Dataset Collection and Processing

Introduction:

Lung adenocarcinoma (LUAD) is one of the most common forms of lung cancer, and its treatment response often depends on the genetic mutations present in key driver genes such as EGFR, KRAS, TP53, and ATM. Different mutations can alter how a gene behaves, disrupt normal cellular pathways, and influence whether a tumor becomes sensitive or resistant to specific therapies. Understanding these mutation patterns is therefore essential for improving therapy selection and advancing precision medicine.

In this project, we analyze mutation sequences of major LUAD genes and extract both deep learning–based embeddings (using DNABERT) and biologically meaningful sequence features. By combining contextual sequence representations with interpretable gene-level descriptors, we aim to predict mutation behaviour, structural impact, and potential therapy class in a structured and explainable manner. This approach helps connect genomic mutation data with clinically relevant therapeutic insights.

Collecting the Mutation Data:

To begin the project, I needed real mutation information from cancer patients rather than artificial sequences. For this purpose I used the cancer genomics resource cBioPortal. From there I selected the lung adenocarcinoma study belonging to the TCGA PanCancer Atlas collection.

Inside the study page there is a downloads section that provides mutation records in MAF format. After downloading it, I obtained a text file named data_mutations.txt.
This file does not contain DNA sequences. Instead, each row describes a mutation observed in a patient sample. It specifies the affected gene and the nucleotide change in coding DNA notation.

From this dataset I filtered only the genes relevant to lung adenocarcinoma such as TP53, KRAS, EGFR, BRAF, ALK and others. I also kept only SNP mutations because they represent single base substitutions and can be reconstructed reliably.

So at this point I had real patient mutation records but still no mutated gene sequences.

Getting the Normal Gene Sequences:

Since the mutation file only tells what changed and not the full sequence, I needed the normal gene sequence as a starting point.

For each gene I searched in the NCBI nucleotide database using its RefSeq ID and downloaded only the CDS (coding sequence) in FASTA format.
This sequence represents the healthy version of the gene and acts as the reference sequence.

Now I had two types of data:

- Mutation coordinates from TCGA
- Normal gene sequence from NCBI

Reconstructing the Mutated Gene:

The mutation dataset gives the position and the nucleotide change.
So I recreated the tumor gene by applying that change on the normal gene.

For every mutation entry:

1. I located the mutation position in the reference CDS
2. I replaced the original nucleotide with the mutated nucleotide
3. I saved the resulting sequence as a mutated gene

This process was repeated for all mutation entries.
Therefore the sequences were not generated randomly — they were reconstructed from real mutation records.

Identifying Mutation Positions:

To record the mutation information in a structured way, each mutated sequence was compared with the normal sequence nucleotide by nucleotide.

The position where the difference first appears was stored along with:

- original base
- mutated base

This produced a clean mutation dataset for analysis.

Redundant Sequence Removal using Cosine Similarity

After reconstructing mutated gene sequences for each patient, many sequences were found to be highly similar because different patients can share identical or near-identical mutation patterns. Such sequences do not contribute additional biological information and may bias downstream analysis.

To address this, each sequence was converted into a numerical vector based on nucleotide composition (frequency of A, T, G and C). Cosine similarity was then calculated between every pair of mutated sequences.

If the similarity between two sequences was greater than 0.80, they were considered redundant and only one representative sequence was retained. Sequences with similarity below this threshold were preserved as unique mutation profiles.

This step ensured that the final dataset contained biologically distinct gene variants rather than repeated mutation patterns, thereby improving the reliability of alignment scoring and comparative analysis.

Pairwise Alignment:

After reconstructing the variants, each mutated gene was aligned with its reference gene.

I used global alignment because the entire gene length needs to be compared.
Local alignment would only compare a matching region, but here the sequences differ only at specific positions while the rest of the gene remains the same.

So comparing the full sequence provides a proper similarity measure.

Alignment Score Calculation:

The alignment score was calculated using a simple similarity count:

match = +1
mismatch = 0
no gap penalty

Since most mutations are single nucleotide substitutions, the score roughly equals the gene length minus the number of changed bases.

This score indicates how similar the mutated gene remains to the normal gene.

Final Result for dataset retrieval :

Mutation data from the MAF file was mapped onto reference coding sequences of selected *Homo sapiens* cancer genes.
For each patient, all reported nucleotide substitutions were incorporated into a single reconstructed sequence. This produced realistic multi-mutation gene sequences instead of treating every mutation independently.

Sequences that did not contain valid nucleotide substitutions were ignored during reconstruction.

After reconstruction and redundancy filtering:

- Each sequence corresponds to a unique patient mutation pattern
- Duplicate mutation profiles were removed
- The dataset preserves mutation diversity instead of mutation repetition
- Sequences remain biologically close to the reference while still representing distinct variants

Features Relevant to the Dataset:

DNABERT-Based Features

When each mutated gene sequence is passed through the pretrained DNABERT model:

- The model outputs a 768-dimensional embedding vector
- Each dimension (Feature_1 to Feature_768) represents a learned contextual sequence pattern

These embeddings capture:

- Local nucleotide dependencies
- Mutation neighborhood context
- Long-range sequence interactions
- Latent structural tendencies
- Mutation signature patterns

Although these features do not have biological names, they encode high-level sequence behavior useful for distinguishing mutation profiles.

These features are especially useful for:

- Mutation behaviour clustering
- Therapy class discrimination
- Pattern recognition across genes

SELECTED FEATURES:

The following features were selected because they directly support predicting:

- Mutation Behaviour
- Structural Damage
- Targetability
- Therapy Class

Mutation Behaviour Features

1. Mutation Burden

Mutation burden reflects the overall genomic instability of a gene. High mutation burden is strongly associated with altered tumor biology and may influence response to immunotherapy and targeted therapy. In LUAD, genes with high mutation accumulation often indicate defective repair mechanisms or carcinogen exposure (e.g., smoking).

Relevance to Therapy Prediction:

- High mutation burden -potential immune checkpoint inhibitor sensitivity
- Indicates overall mutation pressure influencing gene function

2. Mutation Density

Mutation density normalizes mutation count by gene length, allowing fair comparison between genes of different sizes. Larger genes naturally accumulate more mutations; density corrects this bias.

Relevance to Therapy Prediction:

- Identifies genes disproportionately affected by mutations
- Helps detect genes undergoing strong selective pressure

3. Transition/Transversion Ratio (Ti/Tv)

Different mutational processes produce distinct base substitution patterns. LUAD, especially smoking-associated cases, shows characteristic transversion signatures (C>A).

Relevance to Therapy Prediction:

- Indicates underlying mutagenic mechanism
- Links mutation type bias to carcinogenic exposure

- Helps identify mutation signature–specific therapy patterns

4. Trinucleotide Mutation Context

Cancer mutations are not random; they depend on the surrounding nucleotide context. Trinucleotide patterns define mutation signatures.

Relevance to Therapy Prediction:

- Captures signature-level mutation behavior
- Helps identify repair pathway defects
- Distinguishes mutation processes influencing therapy sensitivity

5. CpG Mutation Frequency

CpG sites are prone to methylation-driven mutations (C>T transitions). These mutations influence gene regulation.

Relevance to Therapy Prediction:

- Reflects epigenetic instability
- Associated with aging-related and repair-deficient tumors
- May influence drug response indirectly through gene regulation

6. Hotspot Recurrence Score

Driver mutations tend to recur at specific positions. Recurrent hotspots indicate strong functional selection.

Relevance to Therapy Prediction:

- Hotspot mutations often define therapy targets (e.g., EGFR L858R)
- Distinguishes passenger mutations from driver mutations

Structural Damage Indicators

7. Mutation Type Category (Missense, Nonsense, Frameshift)

Different mutation types produce varying levels of structural disruption.

Relevance to Therapy Prediction:

- Frameshift/nonsense - likely loss of function
- Missense - altered but possibly functional protein
- Determines severity of gene disruption

8. Early Truncation Index

Mutation position relative to gene length influences functional impact.

Relevance to Therapy Prediction:

- Early truncations - severe functional loss
- Late mutations - partial impact
- Influences repair pathway disruption severity

9. Amino Acid Substitution Severity

Physicochemical differences between original and mutated residues determine structural damage.

Relevance to Therapy Prediction:

- Larger chemical shifts - higher structural instability
- Strongly affects protein function and drug binding

10. Domain Disruption Flag

Functional domains are critical for protein activity.

Relevance to Therapy Prediction:

- Mutations in drug-binding domains directly influence sensitivity/resistance
- Domain disruptions alter signaling pathways

Targetability Features

11. Gene Functional Category

Genes have distinct biological roles: oncogene, tumor suppressor, DNA repair gene.

Relevance to Therapy Prediction:

- Oncogenes - often directly targetable
- Tumor suppressors -usually resistant
- Repair genes -synthetic lethality targets

12. DNA Repair Pathway Indicator

Repair genes such as ATM are linked to homologous recombination deficiency.

Relevance to Therapy Prediction:

Repair defects -PARP inhibitor sensitivity

Indicates pathway-based therapeutic vulnerability

13. Loss-of-Function Probability

Combines mutation type and position to estimate functional loss.

Relevance to Therapy Prediction:

- Strong predictor of therapy sensitivity in repair-deficient tumors
- Directly supports classification like “Repair pathway disruption”

Sequence Complexity Features

14. Local GC Content

GC-rich regions influence mutation rate and repair efficiency.

Relevance to Therapy Prediction:

- GC bias linked to mutation susceptibility
- Reflects structural DNA stability

15. Shannon Entropy

Measures variability and sequence randomness.

Relevance to Therapy Prediction:

- High entropy - unstable regions
- Helps differentiate stable vs mutation-prone segments

DNABERT Feature Retrieval

DNABERT is a transformer-based deep learning model trained on genomic sequences. It learns contextual relationships between nucleotides in a similar manner to how BERT models learn contextual relationships between words in natural language.

The goal of using DNABERT is to capture complex mutation patterns and long-range dependencies within gene sequences that may not be detectable through handcrafted statistical features.

Step-by-Step Conceptual Pipeline

1. Tokenization of DNA Sequence into k-mers

Each input DNA sequence is segmented into overlapping k-length nucleotide units (k-mers).
For example, for k = 6:

Sequence:ATCGTACG

Tokenized as:ATCGTA, TCGTAC, CGTACG

This step converts raw nucleotide sequences into token units compatible with the transformer architecture.

Purpose:

- Captures local nucleotide patterns
- Preserves contextual adjacency

2. Passing Tokens Through Pretrained DNABERT

The k-mer tokens are fed into the pretrained DNABERT model.

Inside the model:

- Tokens are converted into embedding vectors.
- Self-attention layers compute contextual relationships between all tokens.
- Each token embedding is influenced by surrounding tokens.

3.Extraction of Hidden States

For each token, DNABERT produces a hidden state vector.

If the model embedding size is 768:

Each token - 768-dimensional vector.

Thus, for a sequence with N tokens:

You obtain an N × 768 matrix.

These hidden states encode high-level sequence information learned from large genomic corpora.

4. Mean Pooling for Fixed-Length Representation

Since sequences vary in length, a pooling operation is applied.

Mean pooling is used:

$$Feature_i = \frac{1}{N}\sum_{j=1}^{N} HiddenState_{j,i}$$

This averages token embeddings across the sequence.

Result:

One fixed 768-dimensional vector per sequence

Independent of sequence length

5. Final Output Representation

Each mutated gene sequence is converted into:

Feature_1 … Feature_768

Each dimension represents a learned latent pattern of mutation behavior and sequence context.

Although not directly biologically interpretable, these embeddings capture:

- Implicit mutation signatures
- Structural tendencies
- Contextual sequence relationships

This process transforms variable-length genomic sequences into standardized feature vectors suitable for classification.

Novelty of our project:

Instead of treating each mutation as a separate sequence, this project reconstructs one complete mutated gene sequence per patient by combining all mutations occurring in that patient. This better reflects real biological conditions where multiple mutations exist together in the same gene. After reconstruction, cosine similarity is used to remove repeated mutation patterns and keep only distinct mutation profiles. Finally, pairwise alignment is applied to verify that the sequences are valid variations of the reference gene. Thus, the project produces a realistic, non-redundant patient-level mutation dataset rather than a simple list of mutations.