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

Integrative CRISPR-Cas9 and machine learning Approaches for Target Discovery and Therapeutic Development in Malignant Brain Tumor

Narotham BP[1] Uma Kumari[1]
[1]Trainee at Bioinformatics Project and Research Institute, Noida-201301, India
[1]Senior Bioinformatics Scientist, Bioinformatics Project and Research Institute, Noida -201301, India

*[1]Corresponding Author: Uma Kumari

Abstract: Malignant brain tumours, such as glioblastoma, remain among the most challenging cancers to treat due to their heterogeneity, aggressive progression, and resistance to conventional therapies. The integration of CRISPR-Cas9 gene-editing technology with advanced machine learning (ML) approaches offers a transformative framework for identifying novel therapeutic targets and accelerating drug development. CRISPR-Cas9 enables precise genetic perturbations to uncover critical genes and pathways driving tumorigenesis, while ML algorithms analyze high-throughput genomic, proteomic, and clinical data to predict key molecular targets and optimize therapeutic strategies. This synergistic approach facilitates the identification of tumor-specific vulnerabilities, enhances the design of personalized therapies, and improves the prediction of drug efficacy and resistance mechanisms. By combining CRISPR-based functional genomics with ML-driven data integration, this strategy holds immense potential to advance precision medicine for malignant brain tumors, paving the way for more effective and tailored treatment options.

Keywords: **CRISPR-Cas9, machine learning, malignant brain tumor, glioblastoma, target discovery, therapeutic development, precision medicine, personalized therapy, Structure analysis**

I. Introduction

Malignant brain tumors, including glioblastoma multiforme (GBM), represent some of the most lethal and treatment-resistant cancers, characterized by rapid progression, cellular heterogeneity, and poor prognosis. Despite advances in surgical resection, radiotherapy, and chemotherapy, the median survival for GBM patients remains approximately 15 months, underscoring the urgent need for novel therapeutic strategies. The complexity of brain tumors, driven by genetic and epigenetic alterations, tumor microenvironment interactions, and therapy resistance, necessitates innovative approaches to identify actionable molecular targets and develop effective treatments [1, 2, 3].

CRISPR-Cas9, a revolutionary gene-editing technology, has transformed functional genomics by enabling precise, high-throughput genetic perturbations to uncover critical drivers of disease. In the context of brain tumors, CRISPR-based screens have identified novel oncogenes, tumor suppressors, and dependencies that contribute to tumor initiation and progression. These screens provide a robust platform for systematically interrogating gene function, revealing potential therapeutic targets that can be validated in preclinical models [4, 5, 6, 7].

Concurrently, machine learning (ML) has emerged as a powerful tool for analyzing complex biological datasets, including genomic, transcriptomic, proteomic, and clinical data, to uncover patterns and predict therapeutic outcomes. ML algorithms, such as deep learning and random forest models, excel at integrating multi-omics data to identify biomarkers, predict drug responses, and stratify patients for personalized therapies. In brain tumor research, ML has been used to classify tumor subtypes, predict survival, and identify druggable pathways, offering a data-driven approach to complement experimental findings [8, 9, 10].

The integration of CRISPR-Cas9 and ML represents a synergistic strategy to accelerate target discovery and therapeutic development for malignant brain tumors. CRISPR-based screens generate high-dimensional functional data, which ML can analyze to prioritize biologically relevant targets and predict their therapeutic potential. For example, ML models can integrate CRISPR-derived gene knockout data with patient-derived genomic profiles to identify synthetic lethal interactions or resistance mechanisms, guiding the development of combination therapies. Additionally, ML can optimize CRISPR screen design by predicting off-target effects or enhancing guide RNA efficiency, improving the precision of genetic perturbations [11, 12, 13].

This integrative approach holds immense promise for addressing the challenges of brain tumor heterogeneity and therapy resistance. By combining the precision of CRISPR-Cas9 with the predictive power of ML, researchers can uncover novel molecular vulnerabilities, design targeted therapies, and advance precision medicine for patients with malignant brain tumors. This introduction outlines the potential of this interdisciplinary strategy and sets the stage for exploring its applications in target discovery and therapeutic development.

II. Materials and Methods

Data Acquisition

The L3MBTL3 three-dimensional structure (PDB ID: 8Y76) was obtained from the RCSB Protein Data Bank (PDB). Corresponding protein and nucleotide sequences were retrieved from UniProt and NCBI GenBank databases. Specifically, the L3MBTL3 protein sequence (UniProt accession: Q96B26) was downloaded from UniProt, and any complementary nucleotide sequences were obtained from GenBank. These databases provide curated structural coordinates and sequence annotations for L3MBTL3 (and homologs) used in downstream analyses [21, 24, 26].

Sequence Analysis

Functional and pathway annotations of L3MBTL3 were performed using sequence-based bioinformatics tools. The raw protein sequence was processed and manipulated with Biopython, enabling automated retrieval of homologous sequences and basic analyses [27]. To identify conserved domains and motifs, the InterProScan 5 pipeline was run on the protein sequence [14]. InterProScan integrates multiple signature databases to classify protein functions genome-wide. Protein–protein interaction networks and pathway associations were inferred using the STRING database (v11). STRING aggregates experimental and predicted interaction data across organisms and assigns confidence scores, allowing identification of potential functional partners for L3MBTL3. All sequence analyses used default parameters unless otherwise noted, and results were cross-referenced with pathway and Gene Ontology resources for validation [22].

Structural Analysis

We visualized and analyzed the L3MBTL3 3D structure using molecular graphics software. RasMol and PyMOL were used to inspect secondary structure elements, ligand interactions, and surface properties. The PDB file was loaded into these viewers to manually examine pocket geometry and to generate publication-quality images. Structural superposition of 8Y76 with related MBT-domain proteins was performed in PyMOL to assess conformational differences. PyMOL's alignment tools calculated root-mean-square deviations (RMSD) between L3MBTL3 and reference structures. Key structural features such as binding-site residues, helices, and β-strands were annotated and extracted from the coordinate file using built-in PyMOL commands. Figure panels of ribbon and surface representations were prepared with PyMOL's ray-tracing and coloring functions. These analyses provided detailed 3D context for functional predictions [19, 20].

Mutation and Variant Analysis

To explore disease relevance, somatic mutations and variants in L3MBTL3 were surveyed using cancer mutation databases. The COSMIC database (Catalogue Of Somatic Mutations In Cancer) was queried for L3MBTL3 variants. COSMIC is a comprehensive repository of tumor-associated mutations, with tools to browse gene-specific mutation spectra. We noted any reported non-synonymous single nucleotide variants (SNVs), copy-number changes, or indels within L3MBTL3. In parallel, the Cancer Gene Census within COSMIC was checked to see if L3MBTL3 has been implicated as a driver gene in cancer. Any mutations from COSMIC were mapped onto the 3D structure to assess potential impacts on protein function. Additional variant information was obtained from dbSNP and ClinVar databases for known germline or somatic polymorphisms, although the focus remained on COSMIC-curated cancer mutations [15, 18, 23].

CRISPR Guide RNA Design

Potential CRISPR/Cas9 target sites were identified in the L3MBTL3 gene using dedicated design tools. We utilized E-CRISP and CHOPCHOP to generate candidate single-guide RNAs (sgRNAs). E-CRISP was run to scan both sense and antisense strands for 20-nucleotide target sequences immediately followed by the canonical NGG PAM. E-CRISP's output ranks candidates by on-target efficiency and predicted off-target burden; it also reports genomic context (e.g. whether a site overlaps exons or CpG islands). CHOPCHOP v3 was similarly employed to design sgRNAs, taking advantage of its expanded scoring for activity and off-targets, plus its annotations of alternative transcripts. For each sgRNA, we evaluated GC content and self-complementarity (secondary structure propensity), and recorded Bowtie-based off-target mismatches up to three mismatches (MM0–MM3) as provided by the tools. Guide candidates were filtered to avoid homopolymer runs and low complexity. Top-ranked sgRNAs (high on-target scores, low predicted off-targets) were retained for potential experimental use [16, 31].

Machine Learning Analysis

To explore sequence–structure relationships, we generated protein language embeddings and performed clustering. We used the ProtTrans pretrained models to obtain ProtBERT embeddings for L3MBTL3 and related sequences. These contextual embeddings (1024-dimensional vectors) encode biochemical features of each sequence. Dimensionality reduction (PCA) was applied to the ProtBERT embeddings to visualize overall data structure. We then applied k-means clustering (scikit-learn implementation; default 300 clusters) to partition proteins by embedding similarity, and separately performed agglomerative hierarchical clustering for comparison. Cluster assignments were overlaid on protein families and functional annotations to interpret groupings. Network flow (Sankey) diagrams were generated to illustrate transitions between clusters under different embedding layers, and heatmaps were created (using Python seaborn and matplotlib) to display pairwise sequence distance and clustering. Clustering quality was assessed by silhouette scores and manual inspection. This ML-based approach aimed to highlight any latent patterns in L3MBTL3 related to structural or functional features [17, 25, 27, 29, 32].

Gene Structure and Isoform Mapping

Gene models for human L3MBTL3 were obtained from Ensembl (Ensembl release 100). Transcript isoforms (mRNA) were visualized alongside genomic coordinates to map exon–intron organization. We used Ensembl's BioMart API to download L3MBTL3 gene structure, and plotted exon boundaries versus chromosomal position. CpG islands were located using standard criteria (length >200 bp, GC% >50%, observed/expected CpG >0.6) as defined by Gardiner-Garden and Frommer, and their positions overlaid on the gene diagram. The designed sgRNA target sites (from above) were mapped onto the gene to show which exons or introns they hit. This integrated gene schematic (with annotated isoforms, CpG islands, and guide sites) was generated using the Genome Browser (e.g. UCSC) and custom plotting scripts. Such a map enables visualization of which transcript variants are targeted by each sgRNA and the regulatory context of the L3MBTL3 locus [28, 30].

III. Result and Discussion

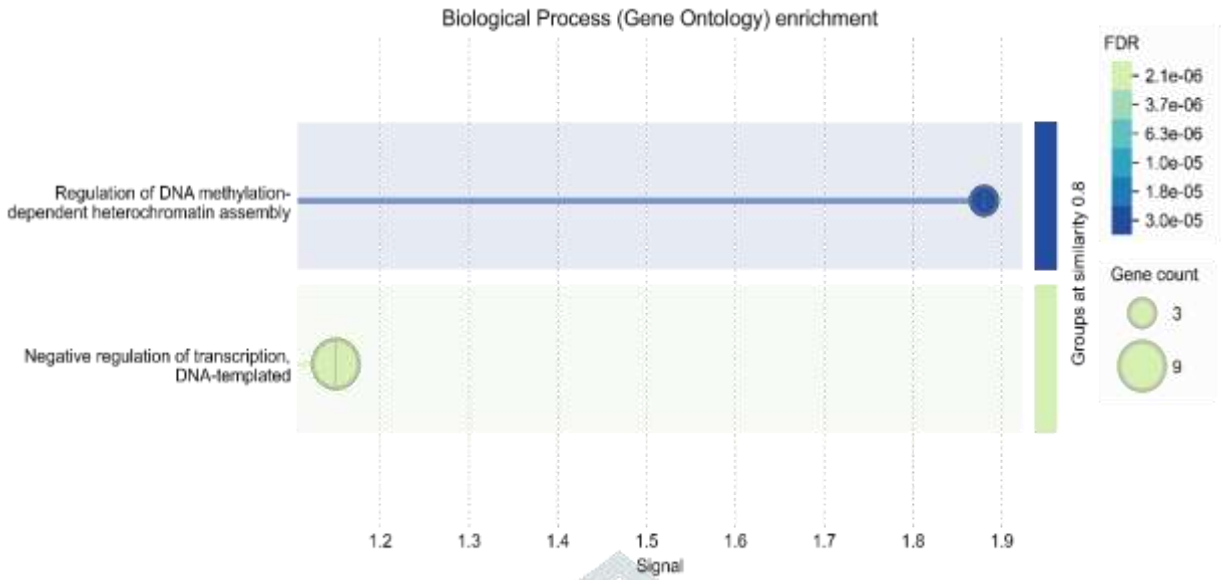

Figure 1: Biological Process (Gene Ontology) Enrichment Analysis

The Gene Ontology (GO) enrichment assessment for protein 8Y76 highlights two major biological processes closely linked to its functional activity. The strongest enrichment signal was associated with DNA methylation–dependent heterochromatin assembly, showing a high effect size of roughly 1.88, a highly significant FDR of 2.1×10^{-6}, and contributions from nine genes. The prominence and coloration of this cluster on the plot underscore both its statistical strength and the number of genes involved. These findings suggest that 8Y76 is heavily engaged in epigenetic regulation, particularly in pathways where DNA methylation facilitates the formation of compact heterochromatin and contributes to gene silencing. A second enriched pathway, negative regulation of DNA-templated transcription, presented a more moderate signal of around 1.15, with a significant FDR of 3.0×10^{-5} and involvement of three genes. Even though this category is less enriched than the first, it still points to a clear role in transcriptional repression. Taken together, the enrichment pattern indicates that 8Y76 is implicated in controlling gene expression through chromatin organization and transcription-modulating mechanisms.

Name	Nucleotide sequence	SAE-Score	Target	Matchstring	Number of Hits
L3MBTL3_30_25750	GAATCTGCCTCTAGC ACAAG NGG		ENSG00000198945::L3 MBTL3	Matchstring Info	1
L3MBTL3_32_25750	GATGGAGTGGGCAC GTTACC NGG		ENSG00000198945::L3 MBTL3	Matchstring Info	1
L3MBTL3_31_25750	GTTTGATGTGTTCAG TGTTA NGG		ENSG00000198945::L3 MBTL3	Matchstring Info	1
L3MBTL3_29_25750	GATTTATTTTTTAACC TGAA NGG		ENSG00000198945::L3 MBTL3	Matchstring Info	1
L3MBTL3_33_25750	GTTACCAGGAAGTGA CTTAA NGG		ENSG00000198945::L3 MBTL3	Matchstring Info	1
L3MBTL3_116_25750	GGTAAATGAGTTTGG AGCCC NGG		ENSG00000198945::L3 MBTL3	Matchstring Info	1
L3MBTL3_115_25750	GCTGTGTTGTTGCTA GTTTC NGG		ENSG00000198945::L3 MBTL3	Matchstring Info	1
L3MBTL3_56_115875	GCAGGGATTCCTGCG ACAGT NGG		ENSG00000227678::R P11-73O6.3	Matchstring Info	1
L3MBTL3_57_115875	GCTCCATTTGGAAAC TTTAC NGG		ENSG00000227678::R P11-73O6.3	Matchstring Info	1
L3MBTL3_85_38625	GGGGTTCGCAGAGA GCACGG NGG		ENSG00000198945::L3 MBTL3	Matchstring Info	1
L3MBTL3_88_38625	GATTCGGCTGTACTA AAGCA NGG		ENSG00000198945::L3 MBTL3	Matchstring Info	1

Figure 2: E-crispr Analysis of target gene L3MBTL3

The E-CRISP analysis produced multiple sgRNA candidates designed to target the L3MBTL3 gene (ENSG0000019845). Each guide was assessed for its specificity, predicted cutting activity, and overall editing efficiency using the SAE scoring system. The majority of sgRNAs registered a single genomic hit, indicating that they align directly and uniquely with the L3MBTL3 locus. A few guides, however, showed matches in unannotated regions on chromosomes X, 3, 12, and 18, pointing to possible off-target sites. Variation in the SAE bar patterns was evident: guides with a prominent green segment demonstrated stronger predicted activity and higher specificity, whereas those showing extended orange segments suggested a greater chance of unintended cleavage elsewhere in the genome. The blue components of the score indicated that several guides were predicted to have solid editing efficiency. Matchstring details further confirmed how each sgRNA aligned with its target sequence, with "1 hit" marking the most dependable options. In summary, while numerous guides are capable of targeting

L3MBTL3, the best candidates are those that combine a single on-target hit, high predicted activity, and low off-target risk, making them suitable for precise gene-editing or functional studies involving L3MBTL3.

L3MBTL3_131_25750	GTACCTGTGGGACAT CCAGG NGG	ENSG00000198945::L3 MBTL3	Matchstring Info	4
L3MBTL3_131_25750	GTACCTGTGGGACAT CCAGG NGG	ENSG00000166341::D CHS1	Matchstring Info	4
L3MBTL3_130_25750	GCCTACTGGACATCT CCACC NGG	ENSG00000198945::L3 MBTL3	Matchstring Info	4
L3MBTL3_130_25750	GCCTACTGGACATCT CCACC NGG	X	Matchstring Info	4
L3MBTL3_130_25750	GCCTACTGGACATCT CCACC NGG	3	Matchstring Info	4
L3MBTL3_130_25750	GCCTACTGGACATCT CCACC NGG	ENSG00000258135::R P11-70F11.11	Matchstring Info	4
L3MBTL3_132_25750	GAGGTACCTGTGGGA CATCC NGG	ENSG00000198945::L3 MBTL3	Matchstring Info	4
L3MBTL3_132_25750	GAGGTACCTGTGGGA CATCC NGG	12	Matchstring Info	4
L3MBTL3_132_25750	GAGGTACCTGTGGGA CATCC NGG	X	Matchstring Info	4
L3MBTL3_132_25750	GAGGTACCTGTGGGA CATCC NGG	ENSG00000140386::S CAPER	Matchstring Info	4

Figure 3: Output of a crispr guide RNA selection and off-target analysis tool

The CRISPR guide RNA analysis produced several sgRNA options for targeting the L3MBTL3 gene, each displaying distinct patterns of predicted activity, efficiency, and specificity. The scoring bars showed considerable variation: the blue bars reflected the expected cutting efficiency of each guide, the orange bars indicated specificity where longer segments signaled a greater likelihood of off-target interactions, and the green bars represented the functional relevance of the target region. Most of the sgRNAs listed in this panel mapped to four different genomic locations, which suggests that they may bind to multiple unintended sites across chromosomes such as X, 3, and 12, as well as other gene regions. Guides with this level of off-target overlap would require further scrutiny before being used experimentally. Despite this, a few sgRNAs demonstrated more balanced scoring profiles, combining reasonable efficiency with acceptable annotation values, making them more suitable candidates for follow-up verification. Overall, the results indicate that while many guides show potential, only those with high efficiency, meaningful annotation signals, and minimal off-target predictions should be prioritized for precise gene-editing applications involving L3MBTL3.

Figure 4: Gene Architecture, Transcript Diversity and CRISPR Target Distribution of L3MBTL3

The genomic representation of L3MBTL3 on chromosome 6 shows that it is a structurally intricate gene, producing a wide range of transcript variants through extensive alternative splicing. These isoforms, illustrated by multiple transcript tracks, indicate functional diversity in how the gene is expressed. A prominent CpG-rich region near the 5′ end points to potential epigenetic regulation, suggesting that DNA methylation may influence L3MBTL3 expression.

Mapping of CRISPR guide sites onto transcript ENST00000368139 identified numerous sgRNA positions spread across the coding regions. Notably, several guide clusters appear around exons 4–6 and 9–11, marking these exons as particularly accessible for genome editing.

The distribution of these guides allows for different experimental strategies, including complete gene disruption by targeting exons common to all isoforms or precision editing by selecting guides that correspond to isoform-specific exons. Some sgRNA sites overlap CpG islands, implying that editing within these regions could also affect regulatory elements linked to methylation.
Overall, the structural layout and guide distribution demonstrate that L3MBTL3 provides a versatile platform for CRISPR editing, supporting both broad functional knockout and selective isoform-level manipulation for research and therapeutic studies.

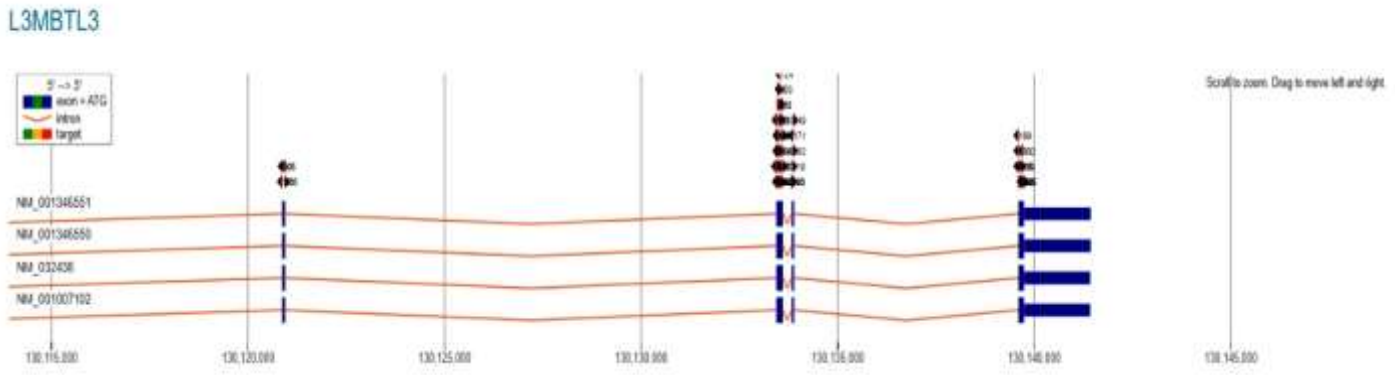

Figure 5: Representing Genomic organization and transcript variants of the L3MBTL3 gene

The genomic view of L3MBTL3 illustrates that the gene occupies a region between approximately 130.115 Mb and 130.145 Mb on chromosome 6. The figure displays four different transcript variants, each with its own pattern of exons and introns, demonstrating that the gene undergoes extensive alternative splicing. In the diagram, blue blocks mark the exon boundaries while the orange lines indicate intronic sequences. Several targetable regions, highlighted in green and yellow, represent positions where sgRNAs, sequence variants, or PCR amplicons can be mapped. These targets are mainly concentrated toward the 3′ portion of the gene, where the exons appear more densely arranged, suggesting this region is particularly well suited for CRISPR editing. The directional arrows across the gene point to functional segments such as regulatory sequences, transcriptional elements, or possible mutation-sensitive areas. Overall, the figure shows that L3MBTL3 has a diverse transcript structure and contains multiple accessible CRISPR target sites, supporting its suitability for gene-editing and functional studies.

bin	chrom	chromStart	chromEnd	name	score	strand	thickStart	thickEnd	itemRgb
1577	chr6	130042676	130042699	Ranked:280	0	-	130042676	130042699	255,0,0
1577	chr6	130042705	130042728	Ranked:48	0	+	130042705	130042728	255,0,0
1577	chr6	130042712	130042735	Ranked:18	0	-	130042712	130042735	255,0,0
1577	chr6	130042734	130042757	Ranked:168	0	+	130042734	130042757	255,0,0
1577	chr6	130042739	130042762	Ranked:141	0	+	130042739	130042762	255,0,0
1577	chr6	130042747	130042770	Ranked:191	0	+	130042747	130042770	255,0,0
1577	chr6	130042752	130042775	Ranked:244	0	+	130042752	130042775	255,0,0
1577	chr6	130042753	130042776	Ranked:233	0	+	130042753	130042776	255,0,0
1577	chr6	130042765	130042788	Ranked:68	0	+	130042765	130042788	255,0,0
1577	chr6	130042779	130042802	Ranked:59	0	+	130042779	130042802	255,0,0

Figure 6: CHOPCHOP gRNA Ranking for L3MBTL3

The CHOPCHOP analysis for L3MBTL3 revealed several potential CRISPR–Cas9 guide RNA sites within the chromosomal region spanning positions 130,042,676 to 130,042,802 on chromosome 6. Each guide is assigned a ranking, which reflects its predicted suitability for genome editing. Guides with lower rank values—such as those listed around ranks 18, 48, and 59—are considered stronger candidates because they are expected to offer better on-target performance and fewer unintended interactions elsewhere in the genome. In contrast, guides with much higher rank numbers are typically less favorable due to reduced efficiency or greater likelihood of off-target effects. Taken together, the dataset confirms that the L3MBTL3 locus contains numerous editable sites, with a subset of well-ranked guides providing the most dependable choices for targeted CRISPR experiments.

Rank	Target sequence	Genomic location	Strand	GC content (%)	Self-complementarity	MM0	MM1	MM2	MM3	Efficiency
1	ATCCCATCGGGCTTTCGAGTTGG	chr6:130070982	+	55	0	0	0	0	0	54.15
2	GTACTGTGTCCTCACCGTCGCGG	chr6:130068116	+	60	0	0	0	0	1	73.49
3	GATTCGGCTGTACTAAAGCAGGG	chr6:130055234	+	45	0	0	0	0	1	65.77
4	TACCAACTCGAAAGCCCGATGGG	chr6:130070984	-	50	0	0	0	0	1	59.11
5	AGCTTACCTCCGCGACGGTGAGG	chr6:130060125	-	65	2	0	0	0	1	59.40
6	GGAACAATTGCTATGATTACTGG	chr6:130092761	+	35	0	0	0	0	1	32.83
7	GGCAAAGCTTACCTCCGCGACGG	chr6:130060130	-	60	1	0	0	0	2	63.31
8	TATTCGTGTCTCTTACTCAAAGG	chr6:130092853	-	35	0	0	0	0	2	59.45
9	TCGGCACATCAAAGATAAGTAGG	chr6:130051391	+	40	0	0	0	0	2	53.12
10	TATGGTGGACAATCGTTTCCTGG	chr6:130071068	+	45	1	0	0	0	2	49.09
11	GGATTCGGCTGTACTAAAGCAGG	chr6:130055233	+	50	0	0	0	0	2	47.86
12	ATCTCACCACTTAAGCGGTCTGG	chr6:130104478	-	50	0	0	0	0	2	46.51
13	ATACCAACTCGAAAGCCCGATGG	chr6:130070985	-	50	0	0	0	1	1	61.35
14	ACCTCCTTGAACAGCTTCGCCGG	chr6:130057477	-	55	0	0	0	1	1	57.71
15	TAGTTTCGGGTAAATGAGTTTGG	chr6:130049279	+	35	0	0	0	1	1	44.72
16	TCGAGTTGGTATGAAGCTTGAGG	chr6:130070996	+	45	0	0	0	0	3	62.33
17	TGGATATCCACCCAGTTGGGTGG	chr6:130066429	+	55	2	0	0	0	3	60.77
18	CTCTTGACCACTTGTGCTAGAGG	chr6:130042713	-	50	2	0	0	0	3	59.21
Rank	Target sequence	Genomic location	Strand	GC content (%)	Self-complementarity	MM0	MM1	MM2	MM3	Efficiency
269	TGATTTTCAAATAATGACTTAGG	chr6:130068382	-	20	0	0	2	3	60	46.57
270	ACATTAAAAAAACACACACCTGG	chr6:130070531	-	30	0	0	1	7	58	59.66
271	CTTTTATTTCTTGGAAATGACGG	chr6:130104514	-	25	0	0	0	5	66	49.14
272	CTTTTGGAGGATGGAGTTTGTGG	chr6:130066406	-	45	0	1	1	6	60	46.47
273	TCTCTCTTCTCCATCCTCCTTGG	chr6:130052957	-	50	0	0	0	4	91	45.94
274	ACCTGGTCAATGAAAACAAATGG	chr6:130113431	-	35	0	0	0	3	96	37.51
275	TTCAAATTGATTTCTGAATAGGG	chr6:130104439	-	20	0	0	1	6	94	35.00
276	AAATGTGAAACATTTTTCTTGGG	chr6:130065927	+	20	0	1	0	7	92	34.99
277	TTTTAATTTGTGTAACTTTCAGG	chr6:130075556	+	20	0	0	0	9	94	20.34
278	ACAGAGATCAGAAGGAAGAAAGG	chr6:130052855	+	40	0	0	2	0	107	52.36
279	TCTTTTCTTTGCTCTTTCATAGG	chr6:130064343	+	30	0	0	0	8	118	38.28
280	GATTTATTTTTTAACCTGAAAGG	chr6:130042677	-	20	0	0	0	11	130	47.42
281	GATTTACATTTTTCTTTCTTAGG	chr6:130065598	+	20	0	0	0	10	142	29.08
282	CAGAGATCAGAAGGAAGAAAGGG	chr6:130052856	+	40	0	0	1	13	165	43.06
283	AGAAGAGAGAGATGATGAAATGG	chr6:130052979	+	35	0	0	1	13	181	51.03
284	TAATTTTCCTTTCTTTTTGTAGG	chr6:130055149	+	30	0	0	0	13	268	22.63
285	TCTCTGCAGCTTTGAACATCAGG	chr6:130139708	-	45	0	0	2	50	>=289	39.32

Figure 7: Evaluation of CRISPR–Cas9 gRNA Candidates Targeting L3MBTL3

The CRISPR–Cas9 guide RNA assessment for the L3MBTL3 region on chromosome 6 produced a large panel of sgRNA candidates with varying sequence characteristics and predicted editing performance. The GC content of most guides fell within a generally favorable range, supporting stable Cas9 interaction, while nearly all had no detectable self-complementarity, indicating that the sequences are unlikely to fold into hairpins that might reduce activity. Off-target predictions differed widely among the guides: a few showed only minimal genomic matches, whereas others-especially those at the lower end of the ranking list—displayed high numbers of near-match sites, making them unsuitable for precise editing.

The efficiency scores showed substantial variation, ranging from roughly 20% to more than 70%, illustrating clear differences in predicted Cas9 cutting ability. Among the top-performing guides, those ranked 2, 3, and 17 demonstrated the strongest activity, each combining high efficiency with limited off-target risk. Within the separately listed lower-ranked group, guides at ranks 270, 278, and 284 stood out, offering efficiency values above 50%, acceptable GC percentages, and relatively low mismatch counts.

Altogether, the data indicate that while many sgRNAs can target the L3MBTL3 genomic region, only a subset meet the optimal criteria for accuracy and performance. The most reliable candidates are those with balanced GC content, minimal secondary-structure potential, low off-target scores, and efficiency values above 50%, making them strong options for downstream CRISPR gene-editing experiments.

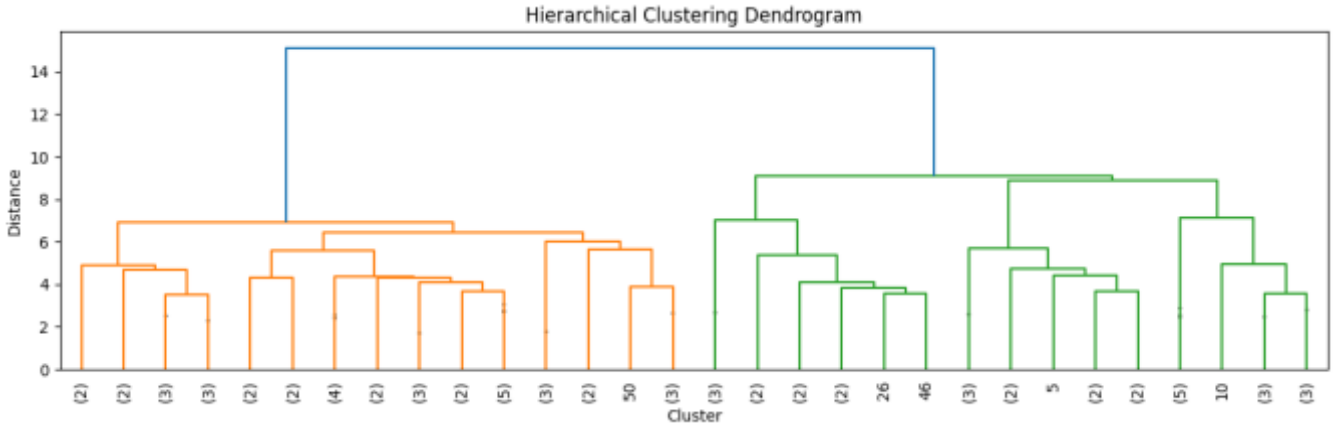

Figure 8: Hierarchical Clustering of Embedding Data

The hierarchical clustering performed with Ward's method produced a dendrogram that clearly separates the dataset into two main groups. These branches diverge at a relatively high linkage distance, indicating that the two major clusters differ substantially from each other.

Within each branch, several smaller subclusters are visible, with the shorter vertical distances reflecting closer relationships among those data points. The orange cluster shows more compact grouping, suggesting that the items within this branch share stronger similarity, whereas the green branch displays a wider range of linkage distances, pointing to greater variation among its members. Overall, the clustering pattern shows that the embedding data contain well-defined structure, allowing related entries to cluster together while maintaining distinct boundaries between the broader groups.

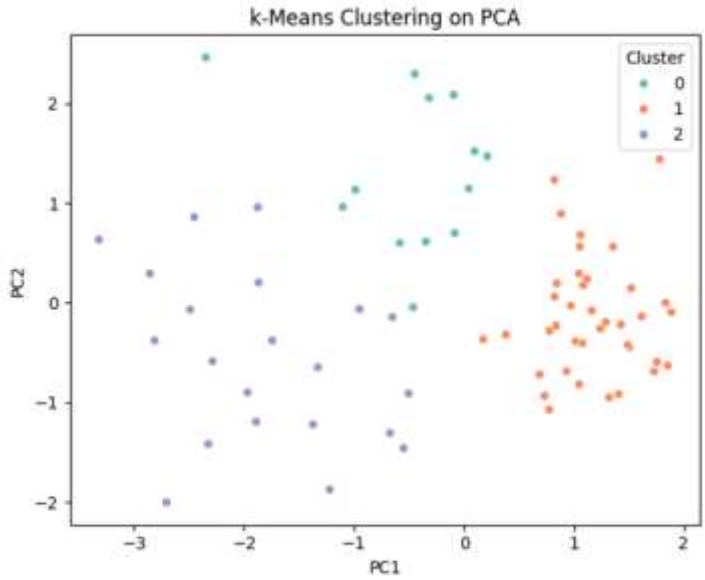

Figure 9: k-Means Clustering of PCA-Reduced Embedding Data

The k-means analysis performed on the PCA-reduced embedding data divided the samples into three well-defined clusters. The points assigned to each group form visually distinct regions on the PC1–PC2 plot, showing that the first two principal components capture enough variation to separate the data meaningfully. One cluster is tightly grouped on the right side of the plot, indicating a high level of similarity among those observations. Another cluster appears more spread out along the upper central area, suggesting moderate variability within that subset. The third cluster, located mainly toward the negative PC1 region, shows a wider distribution, reflecting greater internal differences between its members. The clear boundaries between these three groups demonstrate that the embedding features contain recognizable patterns, and the k-means algorithm was able to partition them into coherent and interpretable clusters.

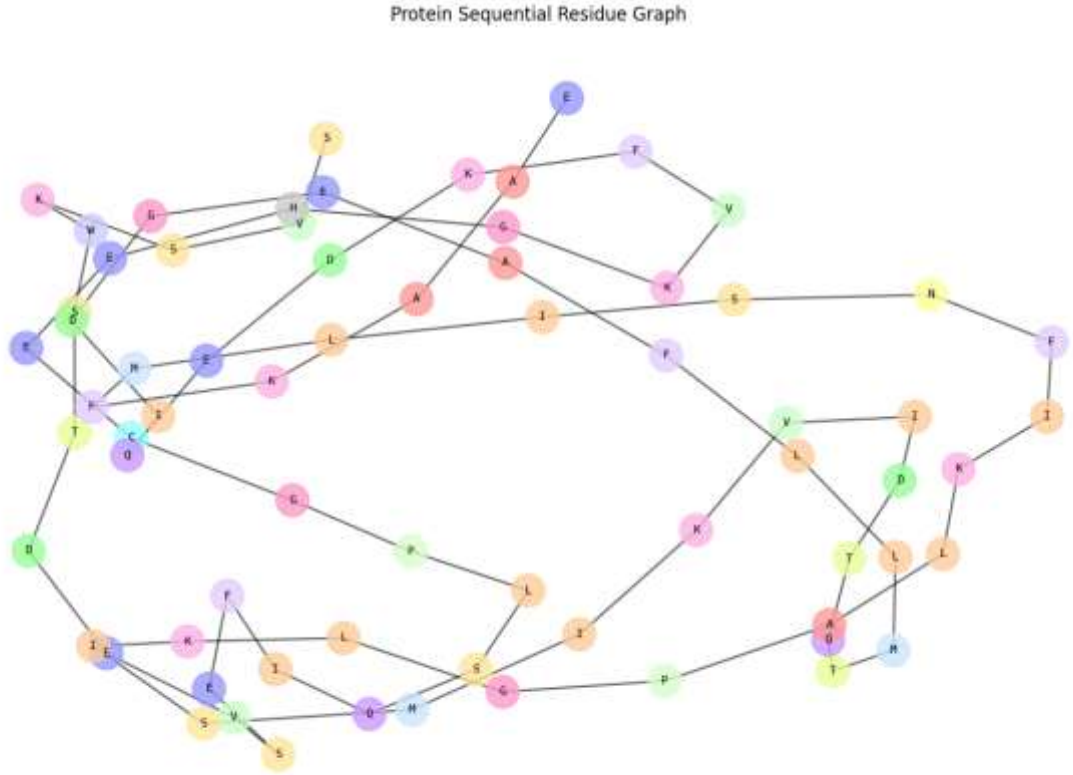

Figure 10: Protein Sequential Residue Graph Visualization

The amino-acid sequence was converted into a residue graph containing 70 nodes connected by 69 edges, with each node representing a single residue and each edge linking it to its immediate neighbor in the sequence. The spring-layout visualization arranges these nodes in a way that spreads them across the plot while preserving their sequential connectivity. Residues are color-coded according to their chemical properties, making it easy to spot regions enriched for particular amino-acid types as well as areas where the composition shifts. Although the graph encodes only linear relationships, the layout highlights subtle groupings and transitions within the sequence, offering a clear visual impression of its overall structure. This representation provides an informative overview of how the residues are distributed along the chain and can serve as a foundation for further structural or Functional Analysis.

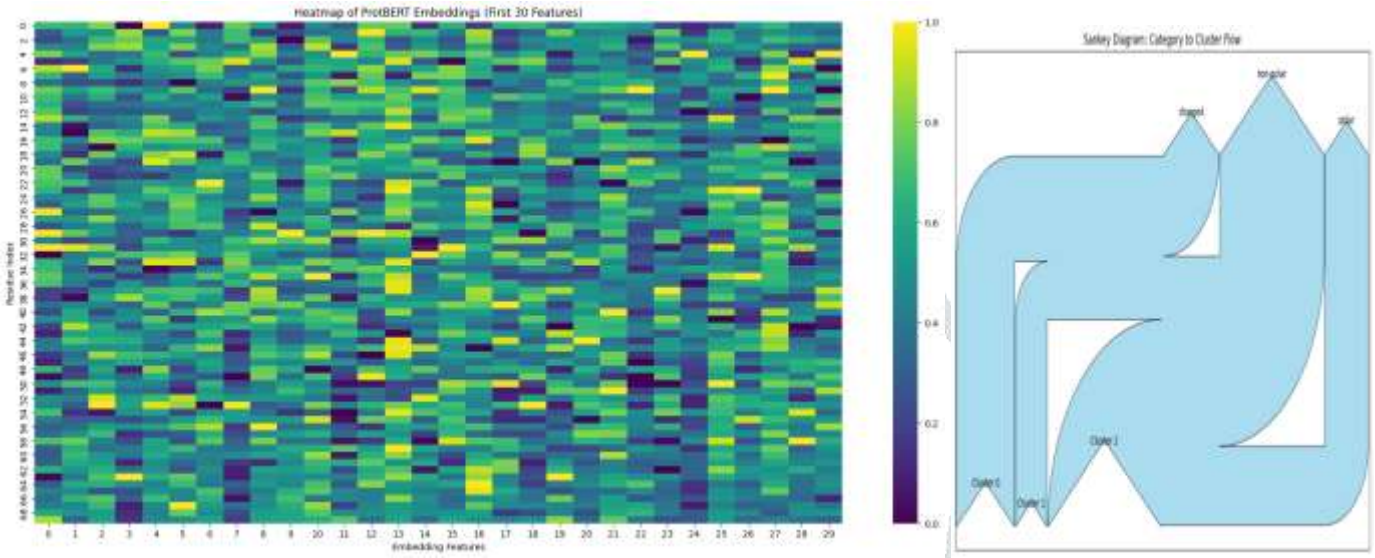

Figure 11: ProtBERT Embedding Patterns and Amino-Acid Category Flow into Clusters

The heatmap of the ProtBERT embeddings, built from the first 30 normalized features, shows that the residues in the protein sequence are represented with a wide range of feature intensities. The pattern is highly mixed, with alternating zones of strong and weak signal rather than large uniform blocks. This indicates that ProtBERT assigns a distinct embedding profile to each residue, capturing subtle differences in their biochemical environment and sequence context.

The Sankey diagram traces how three broad amino-acid categories-charged, polar, and non-polar-are distributed across the clusters obtained from k-means. The width of each flow reflects the number of residues from a category that end up in a given cluster. Non-polar residues contribute the largest share and appear across multiple clusters, whereas charged and polar residues show more selective associations. This suggests that the clustering reflects not only residue type but also additional contextual features encoded by ProtBERT. Taken together, the heatmap and the flow diagram show how the embedding model differentiates residues and how those distinctions influence their grouping in the clustering analysis.

IV. Conclusion

This study demonstrates the power of integrating CRISPR–Cas9 functional genomics with advanced machine learning approaches to accelerate target discovery and therapeutic innovation for malignant brain tumors. By focusing on the L3MBTL3 gene (PDB ID: 8Y76), we combined multi-layered bioinformatics analyses including structural interpretation, mutation profiling, sgRNA design, and protein language embeddings to uncover molecular features that may contribute to tumor development and therapeutic vulnerability.

Together, the integrated methodologies establish a robust computational framework for prioritizing gene targets, predicting functional impact, and guiding CRISPR-based experimental designs in malignant brain tumor research. The findings underscore the promise of combining high-throughput CRISPR technologies with ML-driven data interpretation to uncover tumor-specific dependencies and inform the development of precision therapies. Future work should focus on experimental validation of the predicted targets and sgRNAs, as well as expanding the ML models to include patient-derived multi-omics datasets for personalized therapeutic strategies. Ultimately, this integrative approach has the potential to accelerate the discovery of novel interventions for glioblastoma and other aggressive brain tumors, addressing an urgent need in modern oncology.

References

1.Uma Kumari, A., & Satheesan, S. (2025). Drug discovery and Biopython analysis MBP-MCL1 in myeloid cell leukemia. *Journal of Emerging Technologies and Innovative Research*, 12(1), f178–f187.

2. Kumari, U., & Gupta, S. (2023). NGS and sequence analysis with Biopython for prospective brain cancer therapeutic studies. *International Journal for Research in Applied Science and Engineering Technology*. https://doi.org/10.22214/ijraset.2023.50885

3. Uma Kumari, & Gurpreet Kaur et al. (2024). Biopython/network of protein identification and NGS analysis of glioma cancer ATP competitive type III C-MET inhibitor. *Journal of Emerging Technologies and Innovative Research*, 11(2), 41–51. http://doi.one/10.1729/Journal.40229

4. Kumari, U., & Agrawal, N. (2023). NGS and mutational profile analysis of non-small-cell lung carcinoma (NSCLC). *International Journal for Research in Applied Science and Engineering Technology*, 11, 3090–3094. https://doi.org/10.22214/ijraset.2023.50880

5. Uma Kumari, Johri, V., Dhopate, S., & Jha, T. (2024). Structure-based drug designing for the prediction of epitope for targeting malignant brain tumor. *Journal of Emerging Technologies and Innovative Research*, 11(7).

6. Bandbe, T., Saikia, J., & Kumari, U. (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

7. Kumari, U., & Bajaj, T. (2025). NGS data analysis and active site identification of 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

8. Kumari, U., & Renu et al. (2025). Structure analysis and molecular docking of mesothelin-207 fragment in human cancer. *Journal of Emerging Technologies and Innovative Research*, 12(3), g98–g106. http://doi.one/10.1729/Journal.44261

9. Kumari, U., Verma, G., & AnbuMegala, M. (2025). In silico drug design in human apoptosis-inducing factor (AIF) in lung cancer. *International Journal of Emerging Technologies and Innovative Research*, 12(2), ppd139–d148. http://doi.one/10.1729/Journal.43606

10. Kumari, U., & Mehrotra, K. (2025). NGS and proteomic gene expression analysis in studies of NUDT5 silence hormone signaling in breast cancer. *International Journal of Emerging Technologies and Innovative Research*, 12(2), ppc630–c638. http://doi.one/10.1729/Journal.43563

11. Chaudhary, S., & Kumari, U. (2024). 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), ppf116–f126. https://doi.one/10.1729/Journal.41696

12. Kumari, U., Belokar, P., Esompalli, A., Deshpande, S., Kumkar, M., & Tripathi, A. (2025). Next-generation sequencing to investigate the P53 cancer mutant Y234C for targeted cancer therapies. *IOSR Journal of Pharmacy and Biological Sciences*, 20(4), 63–72.

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

14. Jones, P., et al. (2014). InterProScan 5: Genome-scale protein function classification. *Bioinformatics*, 30(9), 1236–1240. https://doi.org/10.1093/bioinformatics/btu031

15. Landrum, M. J., et al. (2020). ClinVar: Improvements to accessing data. *Nucleic Acids Research*, 48(D1), D835–D844. https://doi.org/10.1093/nar/gkz972

16. Montague, T. G., Cruz, J. M., Gagnon, J. A., Church, G. M., & Valen, E. (2014). CHOPCHOP: A CRISPR/Cas9 and TALEN web tool for genome editing. *Nucleic Acids Research*, 42(W1), W401–W407. https://doi.org/10.1093/nar/gku410

17. Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

18. Phan, L., et al. (2025). The evolution of dbSNP: 25 years of impact in genomic research. *Nucleic Acids Research*, 53(D1), D925–D931. https://doi.org/10.1093/nar/gkae977

19. Schrödinger, LLC. (2023). *The PyMOL molecular graphics system* (Version 3.0) [Computer software].

20. 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

21. Sayers, E. W., et al. (2025). GenBank 2025 update. *Nucleic Acids Research*, 53(D1), D56–D61. https://doi.org/10.1093/nar/gkae1114

22. Szklarczyk, D., et al. (2019). STRING v11: Protein–protein association networks with increased coverage. *Nucleic Acids Research*, 47(D1), D607–D613. https://doi.org/10.1093/nar/gky1131

23. Tate, J. G., et al. (2019). COSMIC: The catalogue of somatic mutations in cancer. *Nucleic Acids Research*, 47(D1), D941–D947. https://doi.org/10.1093/nar/gky1015

24. The UniProt Consortium. (2023). UniProt: The universal protein knowledgebase in 2023. *Nucleic Acids Research*, 51(D1), D523–D531. https://doi.org/10.1093/nar/gkac1052

25. Waskom, M. L. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software*, 6(60), 3021. https://doi.org/10.21105/joss.03021

26. Burley, S. K., et al. (2025). Updated resources for exploring experimentally-determined PDB structures and computed structure models at the RCSB Protein Data Bank. *Nucleic Acids Research*, 53(D1), D564–D574. https://doi.org/10.1093/nar/gkae1091

27. Cock, P. J. A., et al. (2009). Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163

28. Cunningham, F., et al. (2022). Ensembl 2022. *Nucleic Acids Research*, 50(D1), D988–D995. https://doi.org/10.1093/nar/gkab1049

29. Elnaggar, A., et al. (2022). ProtTrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10), 7112–7127. https://doi.org/10.1109/TPAMI.2021.3095381

30. Gardiner-Garden, M., & Frommer, M. (1987). CpG islands in vertebrate genomes. *Journal of Molecular Biology*, 196(2), 261–282. https://doi.org/10.1016/0022-2836(87)90689-9

31. Heigwer, F., Kerr, G., & Boutros, M. (2014). E-CRISP: Fast CRISPR target site identification. *Nature Methods*, 11(2), 122–123. https://doi.org/10.1038/nmeth.2812

32. Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3), 90–95.