AlphaFold2 make more accurate predictions.

AlphaFold2 could also be used with other protein structure prediction models to make more accurate models of oncogenic proteins in cancer research. Many other protein structure models are available, each with their own strengths and weaknesses. We hypothesize that if AlphaFold2 is used with other similar machine learning algorithms, such as Rose TTAFold [48], researchers can improve the accuracy of their predictions for oncogenic proteins. Furthermore, AlphaFold2 can be used to identify specific protein regions that are hard to predict. Then these specific regions could be focused on with other models to help create a more accurate protein model.

Additionally, cancer researchers can use other methods to improve AlphaFold2's ability to predict oncogenic protein structures. Researchers can use other drug discovery engines, such as PandaOmics, which help combine bioinformatic methods with advanced deep learning approaches for target identification, helping optimize the initial steps of drug discovery [49]. Researchers can focus on data from cancer cells and train AlphaFold2 on this data to help AlphaFold2 learn the critical features and patterns of oncogenic proteins essential to understanding cancer progression. Furthermore, incorporating information about mutations could be beneficial because oncogenic proteins often have mutations that change their structure and function [50]. Both tumor suppressors and oncogenes become mutated in order to trigger cancer progression. Incorporating this information could help increase the accuracy of predictions for oncogenic proteins. Finally, developing new algorithms and models that predict the interactions of oncogenic proteins with other biomolecules, such as ligands and cofactors, will improve the accuracy of the model [51].

What Can Be Taken Away from This Model

Based on AlphaFold2's ability to predict oncogenic proteins, we believe a few key lessons can be taken away from this model. The first is the importance of structure and finding the most accurate and efficient way to

predict the structure of proteins to understand their function in biological processes such as cancer. The second is the power of machine learning, as AlphaFold2 is a machine learning model that could improve over time as more training data becomes available. The third is the potential for new cancer treatments using AI and machine learning to accelerate research and development to create beneficial drugs and prevention strategies.

Limitations

AlphaFold2 is a powerful tool for predicting the structure of oncogenic proteins, but it has some limitations that decrease its accuracy.

Complexity of Protein Structure

AlphaFold2 has limitations when finding the structure of intrinsically disordered proteins. These proteins do not have a well-defined structure, and their amino acid sequence can vary significantly. Their lack of a defined structure makes it difficult for AlphaFold2 to predict their structure since it does not have a clear reference point [52]. Furthermore, highly disordered proteins are often involved in complex interactions with other biomolecules, making them challenging to study and predict their structure.

AlphaFold2 has limitations when finding the structure of proteins with multiple conformations (i.e., different ways of folding). This makes it difficult for AlphaFold2 to predict their structure because the model does not know the most stable conformation. In a study to test the ability of algorithms to find the structure of proteins with multiple conformations, AlphaFold2 received a 9.88% success ratio [53]. Multiple conformation proteins can be found in many biological systems because their ability to change conformations is essential for their function. For example, hemoglobin has multiple conformations that allow it to bind more oxygen when needed and release oxygen when it is not [54].

Protein Interactions and Binding Partners

AlphaFold2 may be challenged in predicting proteins in complex with other molecules [55]. Protein interactions are often very complex and can be influenced by many factors, such as the environment, the presence of other proteins, and the binding of small molecules. Due to this, AlphaFold2 also struggles to find accurate structures of proteins that bind with ions, cofactors, ligands, or post-translational modifications such as phosphorylation [46]. Both limitations highlight steps necessary to improve AlphaFold2's modeling, especially for drug design.

Finding binding partners is very complex because of the vast number of possible binding partners, the specificity of binding, and the dynamic process of binding [56]. For example, if a protein is known to interact with a specific DNA sequence, then AlphaFold2 can use this information to improve its model of the protein's structure. This information helps because the DNA sequence or other potential binding partners can provide an image of the binding site's shape and the protein's region that interacts with the DNA. Finding binding partners helps improve information about the presence of other biomolecules and the environment of the protein, helping improve the accuracy of the structure. Finding binding partners also helps scientists understand the function of a protein. For example, the DNA sequence suggests that the protein may be involved in gene regulation.

Mutated and Artificial Proteins

AlphaFold2 is not always accurate for predicting mutated proteins [57]. Mutations can change the structure of proteins in unpredictable ways, and if a mutation changes the amino acid sequence, it makes it more difficult to predict the structure. Furthermore, mutations increase the number of structures that must be studied, each

requiring significant time and resources. However, the difference between mutated and wild-type structures generated by AlphaFold2 is very small. Mutations can also affect the protein's interactions with other biomolecules and the stability of a protein, which can lead to a change in its structure.

AlphaFold2 may find predicting the structures of artificial or man-made proteins challenging because they typically have unusual or unexpected structures. Artificial proteins could adopt structures not found in naturally occurring proteins, so AlphaFold2 is not trained on this data. Artificial proteins are also typically designed to achieve a particular function, creating different structures from naturally occurring proteins.

Current Data Available

AlphaFold2 uses a deep-learning model to predict protein structure from amino acid sequences and is trained on a massive dataset of known protein structures [40]. As more training data becomes available, AlphaFold2 can improve its predictions by recognizing more patterns from the test data. For example, membrane proteins are embedded in the cell membrane and are challenging to study because of their hydrophobicity, so as more membrane protein structures are solved, AlphaFold2 could have more representative data to use in training [58]. Furthermore, there are many other factors other than the amino acid sequence of a protein that determines a protein's structure, which is why AlphaFold2 should be used in conjunction with other algorithms to make accurate predictions.

Limitations in Applications

The limitations of AlphaFold2 are important for pharmaceutical development or industrial applications. If AlphaFold2 is used to predict the structure of a protein involved in a disease, and AlphaFold2 predicts it incorrectly, this could lead to the design of an ineffective drug that has unintended consequences and could waste resources. The standards for rational drug design are high: the predicted structure of a protein should be nearly identical to the experimental structure. This is essential to develop drugs and save large amounts of time and money invested in drug development. As a result, future research needs to address the current limitations of AlphaFold2 in predicting oncogenic proteins to allow for drug design and understanding of disease mechanisms.

Conclusion

Cancer is a group of diseases characterized by uncontrolled cell growth. Oncogenic proteins are proteins that can potentially cause cancer by disrupting normal cellular processes such as cell cycle regulation and DNA repair. AlphaFold2 is a machine-learning algorithm that can predict the 3D structure of proteins and is leading the AI revolution in biology. Finding the structure of oncogenic proteins is essential for understanding their function in cancer.

To determine the efficacy of AlphaFold2 in determining the structures of oncogenic proteins, we compared 26 oncogenic proteins found both experimentally and computationally using AlphaFold2. We used PyMOL, specifically the align command, to get the RMSD values for these proteins to see how well the AlphaFold2 model fit the experimentally found structure.

The RMSD values of the oncogenic proteins ranged from 0.204 Å to 1.980 Å, with an average of 0.627 Å. The results suggest that AlphaFold2 is a promising tool for predicting the 3D structure of oncogenic proteins, but future studies are needed to confirm and improve the accuracy of AlphaFold2's predictions for oncogenic proteins.

AlphaFold2 is a powerful tool for predicting the structure of oncogenic proteins. However, it has limitations in predicting the structure of highly disordered proteins, proteins with multiple conformations, mutated and artificial proteins. Future research is needed to address the current limitations of AlphaFold2 in predicting oncogenic proteins to allow for drug design and prevention methods.

AlphaFold2 can be improved by increasing the amount of training data available, improving the quality of available data, and using it in conjunction with other protein structure prediction models. Another future step could be developing a new algorithm and model to predict oncogenic protein interactions with other biomolecules, which would play a key role in understanding oncogenic proteins' role in cancer.

Acknowledgments

I would like to thank Dr. Daniel Plymire for helping me throughout the entire research process.

References

1. Institute of Medicine (US) Committee on Military Nutrition Research, Poos, M. I., Costello, R., & Carlson-Newberry, S. J. (1999). *Committee on Military Nutrition Research: Activity Report*. National Academies Press (US).

2. Dunham, B., & Ganapathiraju, M. K. (2021). Benchmark Evaluation of Protein-Protein Interaction Prediction Algorithms. *Molecules (Basel, Switzerland)*, *27*(1), 41. https://doi.org/10.3390/molecules27010041

3. Newberry, R. W., & Raines, R. T. (2019). Secondary Forces in Protein Folding. *ACS chemical biology*, *14*(8), 1677–1686. https://doi.org/10.1021/acschembio.9b00339

4. Kuhlman, B., & Bradley, P. (2019). Advances in protein structure prediction and design. *Nature reviews. Molecular cell biology*, *20*(11), 681–697. https://doi.org/10.1038/s41580-019-0163-x

5. Konc, J., & Janežič, D. (2022). Protein binding sites for drug design. *Biophysical reviews*, *14*(6), 1413–1421. https://doi.org/10.1007/s12551-022-01028-3

6. Bergquist, S., Otten, T., & Sarich, N. (2020). COVID-19 pandemic in the United States. *Health policy and technology*, *9*(4), 623–638. https://doi.org/10.1016/j.hlpt.2020.08.007

7. Msemburi, W., Karlinsky, A., Knutson, V., Aleshin-Guendel, S., Chatterji, S., & Wakefield, J. (2023). The WHO estimates of excess mortality associated with the COVID-19 pandemic. *Nature*, *613*(7942), 130–137. https://doi.org/10.1038/s41586-022-05522-2

8. Thorn, C. R., Sharma, D., Combs, R., Bhujbal, S., Romine, J., Zheng, X., Sunasara, K., & Badkar, A. (2022). The journey of a lifetime - development of Pfizer's COVID-19 vaccine. *Current opinion in biotechnology*, *78*, 102803. https://doi.org/10.1016/j.copbio.2022.102803

9. Wrapp, D., Wang, N., Corbett, K. S., Goldsmith, J. A., Hsieh, C. L., Abiona, O., Graham, B. S., & McLellan, J. S. (2020). Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science (New York, N.Y.)*, *367*(6483), 1260–1263. https://doi.org/10.1126/science.abb2507

10. Huang, Y., Yang, C., Xu, X. F., Xu, W., & Liu, S. W. (2020). Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. *Acta pharmacologica Sinica*, *41*(9), 1141–1149. https://doi.org/10.1038/s41401-020-0485-4

11. Harvey, W. T., Carabelli, A. M., Jackson, B., Gupta, R. K., Thomson, E. C., Harrison, E. M., Ludden, C., Reeve, R., Rambaut, A., COVID-19 Genomics UK (COG-UK) Consortium, Peacock, S. J., & Robertson, D. L. (2021). SARS-CoV-2 variants, spike mutations and immune escape. *Nature reviews. Microbiology*, *19*(7), 409–424. https://doi.org/10.1038/s41579-021-00573-0

12. Dill, K. A., Ozkan, S. B., Shell, M. S., & Weikl, T. R. (2008). The protein folding problem. *Annual review of biophysics*, *37*, 289–316. https://doi.org/10.1146/annurev.biophys.37.092707.153558

13. Anfinsen C. B. (1973). Principles that govern the folding of protein chains. *Science (New York, N.Y.)*, *181*(4096), 223–230. https://doi.org/10.1126/science.181.4096.223

14. Kaffe-Abramovich, T., & Unger, R. (1998). A simple model for evolution of proteins towards the global minimum of free energy. *Folding & design*, *3*(5), 389–399. https://doi.org/10.1016/S1359-0278(98)00052-2

15. Li, J., Hou, C., Ma, X., Guo, S., Zhang, H., Shi, L., Liao, C., Zheng, B., Ye, L., Yang, L., & He, X. (2021). Entropy-Enthalpy Compensations Fold Proteins in Precise Ways. *International journal of molecular sciences*, *22*(17), 9653. https://doi.org/10.3390/ijms22179653

16. Sorokina, I., Mushegian, A. R., & Koonin, E. V. (2022). Is Protein Folding a Thermodynamically Unfavorable, Active, Energy-Dependent Process?. *International journal of molecular sciences*, *23*(1), 521. https://doi.org/10.3390/ijms23010521

17. Dobson C. M. (2019). Biophysical Techniques in Structural Biology. *Annual review of biochemistry*, *88*, 25–33. https://doi.org/10.1146/annurev-biochem-013118-111947

18. Smyth, M. S., & Martin, J. H. (2000). x ray crystallography. *Molecular pathology: MP*, *53*(1), 8–14. https://doi.org/10.1136/mp.53.1.8

19. Marion D. (2013). An introduction to biological NMR spectroscopy. *Molecular & cellular proteomics : MCP*, *12*(11), 3006–3025. https://doi.org/10.1074/mcp.O113.030239

20. Murata, K., & Wolf, M. (2018). Cryo-electron microscopy for structural analysis of dynamic biological macromolecules. *Biochimica et biophysica acta. General subjects*, *1862*(2), 324–334. https://doi.org/10.1016/j.bbagen.2017.07.020

21. Holcomb, J., Spellmon, N., Zhang, Y., Doughan, M., Li, C., & Yang, Z. (2017). Protein crystallization: Eluding the bottleneck of X-ray crystallography. *AIMS biophysics*, *4*(4), 557–575. https://doi.org/10.3934/biophy.2017.4.557

22. AlQuraishi M. (2021). Machine learning in protein structure prediction. *Current opinion in chemical biology*, *65*, 1–8. https://doi.org/10.1016/j.cbpa.2021.04.005

23. Meller, A., Ward, M., Borowsky, J., Kshirsagar, M., Lotthammer, J. M., Oviedo, F., Ferres, J. L., & Bowman, G. R. (2023). Predicting locations of cryptic pockets from single protein structures using the PocketMiner graph neural network. *Nature communications*, *14*(1), 1177. https://doi.org/10.1038/s41467-023-36699-3

24. 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. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

25. Mihaly Varadi and others, AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models, *Nucleic Acids Research*, Volume 50, Issue D1, 7 January 2022, Pages D439–D444, https://doi.org/10.1093/nar/gkab1061

26. Marcu, Ş.-B., Tăbîrcă, S., & Tangney, M. (2022). An overview of Alphafold's breakthrough. *Frontiers in Artificial Intelligence*, *5*. https://doi.org/10.3389/frai.2022.875587

27. Bouatta, N., Sorger, P., & AlQuraishi, M. (2021). Protein structure prediction by AlphaFold2: are attention and symmetries all you need?. *Acta crystallographica. Section D, Structural biology*, *77*(Pt 8), 982–991. https://doi.org/10.1107/S2059798321007531

28. Higgins M. K. (2021). Can We AlphaFold Our Way Out of the Next Pandemic?. *Journal of molecular biology*, *433*(20), 167093. https://doi.org/10.1016/j.jmb.2021.167093

29. Matthews, H. K., Bertoli, C., & de Bruin, R. A. M. (2022). Cell cycle control in cancer. *Nature reviews. Molecular cell biology*, *23*(1), 74–88. https://doi.org/10.1038/s41580-021-00404-3

30. Siegel, R. L., Miller, K. D., Wagle, N. S., & Jemal, A. (2023). Cancer statistics, 2023. *CA: a cancer journal for clinicians*, *73*(1), 17–48. https://doi.org/10.3322/caac.21763

31. Ershler W. B. (2003). Cancer: a disease of the elderly. *The journal of supportive oncology*, *1*(4 Suppl 2), 5–10.

32. Disis, M. L., & Cheever, M. A. (1996). Oncogenic proteins as tumor antigens. *Current opinion in immunology*, *8*(5), 637–642. https://doi.org/10.1016/s0952-7915(96)80079-3

33. Williams, G. H., & Stoeber, K. (2012). The cell cycle and cancer. *The Journal of pathology*, *226*(2), 352–364. https://doi.org/10.1002/path.3022

34. Yip, H. Y. K., & Papa, A. (2021). Signaling Pathways in Cancer: Therapeutic Targets, Combinatorial Treatments, and New Developments. *Cells*, *10*(3), 659. https://doi.org/10.3390/cells10030659

35. Weinberg R. A. (1994). Oncogenes and tumor suppressor genes. *CA: a cancer journal for clinicians*, *44*(3), 160–170. https://doi.org/10.3322/canjclin.44.3.160

36. Rudiman, R., Wijaya, A., Sribudiani, Y., Soedjana, H. S., Wiraswati, H. L., Pramaswati, E., Nugraha, P., & Lukman, K. (2023). Identification of KRAS mutation and HER2 expression in Indonesian colorectal cancer population: a cross-sectional study. *Annals of medicine and surgery (2012)*, *85*(5), 1761–1768. https://doi.org/10.1097/MS9.0000000000000694

37. Ramus, S. J., Bobrow, L. G., Pharoah, P. D., Finnigan, D. S., Fishman, A., Altaras, M., Harrington, P. A., Gayther, S. A., Ponder, B. A., & Friedman, L. S. (1999). Increased frequency of TP53 mutations in BRCA1 and BRCA2 ovarian tumours. *Genes, chromosomes & cancer*, *25*(2), 91–96. https://doi.org/10.1002/(sici)1098-2264(199906)25:2<91::aid-gcc3>3.0.co;2-5

38. Batool, M., Ahmad, B., & Choi, S. (2019). A Structure-Based Drug Discovery Paradigm. *International journal of molecular sciences*, *20*(11), 2783. https://doi.org/10.3390/ijms20112783

39. An, X., Tiwari, A. K., Sun, Y., Ding, P. R., Ashby, C. R., Jr, & Chen, Z. S. (2010). BCR-ABL tyrosine kinase inhibitors in the treatment of Philadelphia chromosome positive chronic myeloid leukemia: a review. *Leukemia research*, *34*(10), 1255–1268. https://doi.org/10.1016/j.leukres.2010.04.016

40. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic acids research*, *28*(1), 235–242. https://doi.org/10.1093/nar/28.1.235

41. The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.

42. Kufareva, I., & Abagyan, R. (2012). Methods of protein structure comparison. *Methods in molecular biology (Clifton, N.J.)*, *857*, 231–257. https://doi.org/10.1007/978-1-61779-588-6_10

43. Goodsell D. S. (2005). Visual methods from atoms to cells. *Structure (London, England : 1993)*, *13*(3), 347–354. https://doi.org/10.1016/j.str.2005.01.012

44. Lee, C., Su, B. H., & Tseng, Y. J. (2022). Comparative studies of AlphaFold, RoseTTAFold and Modeller: a case study involving the use of G-protein-coupled receptors. *Briefings in bioinformatics*, *23*(5), bbac308. https://doi.org/10.1093/bib/bbac308

45. Castel, P., Rauen, K. A., & McCormick, F. (2020). The duality of human oncoproteins: drivers of cancer and congenital disorders. *Nature reviews. Cancer*, *20*(7), 383–397. https://doi.org/10.1038/s41568-020-0256-z

46. Perrakis, A., & Sixma, T. K. (2021). AI revolutions in biology: The joys and perils of AlphaFold. *EMBO reports*, *22*(11), e54046. https://doi.org/10.15252/embr.202154046

47. Mou, Y., Huang, P. S., Hsu, F. C., Huang, S. J., & Mayo, S. L. (2015). Computational design and experimental verification of a symmetric protein homodimer. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(34), 10714–10719. https://doi.org/10.1073/pnas.1505072112

48. Drake, Z.C., Seffernick, J.T. & Lindert, S. Protein complex prediction using Rosetta, AlphaFold, and mass spectrometry covalent labeling. *Nat Commun* 13, 7846 (2022). https://doi.org/10.1038/s41467-022-35593-8

49. Ren, F., Ding, X., Zheng, M., Korzinkin, M., Cai, X., Zhu, W., Mantsyzov, A., Aliper, A., Aladinskiy, V., Cao, Z., Kong, S., Long, X., Man Liu, B. H., Liu, Y., Naumov, V., Shneyderman, A., Ozerov, I. V., Wang, J., Pun, F. W., Polykovskiy, D. A., … Zhavoronkov, A. (2023). AlphaFold accelerates artificial intelligence powered drug discovery: efficient discovery of a novel CDK20 small molecule inhibitor. *Chemical science*, *14*(6), 1443–1452. https://doi.org/10.1039/d2sc05709c

50. Keskin Karakoyun, H., Yüksel, Ş. K., Amanoglu, I., Naserikhojasteh, L., Yeşilyurt, A., Yakıcıer, C., Timuçin, E., & Akyerli, C. B. (2023). Evaluation of alphafold structure-based protein stability prediction on missense variations in cancer. *Frontiers in Genetics*, *14*. https://doi.org/10.3389/fgene.2023.1052383

51. Hekkelman, M. L., de Vries, I., Joosten, R. P., & Perrakis, A. (2023). AlphaFill: enriching AlphaFold models with ligands and cofactors. *Nature methods*, *20*(2), 205–213. https://doi.org/10.1038/s41592-022-01685-y

52. Ruff, K. M., & Pappu, R. V. (2021). AlphaFold and Implications for Intrinsically Disordered Proteins. *Journal of molecular biology*, *433*(20), 167208. https://doi.org/10.1016/j.jmb.2021.167208

53. Hou, M.-H., Jin, S.-R., Cui, X.-Y., Peng, C.-X., Zhao, K.-L., Song, L., & Zhang, G.-J. (2023). Protein Multiple Conformations Prediction Using Multi-Objective Evolution Algorithm. bioRxiv 2023.04.21.537776 https://doi.org/10.1101/2023.04.21.537776

54. Ahmed, M. H., Ghatge, M. S., & Safo, M. K. (2020). Hemoglobin: Structure, Function and Allostery. *Sub-cellular biochemistry*, *94*, 345–382. https://doi.org/10.1007/978-3-030-41769-7_14

55. Bryant, P., Pozzati, G., Zhu, W. *et al.* Predicting the structure of large protein complexes using AlphaFold and Monte Carlo tree search. *Nat Commun* 13, 6028 (2022). https://doi.org/10.1038/s41467-022-33729-4

56. Trinkle-Mulcahy, L., Boulon, S., Lam, Y. W., Urcia, R., Boisvert, F.-M., Vandermoere, F., Morrice, N. A., Swift, S., Rothbauer, U., Leonhardt, H., & Lamond, A. (2008). Identifying specific protein interaction partners using quantitative mass spectrometry and Bead Proteomes. *The Journal of Cell Biology*, *183*(2), 223–239. https://doi.org/10.1083/jcb.200805092

57. Buel, G. R., & Walters, K. J. (2022). Can alphafold2 predict the impact of missense mutations on structure? *Nature Structural & Molecular Biology*, *29*(1), 1–2. https://doi.org/10.1038/s41594-021-00714-2

58. Hegedűs, T., Geisler, M., Lukács, G. L., & Farkas, B. (2022). Ins and outs of AlphaFold2 transmembrane protein structure predictions. *Cellular and molecular life sciences : CMLS*, *79*(1), 73. https://doi.org/10.1007/s00018-021-04112-1