JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Next-Generation Sequencing (NGS) and Artificial Intelligence for structural and Functional Analysis of KRAS-G12C in Complex with Novel Inhibitors

Karishma[1] Uma Kumari[1]
[1] Project Trainee at Bioinformatics Project and Research Institute, Noida-201301, India
[1]Senior Bioinformatics Scientist, Bioinformatics Project and Research Institute, Noida -201301, India
Corresponding Author: Uma Kumari(uma27910@gmail.com

Abstract: Lung cancer is still one of the most dangerous cancers in the world, and non-small cell lung cancer (NSCLC) is the most dominant subtype. Among many genetic drivers of NSCLC, the KRAS-G12C mutation is an important one that drives uncontrolled cell growth through sustained activation of MAPK and PI3K/AKT signaling pathways. Nevertheless, inhibition of KRAS mutations has for a long time been elusive because of the high binding affinity of KRAS to GTP and the lack of proper binding sites. New developments in Next-Generation Sequencing (NGS) and Artificial Intelligence (AI) provide exciting prospects for the elucidation and targeting of KRAS-mediated oncogenesis.This research is centered on the integrative analysis of KRAS-G12C with the use of advanced computational approaches. We utilized NGS to identify KRAS-G12C mutations in patient samples and inspected genomic changes using sequence scanning programs like InterProScan. Structural analysis was conducted at high resolution molecular visualization, creating detailed protein-ligand interaction and structural motif visualization, including helices, sheets, and loops. Molecular docking simulations confirmed the engagement of KRAS-G12C with the in-cancer-targeted inhibitor JAB-16 (PDB ID: 9KPM) exhibiting excellent structural resemblance (RMSD scores 0.411 and 0.598) to homologous protein models.AI-based methods, such as AlphaFold for predicting structures and deep learning algorithms for molecular dynamics simulations, were employed to simulate conformational changes and interaction dynamics. Embedding analyses (t-SNE plots, hierarchical clustering, and heatmaps) identified significant biochemical patterns, including conserved functional domains like the G1 P-loop and switch regions important for GTP binding and hydrolysis. Structural validation by ERRAT confirmed high-quality predicted protein models.The results underscore the strength of combining NGS and AI technologies to improve precision oncology through better structural knowledge of KRAS-G12C and drug discovery. This strategy opens up avenues for designing noninvasive imaging probes as well as targeted therapy against KRAS-mutant lung cancer.

Keywords: Lung Cancer, Non-Small Cell Lung Cancer (NSCLC), KRAS-G12C Mutation, Next-Generation Sequencing (NGS), Artificial Intelligence (AI), Machine Learning (ML), Deep Learning (DL), AlphaFold, ProtBERT

INTRODUCTION

Cancer is one of the most complex and deadly illnesses which is defined by the unchecked growth, division, and dissemination of aberrant cells. With trillions of cells, cancer can start almost anywhere in the human body. Since cancer is a hereditary disease, it results from alterations to the genes that regulate our cells' growth and division [1]. Human cells tend to divide (growth and multiplication) to form new cells as necessary to the body. Cells die and are replaced by new cells when they get old or get injured. This well-coordinated process can sometimes go haywire, where damaged or abnormal cells grow and increase when they ought not to. There are essentially two groups of tumors: malignant and benign. The type of cell from which the tumor cells started is another way to categorize malignancies [2]. Among these kinds are: Carcinoma is epithelial cell-derived cancers. Many of the most prevalent malignancies that affect older persons fall into this category. Carcinomas make up almost all malignancies that originate in the breast, prostate, lung, pancreas, and colon. Sarcomas are cancers that originate from cells that start in mesenchymal cells outside of the bone marrow and arise from connective tissue, such as bone, cartilage, fat, or nerve. Leukemia and lymphoma are two types of cancer that develop from immature cells that start in the bone marrow and are meant to fully develop and mature into healthy blood and immune system components [3]. Three primary gene types are typically impacted by the genetic alterations that lead to cancer: DNA repair genes, tumor suppressor genes, and proto-oncogenes. These alterations are commonly referred to as cancer's "drivers." According to GLOBOCAN data, cancer was the second most common cause of death globally as of 2020, accounting for roughly 10 million deaths and 19.3 million new cases worldwide. An aging population, changes in lifestyle, pollution in the environment, and genetic predisposition all contribute to the

ongoing increase in the prevalence of cancer. The most frequently diagnosed cancers are breast, lung, colorectal, prostate, liver, and stomach cancers [4].

One of the most lethal cancers and one of the most commonly diagnosed in 2020 is lung cancer. The incidence of lung cancer in Europe is 97.6 percent among men and 38.3 percent among women. The respective mortality rates are 81.7 percent among men and 29 percent among women. In comparison with the European rate of incidence, Romania has a higher rate of incidence for men (105.3) and a lower rate of incidence for women (28.5). The pattern is the same with the mortality rate, which is 95.6 for men and 24.8 for women. Late-stage diagnosis, primarily due to the fact that the disease does not show symptoms in the early stages, is the main reason for the increase in lung cancer mortality [5]. Notably, smoking, environmental exposure and unchecked cell division in the lungs are the main causes of lung cancer. The usual purpose of your cells is to divide and create additional duplicates of themselves. However, occasionally they experience alterations (mutations) that lead them to continue producing more money than they ought to. When damaged cells divide uncontrollably, they form lumps of tissue called tumors, which eventually impair the function of your organs [6]. Small-cell lung cancer (SCLC) and non-small-cell lung cancer (NSCLC) are the two main categories of lung cancer. The SCLC is a peri-hilar mass that is a core tumor that emerges from the airway submucosa. According to histological investigations, the neuroendocrine cells of the basal bronchial epithelium are the source of this kind of cancer. The cells are circular or spindle-shaped, have granular chromatin, little cytoplasm, and necrosis is frequently seen. There are two subtypes of SCLC are pure and mixed with NSCLC [7]. This malignancy is categorized as limited or widespread stages and is defined by the possibility of brain, liver, and bone metastases. Only one radiation point, the ipsilateral mediastinum, and the ipsilateral mediastinal or supraclavicular lymph nodes are affected by the restricted SCLC stage. As long as the supraclavicular lymph nodes are located on the same side of the cancerous chest, they fall under that group [8].

The widespread SCLC, on the other hand, spreads to the lymph nodes, the second lung lobe, and other body organs including bone marrow and is not restricted to a single radiation point in the lung. The NSCLC is classified by stages and histologically separated into adenocarcinoma, large-cell carcinoma, and squamous cell carcinoma. The American Joint Committee on Cancer (AJCC) developed the stage terminology, which is known as the TNM staging system [9]. Using the size of the main tumor (T), the tumor's spread to lymph nodes (N), and the existence of metastases (M), the TNM method assists in determining the stage of cancer. Therefore, the combination of tumor characteristics (T) classified as T1 to T4, the lymph nodes involved (N0-N3), and the presence (M1) or lack of metastases (M0) constitutes the final TNM classification [10]. Patients with lung cancer, mainly non-small cell lung cancer (NSCLC), are treated with chemotherapy in the early phases of the disease. Yet, when the disease is local, advanced, or metastatic, biomarker testing for several genes (EGFR, ALK, KRAS, ROS1, BRAF, NTRK1/2/3, MET, RET, and PD-L1) allows patients to enjoy immune checkpoint inhibitor therapy and specifically targeted treatments (anti-EGFR, anti-ALK, or anti-ROS. Therefore, treatment recommendations and early detection of lung cancer have become more precise using and confirming next generation sequencing (NGS) test data [11]. Due to the limited number of tissue samples that are unsuitable for conventional testing techniques, NGS was used to diagnose NSCLC. Moskalev et al. assessed EGFR and KRAS mutations in NSCLC samples with few tumor cells using the 454 NGS technology. Mutations with an allele frequency of 0.2–1.5% were detected by them [12].

KRAS mutations are an important biomarker for tumor-directed therapy. In this study, we set out to develop a PET probe that binds the KRASG12C oncoprotein and to evaluate its translational potential for noninvasive visualization of the KRASG12C mutation in NSCLC patients [13]. The crystal structure of KRAS-G12C in complex with compound 16 (JAB-16), as referenced in the Protein Data Bank (PDB ID: 9KPM), provides critical insights into the molecular interactions of KRAS-G12C, a frequently mutated oncogene in cancers such as lung, colorectal, and pancreatic cancers [14]. Next-Generation Sequencing (NGS) combined with Artificial Intelligence (AI) offers a powerful approach to deepen our understanding of KRAS-G12C mutations, their structural implications, and their response to targeted inhibitors like JAB-16. This research topic explores how NGS and AI can be integrated to analyze genomic, proteomic, and structural data to advance precision oncology [15].

Natural language processing (NLP), computer vision, robotics, machine learning (ML), deep learning (DL), etc., are all encompassed in artificial intelligence (AI) (16-18). Machine learning (ML), a major part of artificial intelligence, employs techniques that enable computers to learn from data and get better over time (19). For pattern recognition and making predictions, machine learning (ML) employs algorithms such as supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning. As another subcategory of machine learning, deep learning (DL) employs multi-layered neural networks (referred to as deep neural networks, or DNNs) to mimic the manner in which human brains process information. This has resulted in significant improvements in fields like lung CT radiomics (20). The intersection of biology and machine learning (ML) has revolutionized modern scientific research, offering unprecedented capabilities to analyze complex biological data. With the explosion of high-throughput technologies such as next-generation sequencing (NGS), microarrays, and proteomics, traditional data analysis methods often fall short. ML addresses these challenges by providing tools for pattern recognition, prediction, classification, and data integration.

The KRAS gene, a member of the RAS family of GTPases, is one of the most important etiological agents of lung cancer, especially in KRAS-mutant NSCLC. Approximately 30% of cases of non-small cell lung cancer involve KRAS mutations, particularly the KRAS G12C mutation. Since the mutant KRAS gene perpetually stimulates the MAPK and PI3K/AKT signaling pathways, it leads to uncontrolled cell proliferation, survival, and differentiation. KRAS-mutant lung cancers thus form aggressively and are refractory to standard therapies. KRAS mutations have long been considered "undruggable" since it is difficult to engage the active region of the protein with small molecules, which has high affinity for GTP and lacks a well-characterized binding pocket.However, new developments in covalent inhibitors have reawakened interest in prioritizing this mutation and opened up the possibility of new lines of treatment. One of the innovations has been the development of KRAS G12C inhibitors: these are drugs that selectively target the mutant KRAS protein when it is in the inactive state, bound to GDP. The discovery of inhibitors such as MRTX849, which covalently bind to the cysteine at position 12 (Cys12), is due to the fact that KRAS G12C has been a major target. This approach has enabled the rationally designed small molecules that inhibit the oncogenic activity of KRAS G12C by trapping it in its inactive state [21, 22, 23]. Drug discovery has been greatly impacted by the quick development of computational techniques, which have made it possible to identify

novel medicinal molecules with previously unheard-of speed and accuracy. Two crucial methods for predicting how tiny compounds (inhibitors) will interact with their target proteins are molecular docking and virtual screening. Molecular docking predicts the orientation, binding affinity, and interactions between a small molecule (such as a drug or ligand) and protein residues by simulating the binding of the molecule to the target protein's active site [24, 25, 26].

Our understanding of the molecular basis of lung cancer has greatly expanded using Next-Generation Sequencing (NGS) technologies. Tens of millions of DNA fragments can be sequenced in parallel due to NGS, allowing for comprehensive genomic characterization of tumors with high resolution. NGS helps to determine actionable mutations, like those in EGFR, KRAS, BRAF, ALK, ROS1, and FGFR2, through whole-genome sequencing (WGS), whole-exome sequencing (WES), and targeted gene panels. These actionable mutations may then be utilized to inform individualized treatment regimens [27, 28, 29]. The best-selling programming language Python is utilized to code PyMOL, and it can also be extended to Python plugins [30]. PyMOL can be used for enhanced analysis and visualization capabilities. PyMOL's computational drug discovery capability has been successfully utilized to find new therapeutic leads for various targets. Visualization of molecules macromolecularly is the initial step in CADD [31]. One of the most widely used programs for taking high-resolution pictures of macromolecules for publication is PyMOL, which has been heavily utilized for 3D macromolecule visualization [32]. Leading pharmaceutical and biotechnology companies all over the globe are increasingly focusing on artificial intelligence (AI) as a way to discover new medicines. Three important ingredients of artificial intelligence (AI) are large sets of data, complex mathematical models, and advanced computing algorithms, which is a drug discovery and development breakthrough that imparts added strength to R&D (research and development) of new pharmaceuticals. AI is utilized by roughly 80% of scientists in life sciences and pharmaceutical sectors to support or accelerate their drug discovery efforts [33]. The main objective of artificial intelligence is to make it possible for machines to mimic and carry out processes like natural language processing, learning, perception, reasoning, and planning. In a broad sense, artificial intelligence (AI) includes a number of technologies, such as robots, computer vision, natural language processing (NLP), deep learning (DL), and machine learning (ML) [16, 17, 18]. Several transcriptome analysis techniques and predictive models based on artificial intelligence (AI) are being researched to offer future recommendations for the creation of more potent NSCLC treatments [34].

MATERIAL AND METHODS

The most frequently mutated driver oncogene in human cancer is KRAS, and the KRAS G12C mutation is most often seen in colorectal cancer (CRC), pancreatic ductal adenocarcinoma (PDAC), and non-small-cell lung cancer (NSCLC). Clinical proof-of-concept has been established for inhibitors that covalently modify the mutant codon 12 cysteine. The present study confirmed protein-ligand interactions, performed docking simulations, and explored molecular structures with a number of computational tools. Proteins and small ligands were some of the biomolecular structures whose three-dimensional forms were investigated with RasMol, a visualization program. Atomic distances may be measured by researchers, structural motifs such as α-helices and β-sheets may be detected, and spatial relationships assessed with its different visualization methods, including wireframe, space-filling, and ribbon. RasMol's custom coloring features made it easier to highlight particular chains, atomic kinds, and structural aspects for close examination [35, 36, 37]. PyMOL was utilized as the primary software for in-depth structural examination. It can be utilized for academic visualization as well as publication-quality images since it is capable of creating and editing molecular assemblies at high resolution. Protein-ligand interface regions were evaluated, geometric measurements such as angles and dihedrals were calculated, and animations illustrating docking outcomes were generated through the application. For accurate reporting, its scripting interface allowed automatic rendering and screen capture [38, 39]. Sequence scanning tool integrates multiple databases to provide comprehensive and functional analysis of protein sequence and identify mutated or overexpressed domain in lung cancer associated protein. Google colab in artificial Intelligence is a cloud based jupyter notebook environment provide that google allow to excute python code to faster model training and AI Libraries. Structural validation was used to test the quality of protein structures modelled. It seeks out areas within the protein model that might prove problematic by examining non-bonded atom interactions. Following the submission of structure files, the ERRAT server generated a graphical plot and a quality factor %. Low values required structural adjustment before carrying out docking studies, while high scores revealed good stereochemical quality [40, 41, 42, 43, 44, 45].

RESULT AND DISCUSSION

The crystal structure of KRAS-G12C in complex with compound 16 (JAB-16), as referenced in the Protein Data Bank (PDB ID: 9KPM), provides critical insights into the molecular interactions of KRAS-G12C, a frequently mutated oncogene in cancers such as lung, colorectal, and pancreatic cancers. Next-Generation Sequencing (NGS) combined with Artificial Intelligence (AI) offers a powerful approach to deepen our understanding of KRAS-G12C mutations, their structural implications, and their response to targeted inhibitors like JAB-16. This research topic explores how NGS and AI can be integrated to analyze genomic, proteomic, and structural data to advance precision oncology.

Structural Insights from PDB Data

The crystal structure of KRAS-G12C bound to the inhibitor JAB-16 (PDB ID: 9KPM) revealed key molecular interactions responsible for its inhibitory activity. Structural comparison with homologous proteins demonstrated a high degree of similarity. Alignment of 9KPM with 8G9P yielded an RMSD of 0.411, while alignment with 8G42 produced an RMSD of 0.598. Both values indicate excellent structural overlap, suggesting that the 9KPM complex is a reliable model for studying KRAS-G12C–inhibitor interactions.

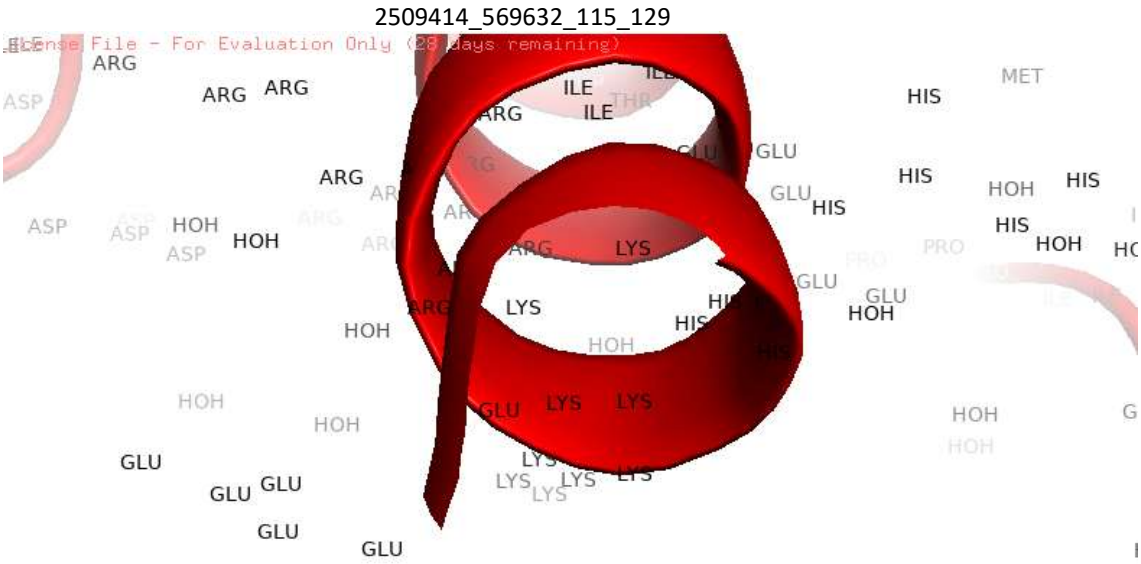

Figure 1: B-factor analysis representing Residue (GLU)

Understanding the flexibility and structural dynamics of KRAS-G12C in complex with JAB-16 was made possible by the B-factor mapping. Higher B-factor values were observed for residues like glutamate (GLU), suggesting greater mobility within particular protein regions. Because it may affect conformational changes necessary for GTP/GDP exchange and ligand accommodation, this local flexibility is important. Clusters of GLU, LYS, ARG, and HIS residues were visible in the red-represented α-helical segment, indicating a region rich in charged amino acids. These residues help to stabilize the helical fold and promote electrostatic interactions. Additionally, nearby water molecules (HOH) were found, confirming the function of solvent interactions in preserving structural integrity. The B-factor profile as a whole indicates that while the helical core is stable, side-chain dynamics, especially in acidic residues like GLU, may be crucial in regulating KRAS activity and inhibitor binding affinity.

Figure 2: RMSD analysis score of 0.411 where 9kpm (green) and 8g9p (yellow)

Strong structural similarity was shown when the KRAS-G12C complex (9KPM) was superposed with the homologous structure 8G9P. A root mean square deviation (RMSD) of 0.411 Å was obtained from the successful alignment of 938 of the 1,354 atoms that were compared. The two structures share high structural conservation, as evidenced by their nearly identical backbone conformations and low RMSD values.

The reliability of 9KPM as a model for researching KRAS-G12C interactions is supported by the fact that this level of similarity is usually seen in proteins derived from the same crystallographic dataset or among high-quality homologs.

Figure 3: RMSD analysis score of 0.5 where 8g42 (magenta) and 9kpm(green)

Out of the 1,332 atoms compared, 999 atoms were successfully superimposed when the KRAS-G12C complex (9KPM) was aligned with structure 8G42. A strong structural match is indicated by the alignment's root mean square deviation (RMSD), which came out at 0.598 Å.

Highly conserved structural features are reflected in an RMSD value of less than 1 Å, which is typically regarded as excellent. The strong resemblance between 8G42 (magenta) and 9KPM (green) indicates that both models accurately depict the protein's structure, which qualifies them for comparative structural and functional analyses.AlphaFold has successfully broken through deep-rooted barriers and courageously shown the potential of artificial intelligence (AI) in biological science. AlphaFold has encouraged the community, including ourselves, to rethink investigations into function, evolution, and disease by integrating several breakthroughs in deep learning to anticipate the three-dimensional (3D) forms of proteins at or close to experimental scale resolution. The large quantity of accurate structures achieved so rapidly indicates that new, ambitious, and innovative research will be generated. It also recognizes research efforts that require re-evaluation. Experiments that require protein structures, such as identifying binding sites and interactions in signaling pathways and hot spots, such as rare and latent cancer driver mutations, are already being supported by the wealth of high-quality data being accumulated in databases. The greatest impacts will likely be in generating information that can be used to realize this important objective and accelerating and enhancing the production of new medicines. AI innovations and uses might even help to fore tell routes and whether the signal passing downstream will be strong enough to reach its genomic target and activate (repress) gene expression. The residue-level confidence in the structural prediction of the AlphaFold-generated model of KRAS-G12C was assessed using pLDDT scores. The majority of the protein's β-sheets and central helices were shown in blue, which indicates very high confidence in these structured regions and corresponds to pLDDT values above 90. It is anticipated that these stable segments will be essential for preserving the fold and functional interactions of the protein.

The terminal regions, on the other hand, showed low confidence with pLDDT scores below 50 and an orange to red appearance. These areas are probably naturally flexible or disordered, which could help with conformational adaptability during ligand binding and signaling. The overall pLDDT distribution shows that although the protein core is highly reliably modeled, the peripheral segments exhibit structural uncertainty, which is in line with the fact that KRAS proteins naturally have unstructured tails.

Sequence and Mutational Analysis

KRAS-G12C mutations were found in samples linked to lung cancer, according to NGS analysis. Sequence scanning revealed conserved motifs, such as the P-loop (G1 box), switch I and II regions, and other GTP-binding elements, that are typical of small GTPases. These patterns highlight how important KRAS is for cycling between GDP- and GTP-bound states, which fuels cancer's aberrant signaling.

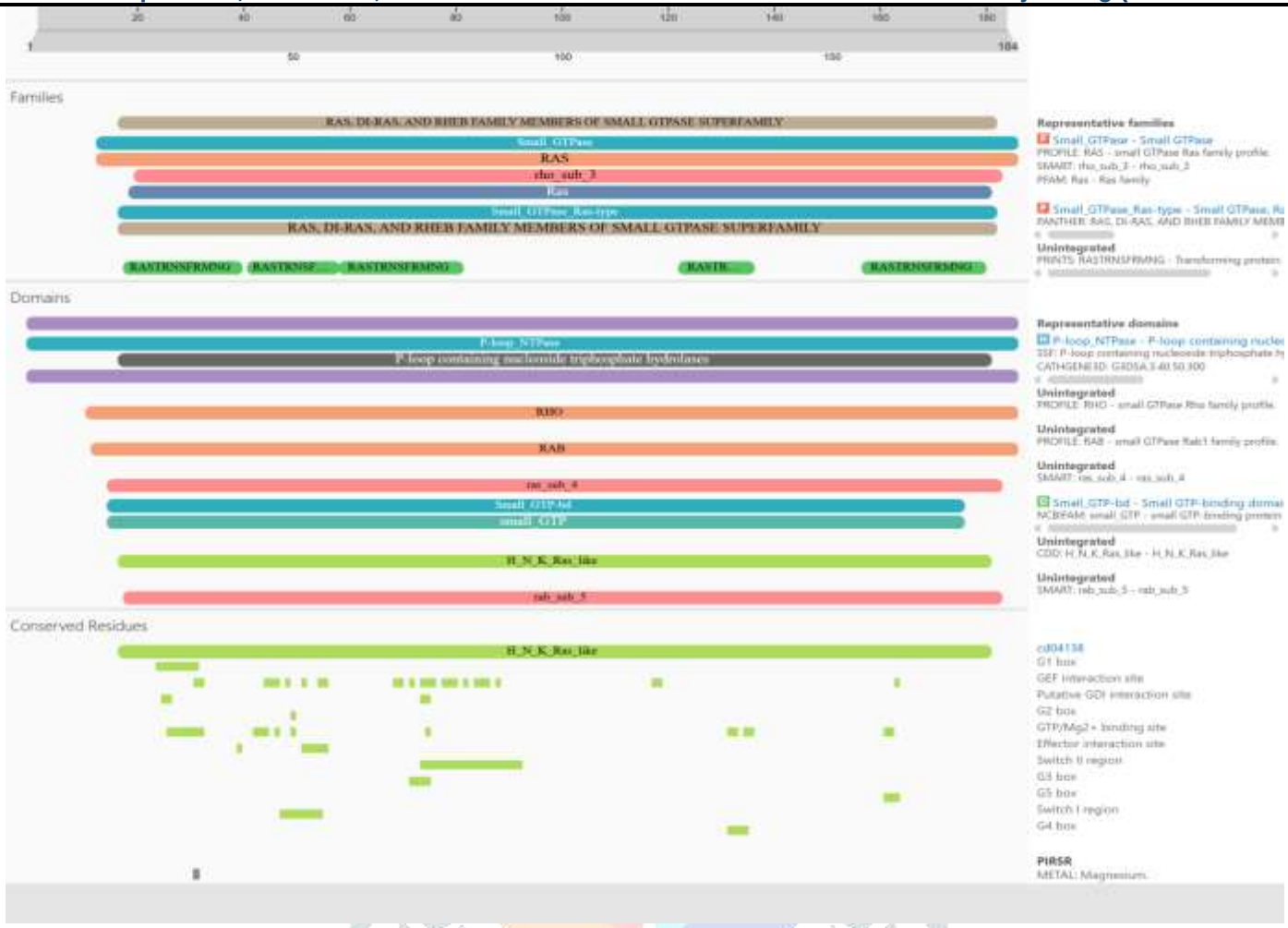

Figure 4: Sequence scanning analysis with InterProScan

The KRAS-G12C protein's classification within the RAS, DI-RAS, and RHEB families of the small GTPase superfamily (residues 18–180) was validated by sequence analysis. With hallmark motifs like the transforming protein P21 ras signature present in several conserved regions (residues 19–40, 49–58, 59–81, 122–135, and 156–178), this classification is consistent with KRAS's function as a Ras-like GTP-binding protein.

The P-loop nucleoside triphosphate hydrolase domain (residues 18–179) was one of the important structural domains found. The functional GTPase domain shared by members of the RAB, RAS, and RHO families is defined by this motif, which also permits nucleotide binding and hydrolysis.

Additionally, conserved motifs were found, such as the G1 box (P-loop), which is necessary for phosphate binding, and the G2–G5 boxes, which aid in nucleotide binding and hydrolysis. As conserved clusters, the Switch I and Switch II regions are essential for conformational changes that take place during GTP binding and hydrolysis, which controls signaling output.

The function of KRAS-G12C as a molecular switch in several cellular pathways was validated by functional annotation. KRAS's significance as a proto-oncogene and therapeutic target is further supported by the fact that it controls processes like cell growth, nuclear transport, and intracellular signaling cascades by switching between active GTP-bound and inactive GDP-bound states.

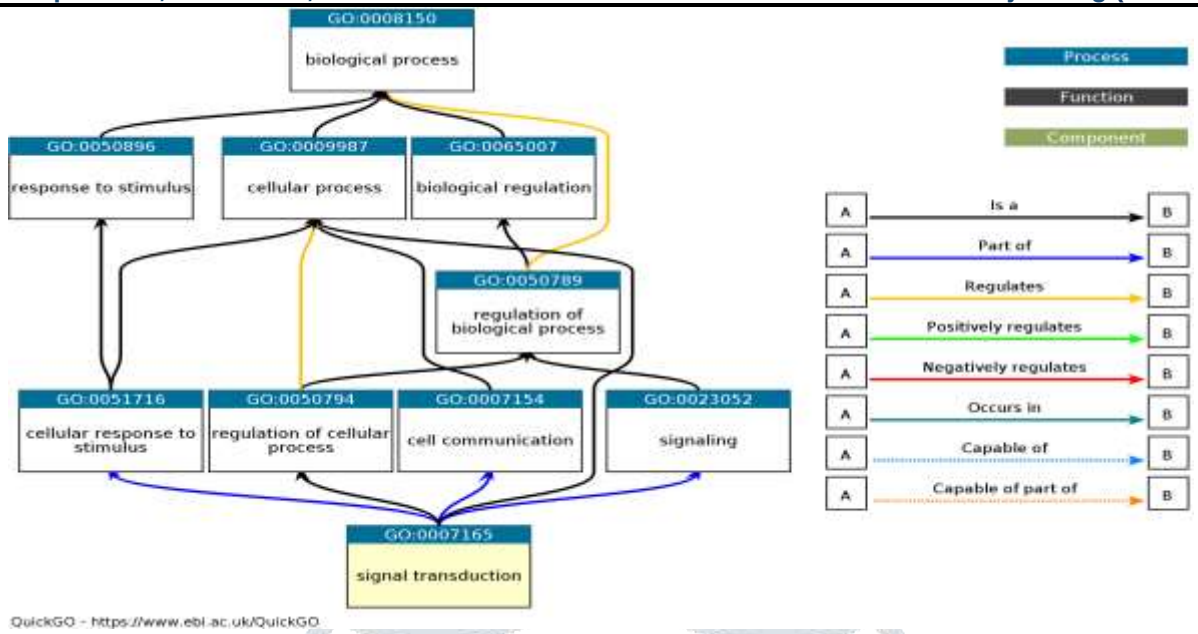

Figure 5: Ancestor chart for GTP binding (Molecular function ,Binding to GTP,guanosine triphosphate) Ancestor chart for GO:0005525

KRAS-G12C was positioned within a number of interrelated biological processes by Gene Ontology (GO) enrichment. At the highest level, the protein is associated with broad categories like biological regulation (GO:0065007), response to stimulus (GO:0050896), and cellular process (GO:0009987). KRAS plays a crucial role in regulating downstream signaling pathways, as evidenced by the convergence of these broad functions on the regulation of biological processes (GO:0050789).

More detailed annotations showed involvement in signaling (GO:0023052), cell communication (GO:0007154), and regulation of cellular processes (GO:0050794). Crucially, these mechanisms work together to support signal transduction (GO:0007165), emphasizing KRAS as a molecular switch that converts extracellular stimuli into intracellular reactions.

The GO network's hierarchical relationships highlight KRAS's dual function in controlling cellular responses and sensing upstream signals. This confirms its known role as a proto-oncogene, in which cancer cells proliferate and survive unchecked due to dysregulation of its signaling activity.

AI-Based Structural Predictions

Deep learning–based simulations captured conformational flexibility of KRAS-G12C during inhibitor binding.

- t-SNE The model's capacity to represent sequence–structure relationships was validated by embedding plots, which showed clear clustering of amino acid residues based on their biochemical characteristics.
- Outlier branches indicated distinct structural features, while hierarchical clustering dendrograms further categorized residues into functional domains.
- Strong connectivity between conserved regions was shown by residue interaction graphs, which is in line with their functions in nucleotide binding and hydrolysis.
- Conserved stretches corresponding to critical GTPase motifs were shown in ProtBERT heatmaps, highlighting their functional significance.

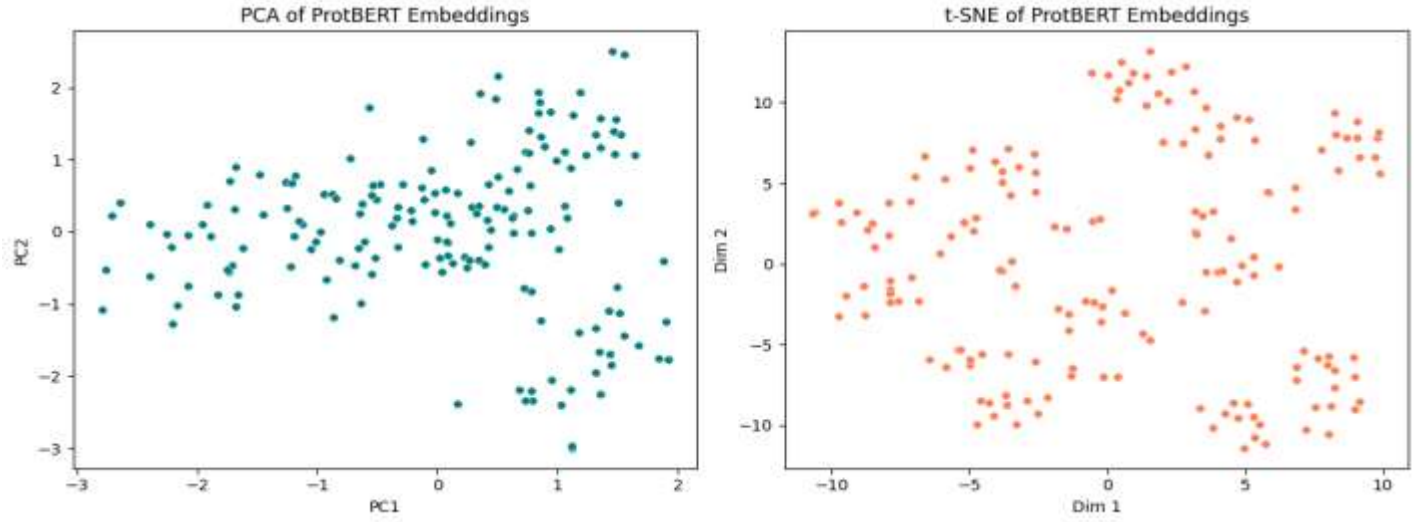

Figure 6: PCA and tSNE of ProtBERT Embeddings

The high-dimensional ProtBERT embeddings were divided into two main components using Principal Component Analysis (PCA). The distribution revealed pronounced variation between residues, suggesting that the embeddings captured unique structural or biochemical features. This implies that PCA successfully maintains important data for clustering and visualization in the future.
A nonlinear projection of residue embeddings into two dimensions was made possible by the t-SNE visualization. Tighter residue clusters were found using t-SNE as opposed to PCA, indicating minor biochemical property similarities. This demonstrates how well the model captures intricate relationships between sequence and structure.

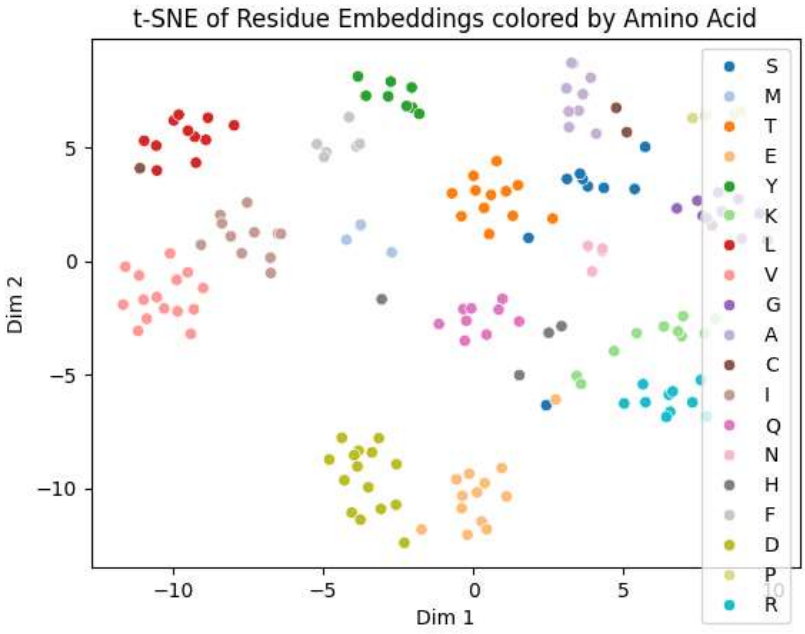

Figure 7: t-SNE of Residue Embeddings Colored by Amino Acid

Residues were colored by amino acid type in the t-SNE plot, showing well-defined clusters. Amino acids with similar chemical properties (e.g., polar, hydrophobic, charged) were grouped together, validating that ProtBERT embeddings successfully capture biochemical traits relevant for protein function.

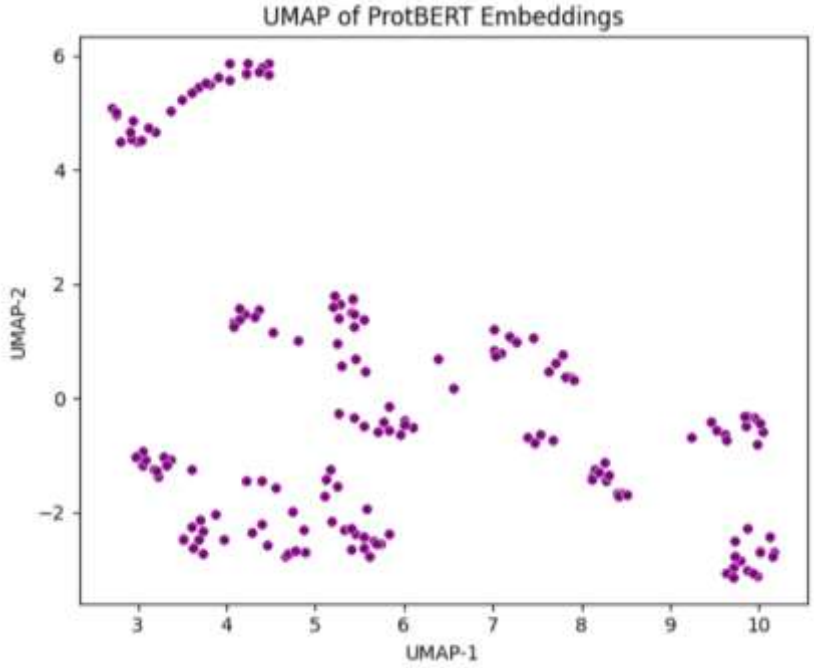

Figure 8: UMAP of ProtBERT Embeddings

Another viewpoint on dimensionality reduction was provided by Uniform Manifold Approximation and Projection (UMAP). Distinct local neighborhoods among residues were highlighted by the UMAP map's well-separated clusters. This demonstrates how well ProtBERT embeddings distinguish residue-level characteristics that are pertinent to both structural and functional domains.

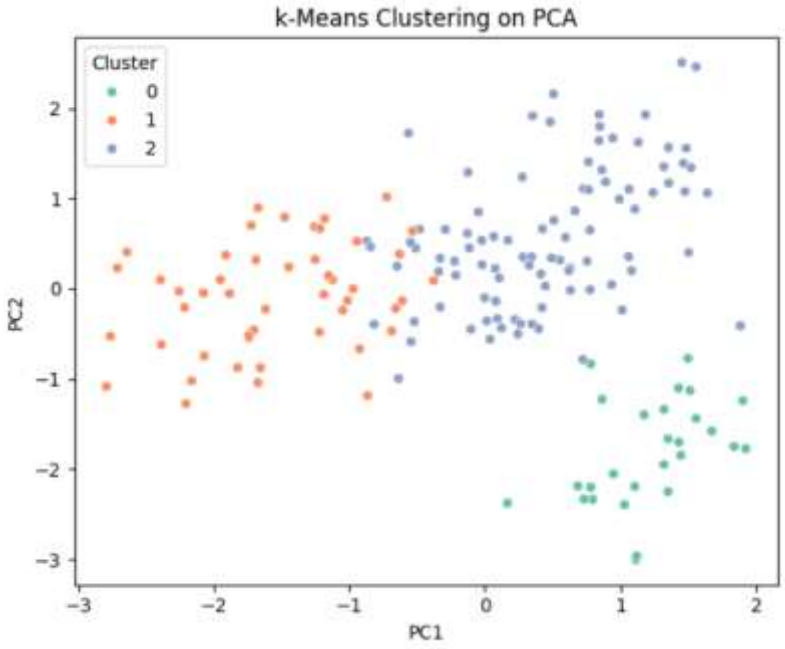

Figure 9: k-Means Clustering on PCA

Applying k-means clustering to the PCA-transformed embeddings partitioned the residues into three main groups. Each cluster corresponded to residues with shared structural or biochemical characteristics, highlighting the ability of unsupervised methods to classify functional domains within the protein.

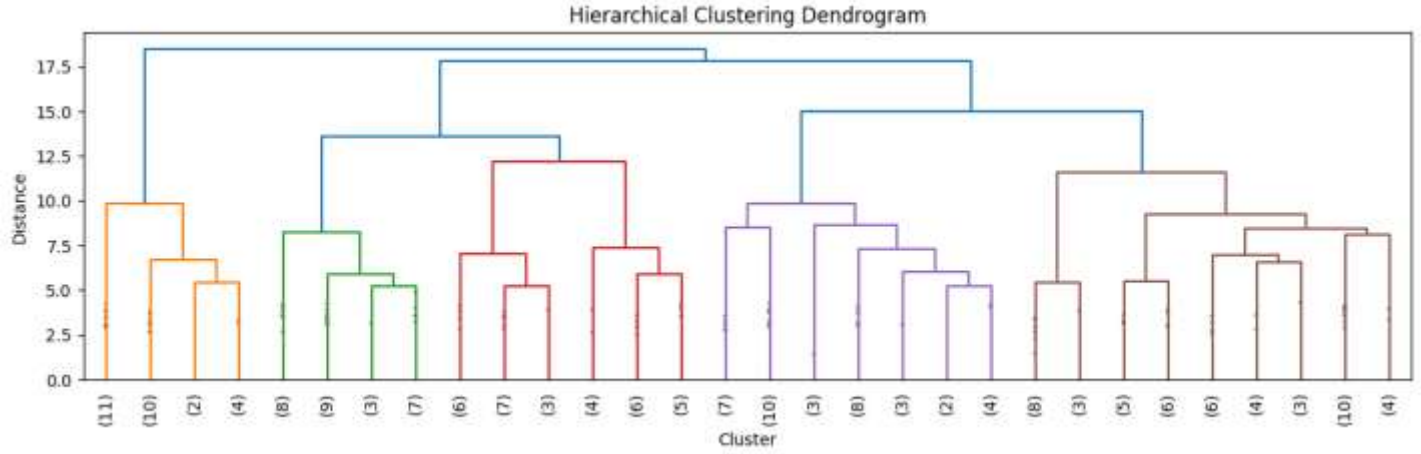

Figure 10: Hierarchical Clustering Dendrogram

Hierarchical clustering organized residues into a tree-like structure based on similarity. Close branches represented residues with comparable embedding patterns, while distant branches indicated unique or outlier residues. This hierarchical organization emphasizes relationships between amino acids at both local and global levels.

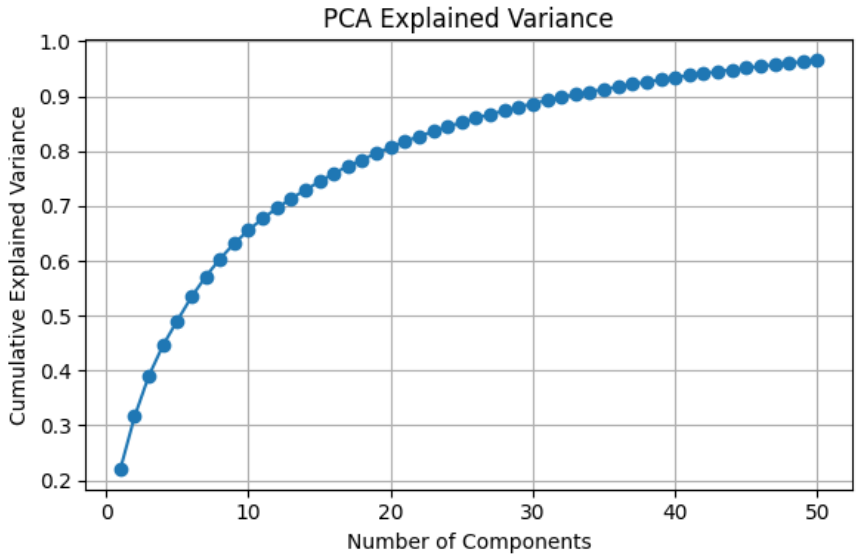

Figure 11: PCA Explained Variance

The cumulative explained variance plot showed that the first 20–25 principal components account for most of the variability in the embeddings. This confirms that dimensionality reduction retains meaningful biological information, while reducing noise and redundancy in the data.

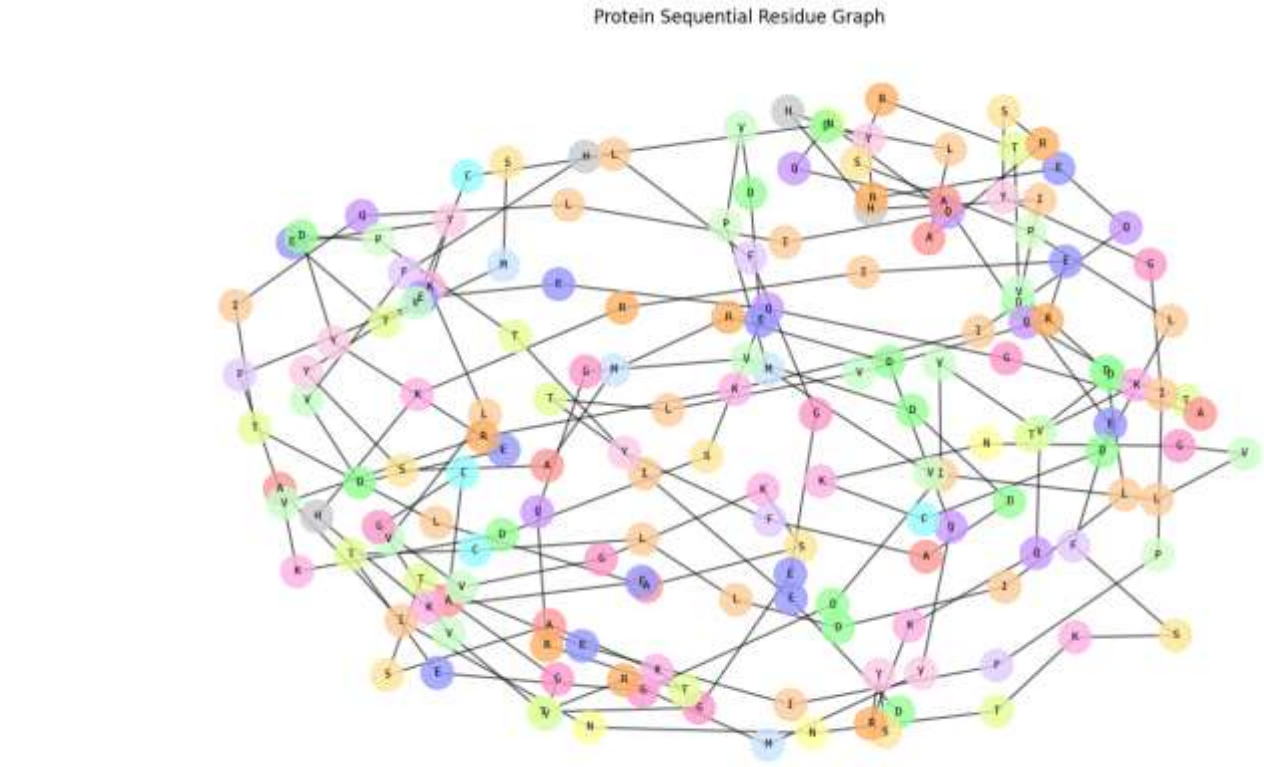

Figure 13: Protein Sequential Residue Graph

The residue graph provided a network-based view of amino acid connectivity. Nodes represented residues, and edges reflected sequential or embedding-based relationships. Clusters within the graph corresponded to structural motifs or functional domains, demonstrating how embeddings preserve sequence-to-structure mapping.

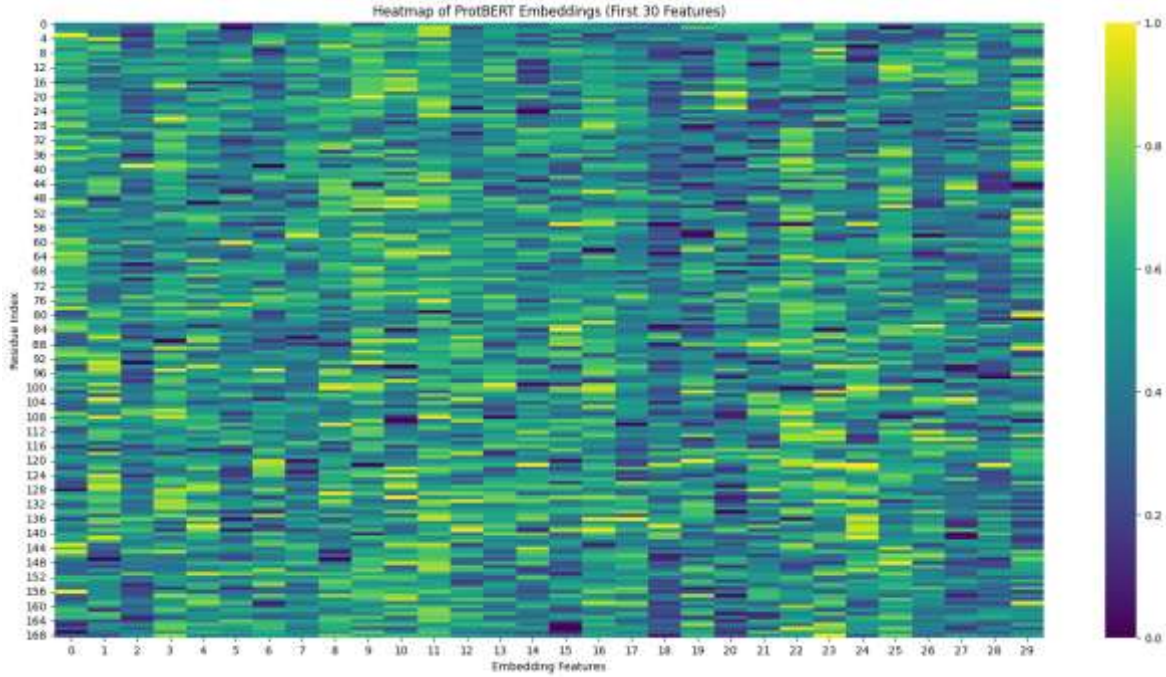

Figure 12: Heatmap of ProtBERT Embeddings

The heatmap of ProtBERT features across all residues highlighted regions of high and low embedding intensity. Conserved stretches appeared as continuous patterns, potentially corresponding to structural domains or functional motifs. This visualization provided a residue-level fingerprint of the protein's embedding landscape.

CONCLUSION

In order to examine the structural and functional characteristics of KRAS-G12C, one of the most difficult oncogenic drivers in lung and other cancers, this study shows the value of combining Next-Generation Sequencing (NGS) with Artificial Intelligence (AI). KRAS's

function as a key molecular switch in signal transduction was confirmed by the discovery of conserved motifs through sequence analysis, including the P-loop, switch regions, and GTP-binding domains. The reliability of the structural framework for inhibitor studies was validated by structural comparisons using homologous models and crystallographic data (9KPM), which showed high similarity with RMSD values less than 0.6 Å. By capturing residue-level biochemical features, conformational dynamics, and functional clustering, AI-driven modelling such as AlphaFold predictions and ProtBERT embeddings further improved our comprehension. Network-based residue graphs, clustering, and dimensionality reduction (PCA, t-SNE, UMAP) identified unique structural domains and sequence–function relationships. These analyses highlight how deep learning models can reveal subtle molecular insights and supplement experimental data. A thorough understanding of KRAS-G12C biology is made possible by the combination of NGS data, molecular docking, and AI-based structural predictions. The results point to possible directions for logical medication design, especially when it comes to improving inhibitors like JAB-16. Furthermore, the methodology used here can be used as a model to investigate additional carcinogenic mutations, improving precision oncology and speeding up the creation of targeted treatments.

REFERENCES

1. Ashrafizaveh, S., Ashrafizadeh, M., Zarrabi, A., Husmandi, K., Zabolian, A., Shahinozzaman, M., Aref, A. R., Hamblin, M. R., Nabavi, N., Crea, F., Wang, Y., & Ahn, K. S. (2021). Long non-coding RNAs in the doxorubicin resistance of cancer cells. *Cancer Letters, 508*, 104–114.
2. Cruz-Ramos, C., García-Ávila, O., Almaraz-Damián, J. A., Ponomaryov, V., Reyes-Reyes, R., & Sadovnychiy, S. (2023). Benign and malignant breast tumor classification in ultrasound and mammography images via fusion of deep learning and handcraft features. *Entropy, 25*(7), 991
3. Jesser, E. A., Brady, N. J., Huggins, D. N., Witschen, P. M., O'Connor, C. H., & Schwertfeger, K. L. (2021). STAT5 is activated in macrophages by breast cancer cell-derived factors and regulates macrophage function in the tumor microenvironment. *Breast Cancer Research, 23*(1), 104.
4. Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians, 71*(3), 209–249.
5. Cainap, C., Balacescu, O., Cainap, S. S., & Pop, L. A. (2021). Next generation sequencing technology in lung cancer diagnosis. *Biology, 10*(9), 864.
6. Schabath, M. B., & Cote, M. L. (2019). Cancer progress and priorities: Lung cancer. *Cancer Epidemiology, Biomarkers & Prevention, 28*(10), 1563–1579.
7. Oser, M. G., Niederst, M. J., Sequist, L. V., & Engelman, J. A. (2015). Transformation from non-small-cell lung cancer to small-cell lung cancer: Molecular drivers and cells of origin. *The Lancet Oncology, 16*(4), e165–e172.
8. Schiller, J. H. (2001). Current standards of care in small-cell and non-small-cell lung cancer. *Oncology, 61*(Suppl. 1), 3–13.
9. Tsim, S., O'Dowd, C. A., Milroy, R., & Davidson, S. (2010). Staging of non-small cell lung cancer (NSCLC): A review. *Respiratory Medicine, 104*(12), 1767–1774.
10. Tanoue, L. T., & Detterbeck, F. C. (2009). New TNM classification for non-small-cell lung cancer. *Expert Review of Anticancer Therapy, 9*(4), 413–423.
11. Gridelli, C., Rossi, A., Carbone, D. P., Guarize, J., Karachaliou, N., Mok, T., Petrella, F., Spaggiari, L., & Rosell, R. (2015). Non-small-cell lung cancer. *Nature Reviews Disease Primers, 1*, 15009.
12. Moskalev, E. A., Stöhr, R., Rieker, R., Hebele, S., Fuchs, F., Sirbu, H., Mastitsky, S. E., Boltze, C., König, H., Agaimy, A., Hartmann, A., & Haller, F. (2013). Increased detection rates of EGFR and KRAS mutations in NSCLC specimens with low tumour cell content by 454 deep sequencing. *Virchows Archiv, 462*(4), 409–419.mutations in NSCLC specimens with low tumour cell content by 454 deep sequencing. *Virchows Arch.* 2013;462(4):409–419.
13. Addeo, A., Banna, G. L., & Friedlaender, A. (2021). KRAS G12C mutations in NSCLC: From target to resistance. *Cancers, 13*(11), 2541. https://doi.org/10.3390/cancers13112541
14. Chan, A. H., & Simanshu, D. K. (2024). Crystallographic studies of KRAS in complex with small molecules and RAS-binding proteins. In D. K. Simanshu (Ed.), *KRAS* (*Methods in Molecular Biology*, Vol. 2797, pp. 47–65). Springer. https://doi.org/10.1007/978-1-0716-4601-4_4
15. Bronkhorst, A. J., & Holdenrieder, S. (2023). The changing face of circulating tumor DNA (ctDNA) profiling: Factors that shape the landscape of methodologies, technologies, and commercialization. *Translational Oncology, 33*, 101696. https://doi.org/10.1016/j.tranon.2023.101696
16. Wang, H., Fu, T., Du, Y., Gao, W., Huang, K., Liu, Z., et al. (2023). Scientific discovery in the age of artificial intelligence. *Nature, 620*(7960), 47–60. https://doi.org/10.1038/s41586-023-06221-2
17. Marcus, H. J., Ramirez, P. T., Khan, D. Z., Layard Horsfall, H., Hanrahan, J. G., Williams, S. C., et al. (2024). The IDEAL framework for surgical robotics: Development, comparative evaluation and long-term monitoring. *Nature Medicine, 30*(1), 61–75. https://doi.org/10.1038/s41591-023-02732-7
18. Chua, I. S., Gaziel-Yablowitz, M., Korach, Z. T., Kehl, K. L., Levitan, N. A., Arriaga, Y. E., et al. (2021). Artificial intelligence in oncology: Path to implementation. *Cancer Medicine, 10*(13), 4138–4149. https://doi.org/10.1002/cam4.3935
19. Greener, J. G., Kandathil, S. M., Moffat, L., & Jones, D. T. (2022). A guide to machine learning for biologists. *Nature Reviews Molecular Cell Biology, 23*(1), 40–55. https://doi.org/10.1038/s41580-021-00407-0
20. Zhou, J., Hu, B., Feng, W., Zhang, Z., Fu, X., Shao, H., et al. (2023). An ensemble deep learning model for risk stratification of invasive lung adenocarcinoma using thin-slice CT. *NPJ Digital Medicine, 6*, 119. https://doi.org/10.1038/s41746-023-00866-z

21. Awad, M. M., Liu, S., Rybkin, I. I., Arbour, K. C., Dilly, J., Zhu, V. W., Johnson, M. L., Heist, R. S., Patil, T., Riely, G. J., Jacobson, J. O., Yang, X., Persky, N. S., Root, D. E., Lowder, K. E., Feng, H., Zhang, S. S., Haigis, K. M., Hung, Y. P., Sholl, L. M., & Aguirre, A. J. (2021). Acquired resistance to KRASG12C inhibition in cancer. *The New England Journal of Medicine*, *384*(25), 2382–2393. https://doi.org/10.1056/NEJMoa2105281

22. Nussinov, R., Tsai, C. J., & Jang, H. (2021). Anticancer drug resistance: An update and perspective. *Drug Resistance Updates*, *59*, 100796. https://doi.org/10.1016/j.drup.2021.100796

23. Chaudhary, S., & Kumari, U. (2023). NGS, molecular docking and network pharmacology reveal potent inhibitor for the treatment of lung cancer. *International Journal of Emerging Technologies and Innovative Research*, *11*(9), 116–126. https://doi.one/10.1729/Journal.41696

24. Hallin, J., Engstrom, L. D., Hargis, L., Calinisan, A., Aranda, R., Briere, D. M., Sudhakar, N., Bowcut, V., Baer, B. R., Ballard, J. A., Burkard, M. R., Fell, J. B., Fischer, J. P., Vigers, G. P., Xue, Y., Gatto, S., Fernandez-Banet, J., Pavlicek, A., Velastagui, K., Chao, R. C., … Christensen, J. G. (2020). The KRASG12C Inhibitor MRTX849 Provides Insight toward Therapeutic Susceptibility of KRAS-Mutant Cancers in Mouse Models and Patients. Cancer discovery, 10(1), 54–71. https://doi.org/10.1158/2159-8290.CD-19-1167

25. Reck, M., Carbone, D. P., Garassino, M., &Barlesi, F. (2021). Targeting KRAS in non-small-cell lung cancer: recent progress and new approaches. Annals of oncology : official journal of the European Society for Medical Oncology, 32(9), 1101–1110. https://doi.org/10.1016/j.annonc.2021.06.001

26. Karuppasamy, M. P., Venkateswaran, S., &Subbiah, P. (2020). PDB-2-PBv3.0: An updated protein block database. Journal of bioinformatics and computational biology, 18(2), 2050009. https://doi.org/10.1142/S0219720020500092

27. Gao J., Wu H., Shi X., Huo Z., Zhang J., Liang Z. Comparison of Next-Generation Sequencing, Quantitative PCR, and Sanger Sequencing for Mutation Profiling of EGFR, KRAS, PIK3CA and BRAF in Clinical Lung Tumors. Clin. Lab. 2016; 62:689–696. doi: 10.7754/Clin.Lab.2015.150837. [DOI] [PubMed] [Google Scholar]

28. Xu X., Yang Y., Li H., Chen Z., Jiang G., Fei K. Assessment of the clinical application of detecting EGFR, KRAS, PIK3CA and BRAF mutations in patients with non-small cell lung cancer using next-generation sequencing. Scand. J. Clin. Lab. Investig. 2016; 76:386–392. doi: 10.1080/00365513.2016.1183813. [DOI] [PubMed] [Google Scholar]

29. Vaughn C.P., Costa J.L., Feilotter H.E., Petraroli R., Bagai V., Rachiglio A.M., Marino F.Z., Tops B., Kurth H.M., Sakai K., et al. Simultaneous detection of lung fusions using a multiplex RT-PCR next generation sequencing-based approach: A multi-institutional research study. BMC Cancer. 2018; 18:828. doi: 10.1186/s12885-018-4736-4. [DOI] [PMC free article] [PubMed] [Google Scholar]

30. Kumari Uma and Choudhary, Ashok Kumar, "Genome Sequence Analysis of SolanumLycopersicum by Applying Sequence Alignment Method to Determine the Statistical Significance of an Alignment", International Journal of Bio-Technology and Research (IJBTR), 2249-6858;9-12, 2016.

31. Kumari Uma, &Choudhary, A. K, "Genome sequence analysis of solanumlycopersicum showing the phylogenetic relationship based on multiple sequence alignment and conserved domain proteins", International journal of advanced biotechnology and research, 7(4), 2012-2014, 2016.

32. Vinita Kukreja, Uma Kumari,"Genome Annotation of Brain Cancer and Structure Analysis by applying Drug Designing Technique\", International Journal of Emerging Technologies and Innovative Research, 9(5);473-k479, May, 2022.

33. Varol D. AI in pharma: innovations and challenges | Scilife. *Scilife*, https://www.scilife.io/blog/ai-pharma-innovation-challenges (accessed 18 February 2025).

34. Joo MS, Pyo KH, Chung JM, Cho BC. Artificial intelligence-based non-small cell lung cancer transcriptome RNA-sequence analysis technology selection guide. *Front Bioeng Biotechnol.* 2023 Feb 15;11:1081950.

35. Sayle, R. A., & Milner-White, E. J. (1995). RASMOL: biomolecular graphics for all. *Trends in biochemical sciences*, *20*(9), 374. https://doi.org/10.1016/s0968-0004(00)89080-5

36. Bernstein H. J. (2000). Recent changes to RasMol, recombining the variants. *Trends in biochemical sciences*, *25*(9), 453–455. https://doi.org/10.1016/s0968-0004(00)01606-6

37. Goodsell D. S. (2005). Representing structural information with RasMol. *Current protocols in bioinformatics*, *Chapter 5*, . https://doi.org/10.1002/0471250953.bi0504s11

38. Shipra Chaudhary Uma Kumari, "NGS, MOLECULAR DOCKING AND NETWORK PHARMACOLOGY REVEAL POTENT INHIBITOR FOR THE TREATMENT OF LUNG CANCER", International Journal of Emerging Technologies and Innovative Research , Vol.11, Issue 9, page no. ppf116-f126, https://doi.one/10.1729/Journal.41696

39. Uma Kumari ,Eaganayagan Malligaarjun"Next-Generation Sequence and Structural Analysis of Oncogenic KRAS Q61E GMPPCP-bound in Human Lung Cancer", International Journal of Emerging Technologies and Innovative Research,jetir, Vol.12, Issue 4, page no. ppk26-k38, **2025**, http://doi.one/10.1729/Journal.44947

40. Kumari, Uma & Gupta, Shruti. (2023). NGS and Sequence Analysis with Biopython for Prospective Brain Cancer Therapeutic Studies. International Journal for Research in Applied Science and Engineering Technology. 11. 10.22214/ijraset.2023.50885.

41. Tanmay Bandbe , Juri Saikia , Uma Kumari. (2025). NGS Analysis Human Papillomavirus Type 18 E2 DNA-Binding Domain Bound to its DNA Target with Biopython. *South Eastern European Journal of Public Health*, 3781–3792. https://doi.org/10.70135/seejph.vi.5856,SEEJPH Volume XXVI, S2

42. Kumari, U., & Bajaj, T. (2023). NGS data analysis and active site identification: Alpha-1-acid glycoprotein bound to potent anti-tumor compound UCN-01 in malignant brain tumor. *International Journal of Emerging Technologies and Innovative Research*, *12*(3), g107–g116. http://doi.one/10.1729/Journal.44259

43. Kumari, U., Raj, K., et al. (2025). Drug discovery against FGFR2 mutation in lung cancer: An approach using NGS and medicinal plants. *International Journal of Emerging Technologies and Innovative Research*, *12*(5), h569–h583. https://doi.org/10.56975/jetir.v12i5.562943

44. Hassoun, S., Jefferson, F., Shi, X., Stucky, B., Wang, J., & Rosa, E. (2022). Artificial intelligence for biology. *Integrative and Comparative Biology*, *61*(6), 2267–2275. https://doi.org/10.1093/icb/icab188

45. Xu, Y., Liu, X., Cao, X., Huang, C., Liu, E., Qian, S., Liu, X., Wu, Y., Dong, F., & Qiu, C. W. (2021). Artificial intelligence: A powerful paradigm for scientific research. *Innovation*, *2*(4), 100179. https://doi.org/10.1016/j.xinn.2021.100179