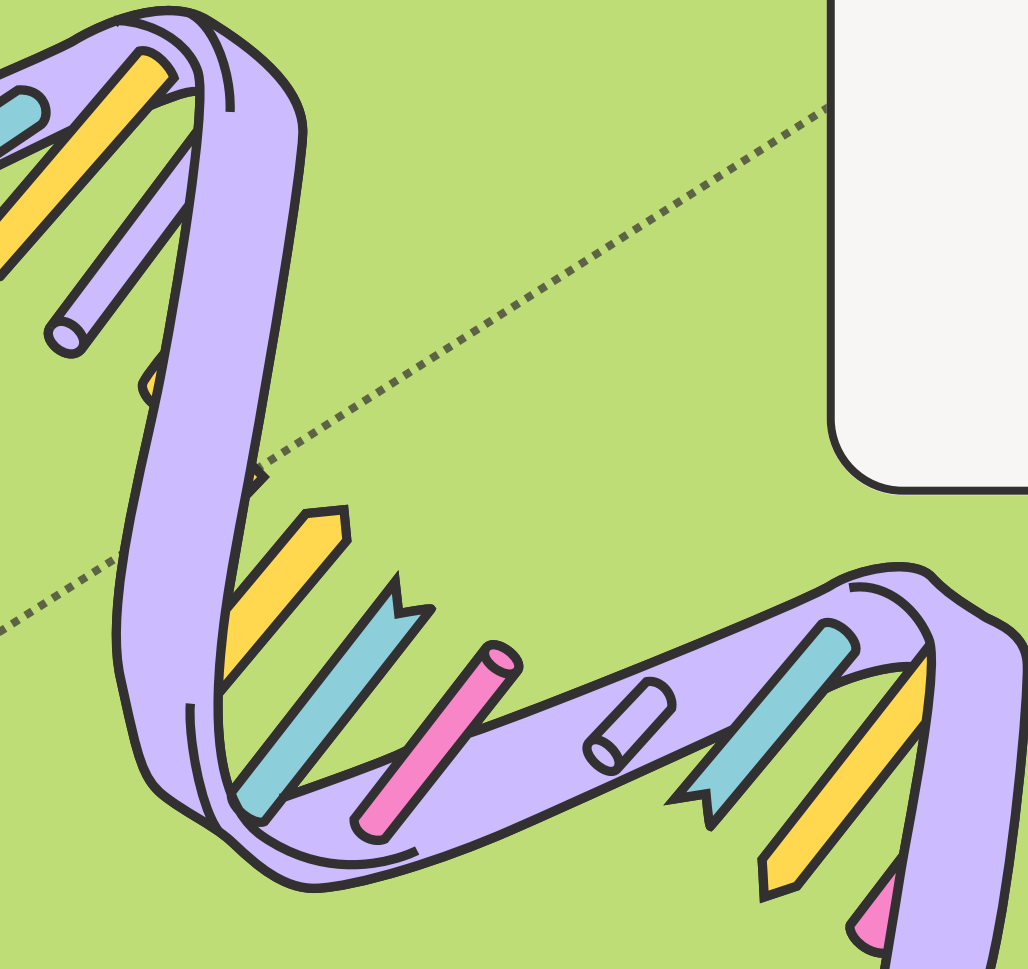
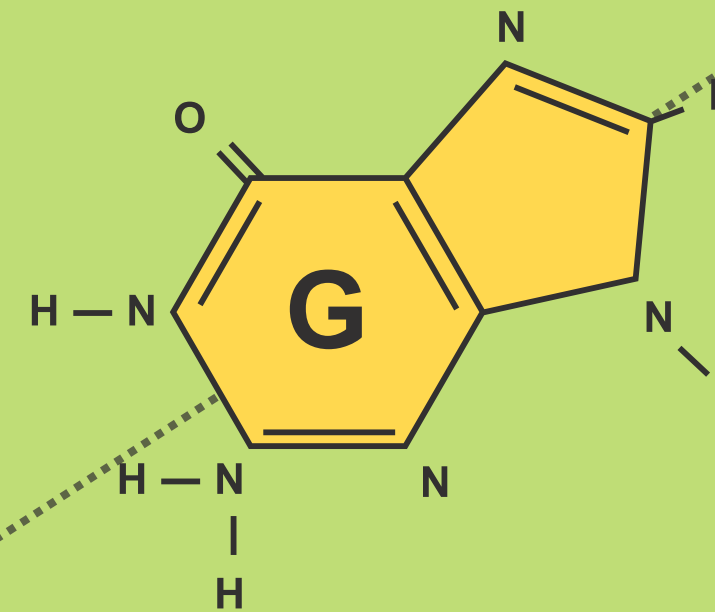
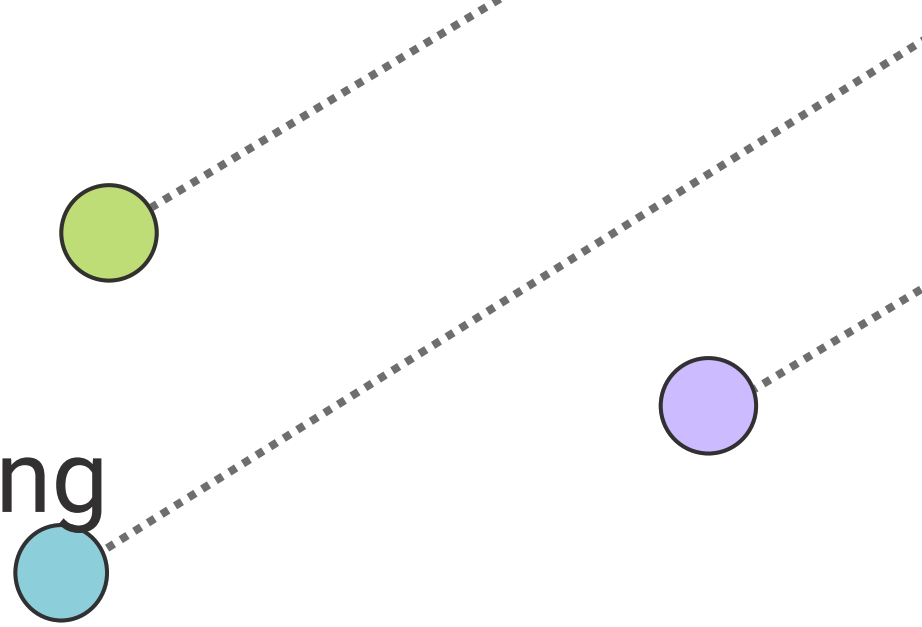


# **HIV'S** GENOMIC EVOLUTION AND ITS IMPACT

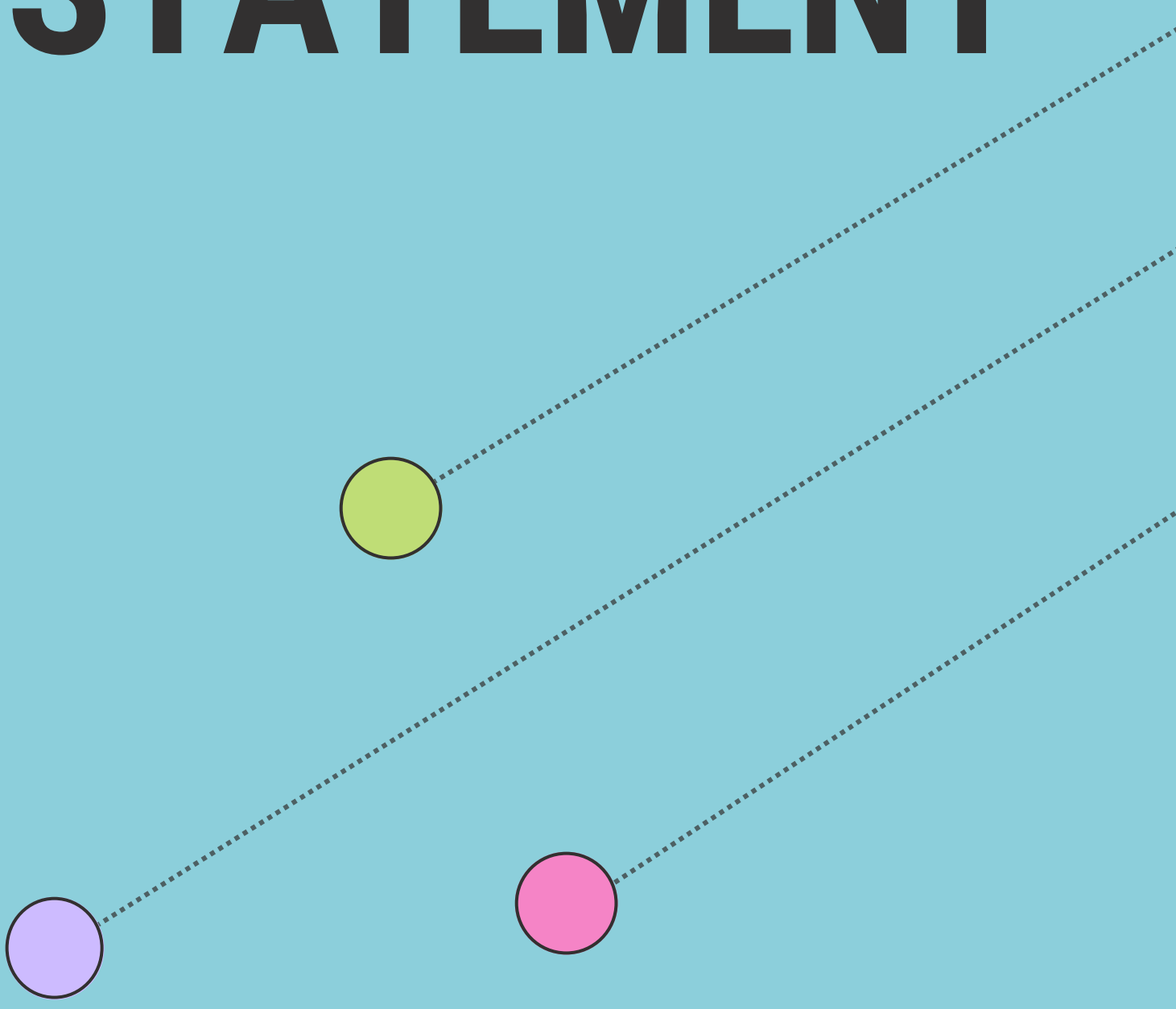


# QUICK RECALL:

- Human Immunodeficiency Virus (HIV) is a rapidly mutating retrovirus responsible for AIDS.
- Its high mutation rate leads to genetic diversity, affecting drug resistance and vaccine development.
- Understanding HIV's genomic evolution helps in tracking transmission, drug resistance and treatment strategies.
- HIV exists as multiple subtypes and recombinant forms, making it difficult to develop a universal vaccine.



# PROBLEM STATEMENT

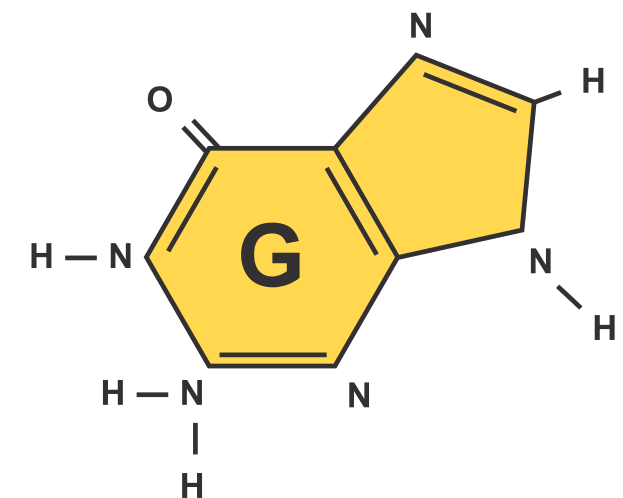
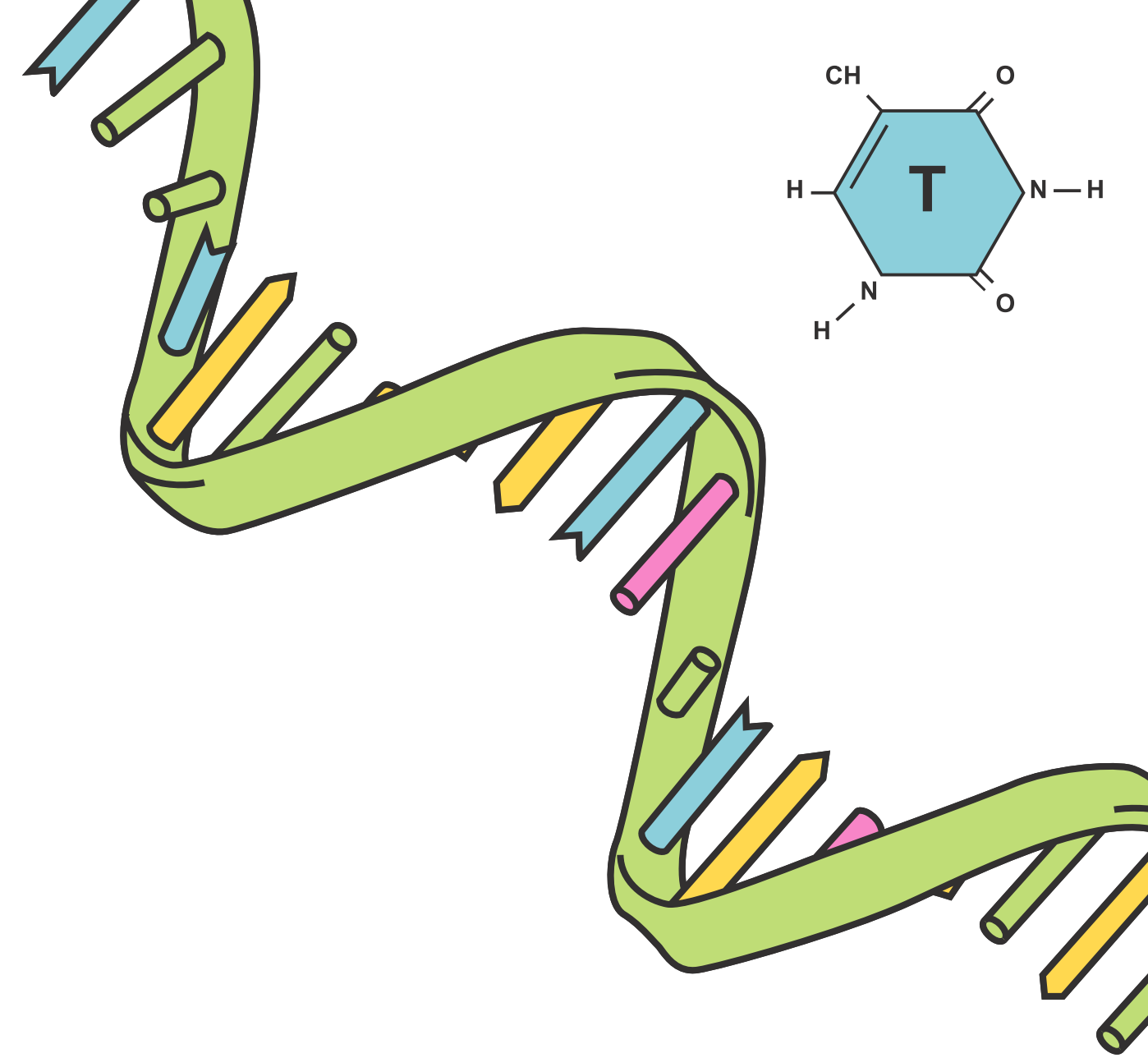


## Why Study HIV's Genomic Evolution?

- 1 HIV evolves through mutations and recombination, impacting treatment efficacy.
- 2 Drug resistance arises due to genetic variations, making treatment challenging.
- 3 Analyzing HIV sequences helps identify key mutations and predict future evolutionary trends.
- 4 HIV's rapid mutation hinders the creation of a universal vaccine, complicating prevention efforts.

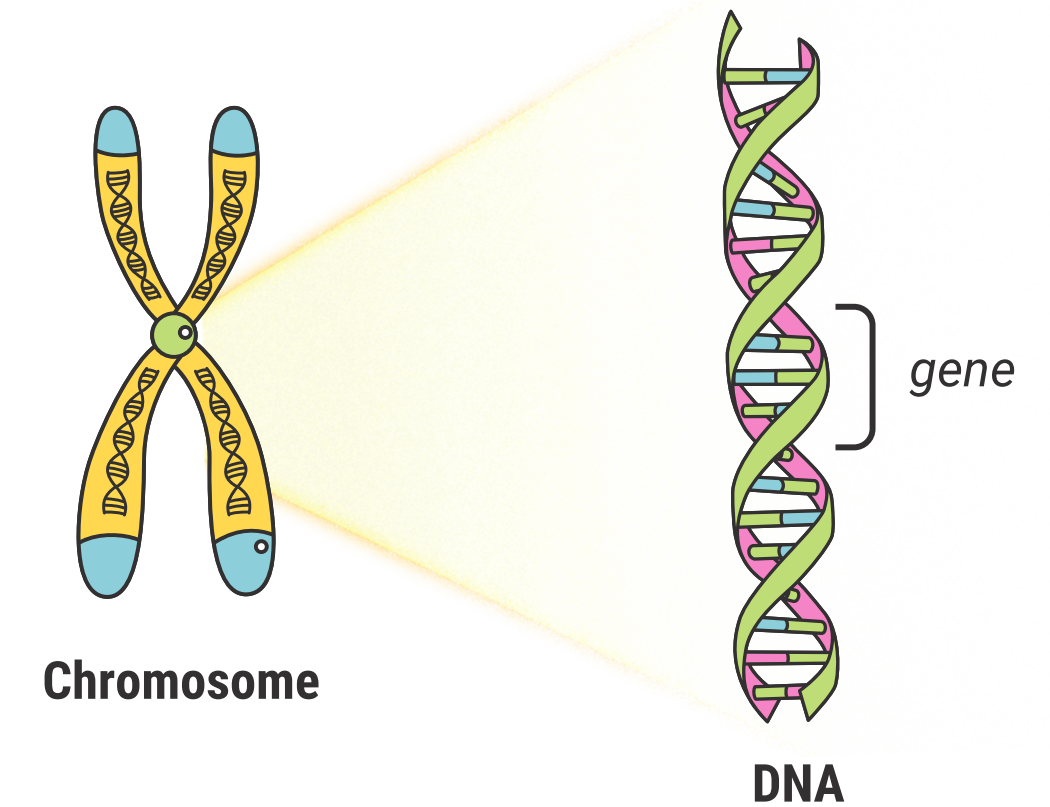
# METHODOLOGY & TOOLS

- Collecting HIV genome sequences from NCBI GenBank.
- Predicting Mutation Impact with SIFT.
- Using MEGA (Molecular Evolutionary Genetics Analysis) for phylogenetic tree construction to track evolution.
- Identifying drug-resistant mutations with HIV Drug Resistance Database.
- BioPython for automated sequence analysis and ExPASy tools for protein structure and functional analysis of HIV mutations.



# EXPECTED RESULTS

Our analysis aims to provide key insights into HIV's genomic evolution and its impact on treatment strategies.



## Mutations

Identification of common mutations in HIV genes affecting drug resistance.

## Evolution

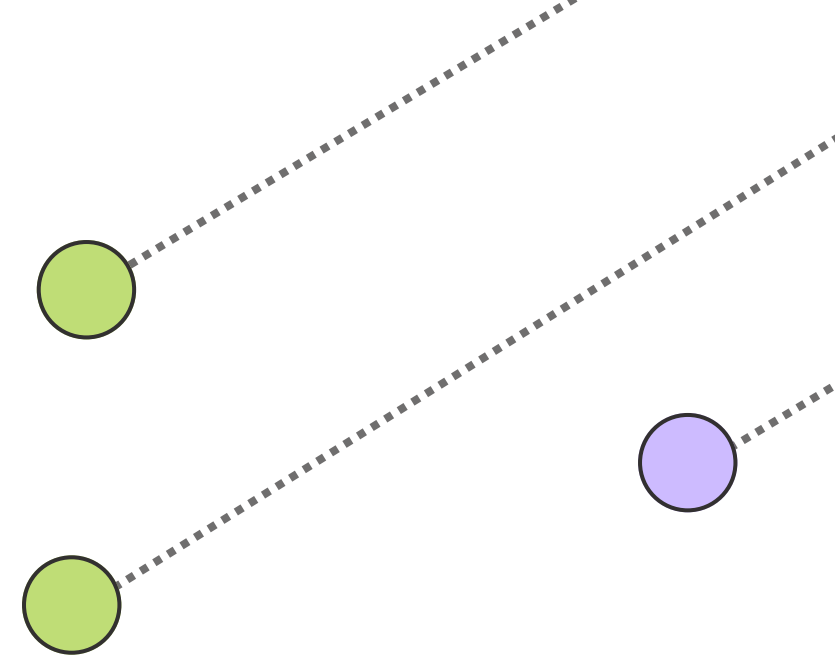
Phylogenetic analysis revealing the evolutionary trajectory of different HIV strains.

## Adaptation

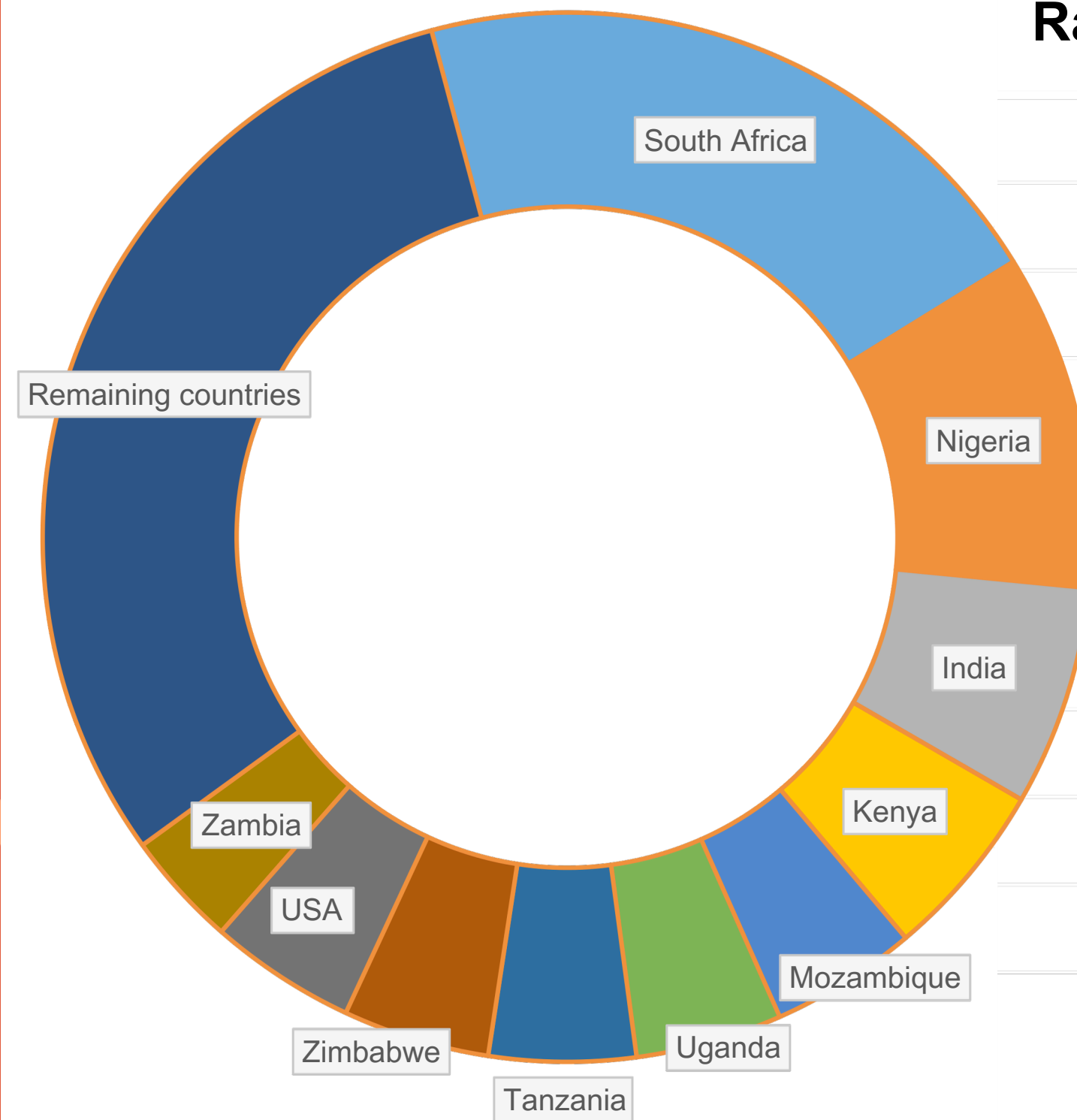
Insights into how HIV adapts to antiretroviral treatments, aiding in future drug development

# GOAL OF OUR PROJECT:

- How geography influences HIV genetic diversity.
- What that tells us about the spread, mutations, and possibly drug resistance.
- Finding region-specific evolutionary signatures or common hotspots in the genome.



# Top 10 countries: People living with HIV



Rank Country		% of people with HIV in the world
1	South Africa	18%
2	Nigeria	9%
3	India	6%
4	Kenya	5%
5	Mozambique	4%
6	Uganda	4%
7	Tanzania	4%
8	Zimbabwe	4%
9	USA	4%
10	Zambia	3%
Remaining countries		39%



# SELECTING THE GAG-POL REGION:

## *WHY GAG-POL?*

- After choosing five diverse countries for our analysis, we needed to select a genomic region to focus on.

### 1. HIV's Lifecycle

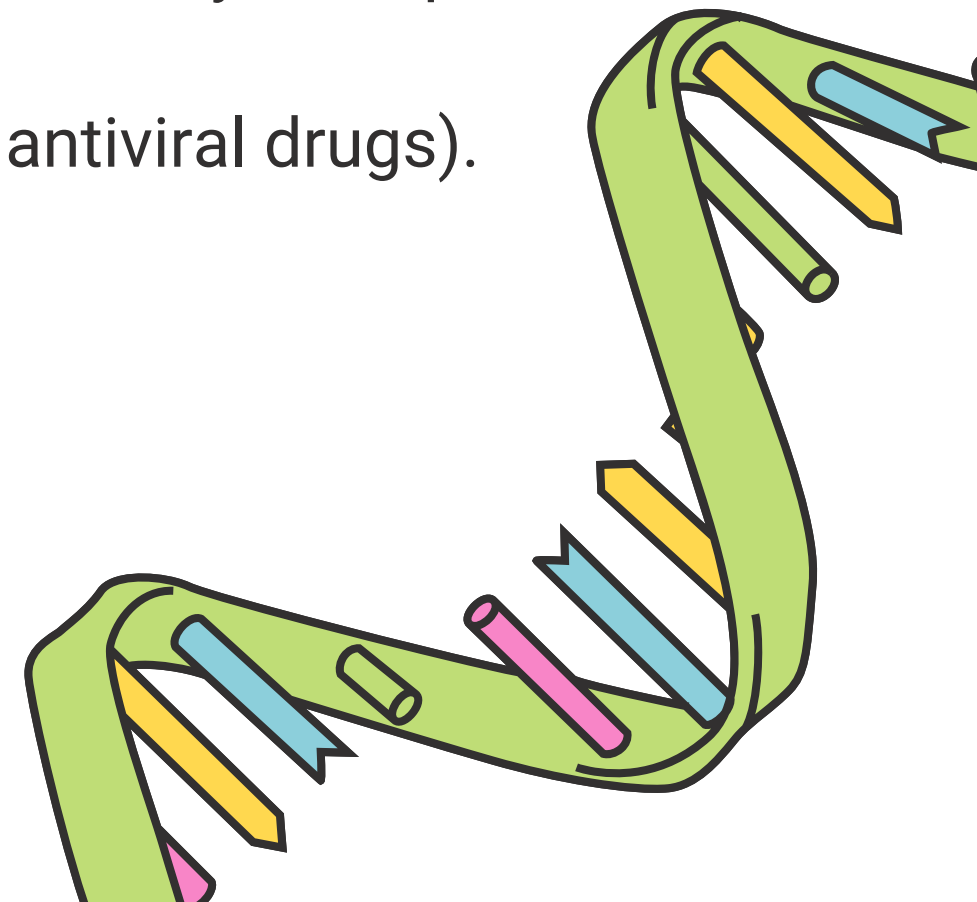
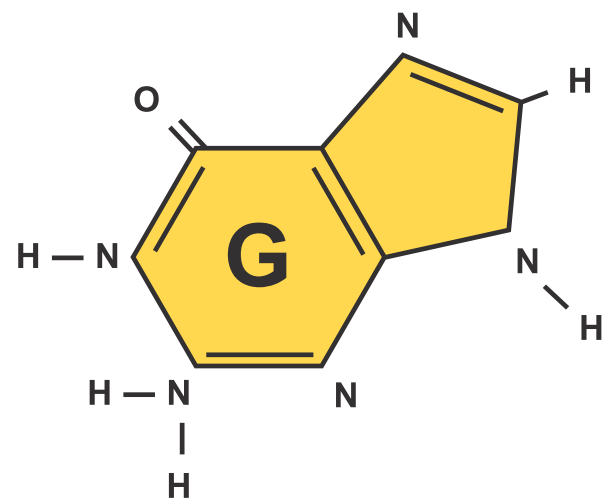
The Gag-Pol polyprotein is essential for virus replication, assembly and maturation.

It encodes:

- a) Gag: matrix (MA), capsid (CA), nucleocapsid (NC) — structural proteins
- b) Pol: protease (PR), reverse transcriptase (RT), integrase (IN) — enzymatic machinery for replication.

### 2. Evolutionary Insights

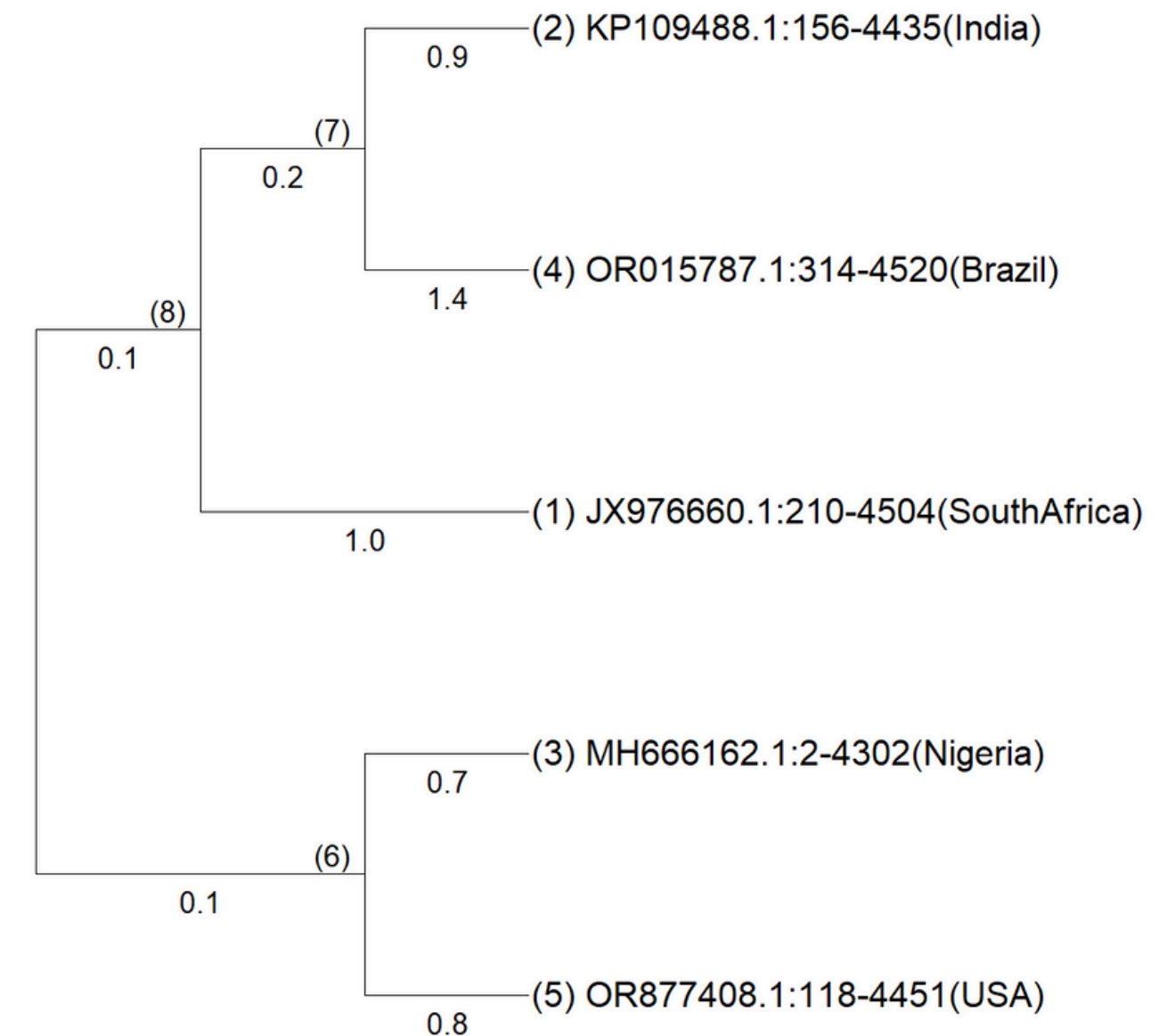
Gag-Pol accumulates mutations under selective pressure (e.g., immune response, antiviral drugs).





# EVOLUTION: PHYLOGENETIC TREE

- Strain (1) from South Africa diverged early from the rest, suggesting it is more distinct.
- Strains (2) and (4) (India and Brazil) are grouped together, indicating they share a more recent common ancestor.
- Strains (3) and (5) (Nigeria and USA) form another distinct clade, suggesting closer evolutionary relationships
- India and Brazil share a more recent common ancestor, as do the strains from Nigeria and the USA
- The strain from South Africa is more distinct in this dataset



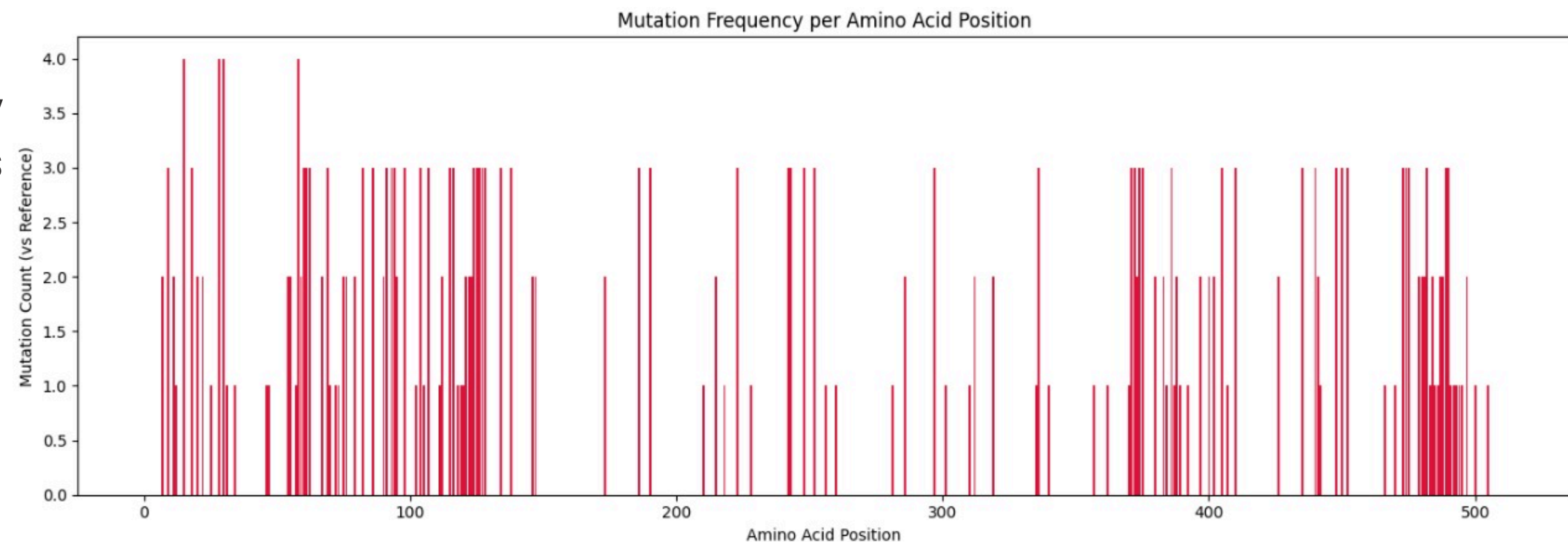
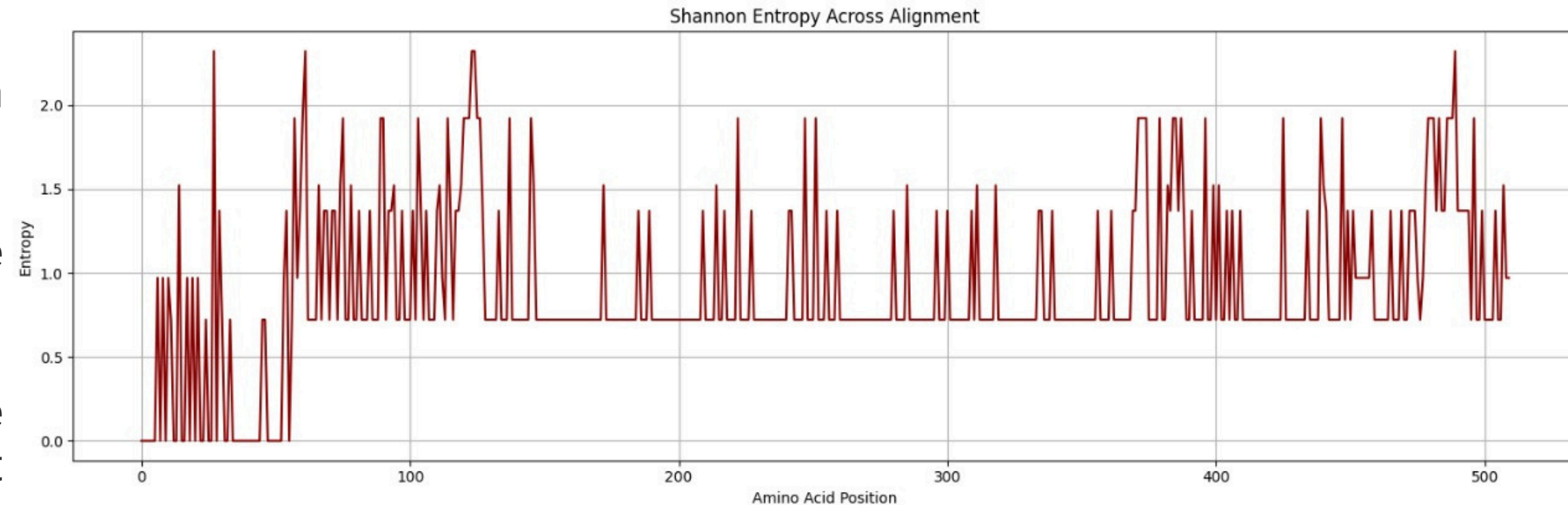
# ENTROPY AND MUTATION FREQUENCY

The Entropy plot shows the entropy at each amino acid position measures variability:

- Low entropy ( $\approx 0$ ): Highly conserved site (same amino acid in all sequences).
- High entropy (up to  $\sim 2.5$ ): Highly variable site (many different amino acids occur at that position).
- Flat line at a non-zero value, it likely means the distribution of amino acids is constant across those positions.
- Entropy is calculated using:

$$H = - \sum p_i \log_2 p_i$$

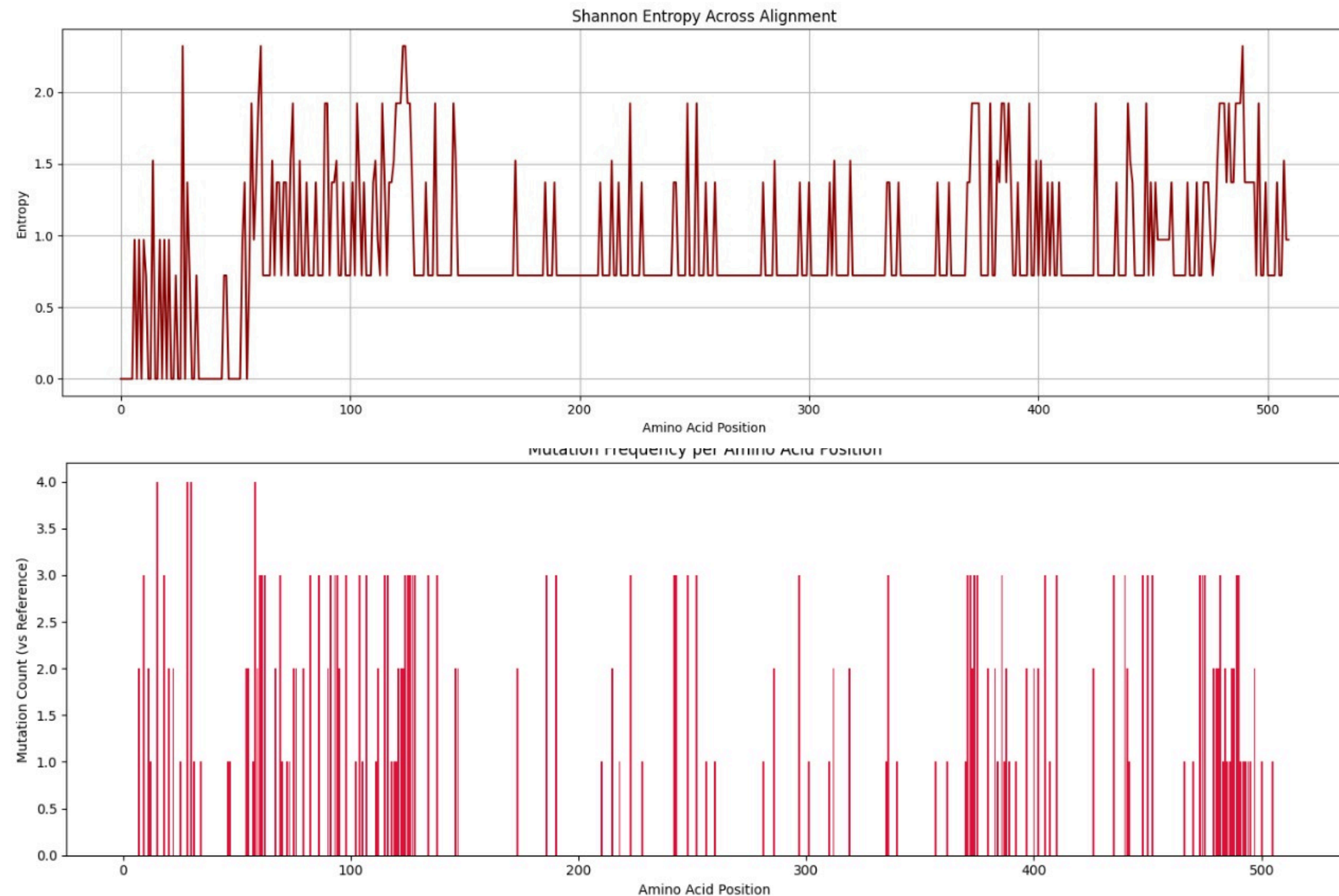
Where  $p_i$  is the frequency of each amino acid at a given position.



# ENTROPY AND MUTATION FREQUENCY

## Interpretation:

- **Gag** regions around positions 20–40, 80–100, 120–140 and 480–510 show sharp spikes in entropy ( $>2.0$ ), indicating high variability indicating likely hotspots of genetic variation across strains.(p17 and p24)
- Flat/near-zero regions (like 0–10, ~350–375) are conserved regions — often functionally important and intolerant to mutations.
- This bar plot visualizes the frequency of mutations at each amino acid position across all strains.



# BEFORE VS AFTER NORMALISATION

For any value  $x$ , its Z-score is:

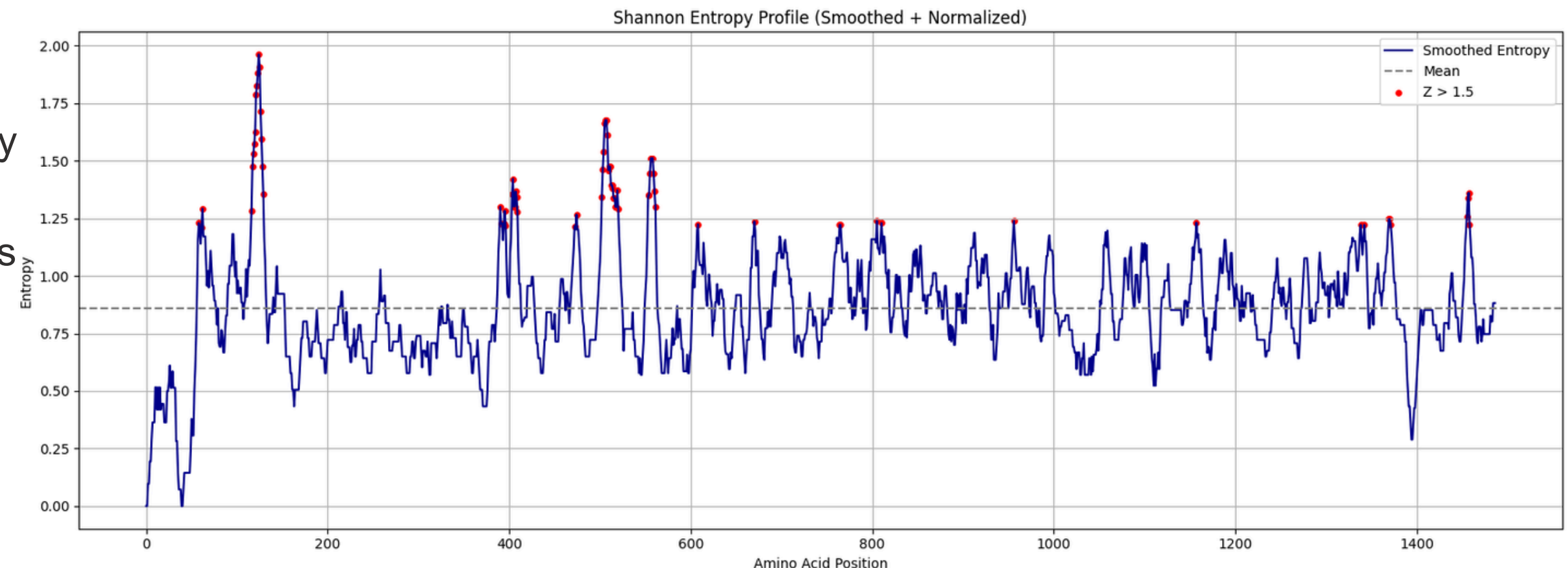
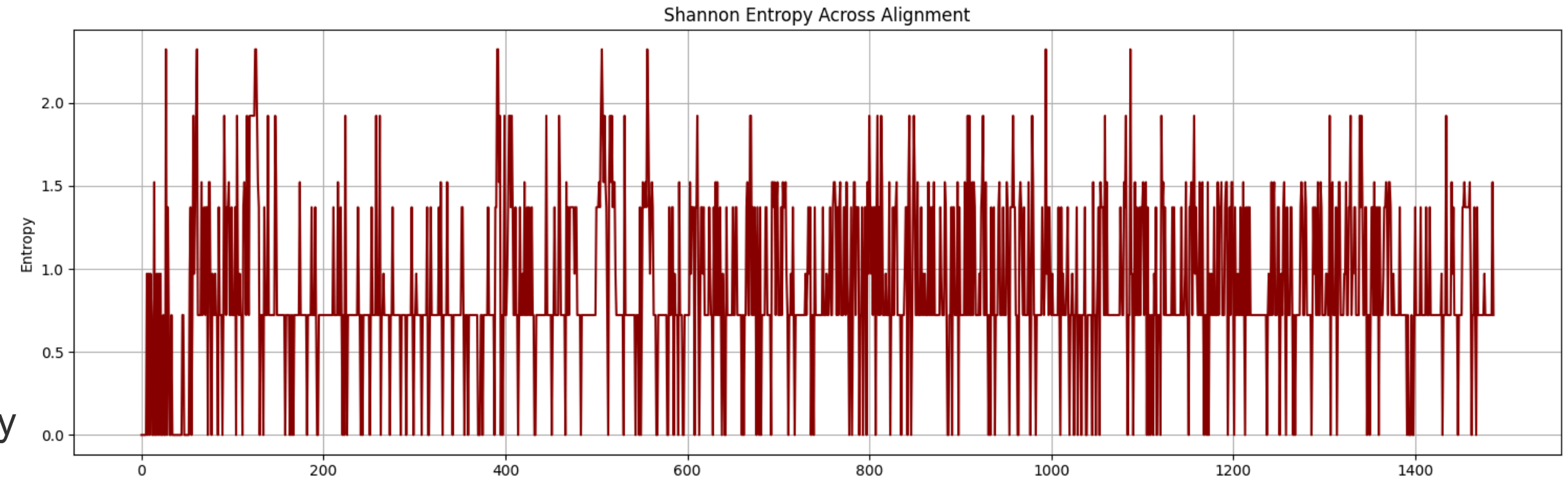
$$z = \frac{x - \mu}{\sigma}$$

Where:

- $x$  the entropy value at a specific position,
- $\mu$ : the mean of all entropy values,
- $\sigma$ : the standard deviation of the entropy values.

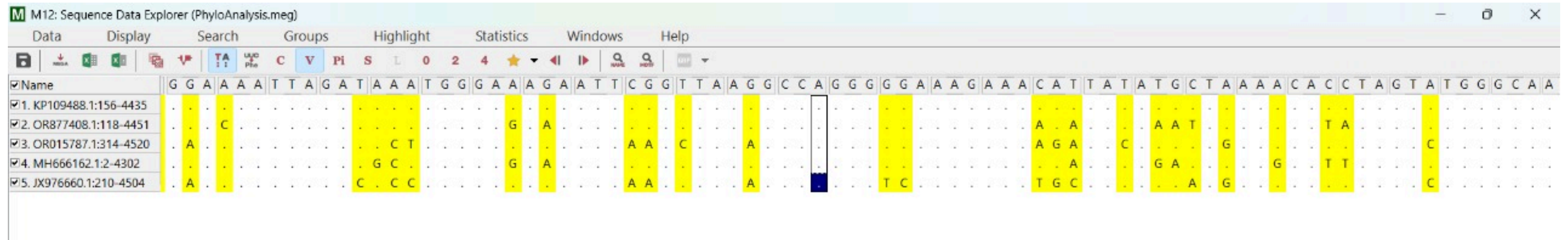
By using Z-normalization:

- Scaled all entropy values to a standard unit.
- Flag regions where entropy is significantly higher than average (e.g.,  $Z > 1.5$ ).
- This avoids arbitrary thresholds and gives a statistically robust method of detection.





# MULTI SEQUENCE ALIGNMENT



## Process:

- Performed MSA on the Gag-Pol gene sequences using tools like Clustal Omega or MAFFT.
- Aligned sequences against a reference strain (India) to ensure consistent positional comparison.

## Outcomes:

- Detected conserved vs. highly variable regions across strains.
- Enabled computation of mutation frequency plots, highlighting hotspots of amino acid variation.
- Formed the basis for Shannon entropy analysis, helping quantify variability per residue.
- Generated mutation matrices for heatmap visualizations to compare strain-level mutation densities.

# SIFT PROBABILITY MATRIX

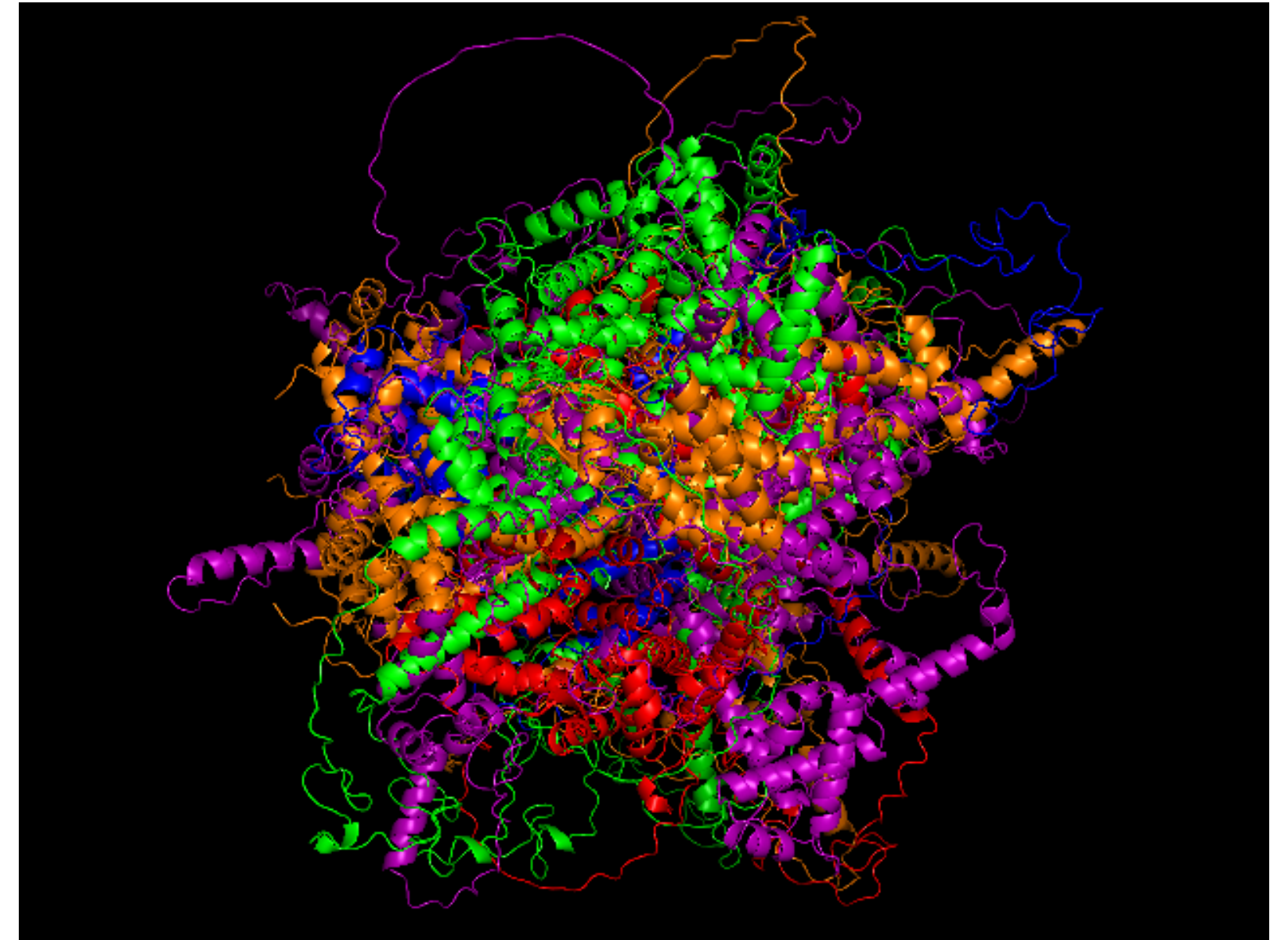
Predict Not Tolerated	Position	Seq Rep	Predict Tolerated
ywvtsrqpnmlkingfedca	1M	1.00	M
ywvtsrqpnmlkihfedca	2G	1.00	G
ywvtsrqpnmlkingfedc	3A	1.00	A
ywvtsrqpnmlkingfedca	4R	1.00	R
ywvtsrqpnmlkingfedc	5A	1.00	A
ywvtrqpnmlkingfedca	6S	1.00	S
hwqdnprgcksyfnta	7V	1.00	LIIV
ywvtsrqpnmlkingfedca	8L	1.00	L
w	9R	1.00	cfmyihvlpdgnqetakSR
ywvtsrqpnmlkihfedca	10G	1.00	G
w	11E	1.00	cmfiyhlpvrtqksnaGdE
cwfmfiyihpgldnsta	12K	1.00	erQK
ywvtsrqpnmlkingfedca	13L	1.00	L
ywvtsrqpnmlkingfedca	14D	1.00	D
w	15T	1.00	cfymhiplvgndqresATK
yvtsrqpnmlkingfedca	16W	1.00	W
ywvtsrqpnmlkingfdca	17E	1.00	E
cwfmdivgphslnta	18R	1.00	eqKR
ywvtsrqpnmlkhgfedca	19I	1.00	I
cwfmdivgphslnta	20K	1.00	eqKR
ywvtsrqpnmlkingfedca	21L	1.00	L
cwfmdivgphslnta	22K	1.00	eqKR
ywvtsrqpnmlkingfedca	23P	1.00	P
ywvtsrqpnmlkihfedca	24G	1.00	G
wmfihcrlrqvekp	25G	1.00	tdnaSG
ywvtsrqpnmlingfedca	26K	1.00	K
ywvtsrqpnmlingfedca	27K	1.00	K
	28R	1.00	wCmfipHvygdIntsQaeRK
wvtsrqpnmlkingfedca	29Y	1.00	Y
cw	30M	1.00	fyidvpHgMnlstaeqRK
dhgncwsrkpyqtaf	31L	1.00	viML
ywvtsrqpnmlingfedca	32K	1.00	K
ywvtsrqpnmlkigfedca	33H	1.00	H
dhgncwsrkpyqta	34L	1.00	fmvIL
ywtsrqpnmlkingfedca	35V	1.00	V
yvtsrqpnmlkingfedca	36W	1.00	W
ywvtsrqpnmlkingfedc	37A	1.00	A
ywvtrqpnmlkingfedca	38S	1.00	S
ywvtsrqpnmlkingfedca	39R	1.00	R
ywvtsrqpnmlkingfdca	40E	1.00	E
ywvtsrqpnmlkingfedca	41L	1.00	L
ywvtsrqpnmlkingfdca	42E	1.00	E
ywvtsrqpnmlkingfedca	43R	1.00	R
ywvtsrqpnmlkingedca	44F	1.00	F
ywvtsrqpnmlkingfedc	45A	1.00	A
dhwgncersqpkya	46L	1.00	fmiVL
wmifcvlyrphqatk	47D	1.00	esgDN
ywvtsrqpnmlkingfedca	48P	1.00	P
ywvtsrqpnmlkihfedca	49G	1.00	G
ywvtsrqpnmlkingfedca	50L	1.00	L
ywvtsrqpnmlkingfedca	51L	1.00	L
ywvtsrqpnmlkingfdca	52E	1.00	E
ywvtsrqpnmlkingfedca	53T	1.00	T
whyfmicrqlndke	54S	1.00	vptgSA
wcfmy	55E	1.00	ihvlprrtsGqankDE
ywvtsrqpnmlkihfedca	56G	1.00	G
whdnrqyekfmgps	57V	1.00	tliaVC
wcfym	58N	1.00	ihvplgdtSaeNQRK
cwfmfiyiv	59K	1.00	hpgldntsaerKQ

pos	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
1M	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2G	1.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3A	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4R	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
5A	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
6S	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
7V	1.00	0.03	0.01	0.00	0.01	0.02	0.01	0.00	0.74	0.01	0.12	0.02	0.01	0.01	0.00	0.01	0.01	0.03	1.00	0.00
8L	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
9R	1.00	0.43	0.05	0.25	0.35	0.06	0.28	0.13	0.10	0.69	0.18	0.06	0.30	0.18	0.31	1.00	0.79	0.42	0.16	0.03
10G	1.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
11E	1.00	0.54	0.05	0.69	1.00	0.06	0.59	0.13	0.08	0.43	0.14	0.05	0.46	0.23	0.34	0.23	0.45	0.27	0.14	0.02
12K	1.00	0.05	0.00	0.03	0.08	0.01	0.03	0.03	0.02	1.00	0.03	0.01	0.04	0.03	0.30	0.14	0.04	0.04	0.02	0.00
13L	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
14D	1.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
15T	1.00	0.75	0.06	0.40	0.63	0.06	0.29	0.11	0.17	1.00	0.25	0.09	0.35	0.23	0.41	0.45	0.64	0.91	0.29	0.02
16W	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
17E	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
18R	1.00	0.05	0.00	0.02	0.06	0.01	0.03	0.03	0.02	0.80	0.04	0.01	0.04	0.03	0.09	1.00	0.04	0.05	0.02	0.01
19I	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
20K	1.00	0.05	0.00	0.02	0.06	0.01	0.03	0.03	0.02	0.81	0.04	0.01	0.04	0.03	0.09	1.00	0.04	0.05	0.02	0.01
21L	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
22K	1.00	0.05	0.00	0.02	0.06	0.01	0.03	0.03	0.02	0.81	0.04	0.01	0.04	0.03	0.09	1.00	0.04	0.05	0.02	0.01
23P	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
24G	1.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25G	1.00	0.17	0.02	0.07	0.04	0.01	1.00	0.02	0.01	0.04	0.02	0.01	0.07	0.05	0.03	0.03	0.34	0.05	0.03	0.00



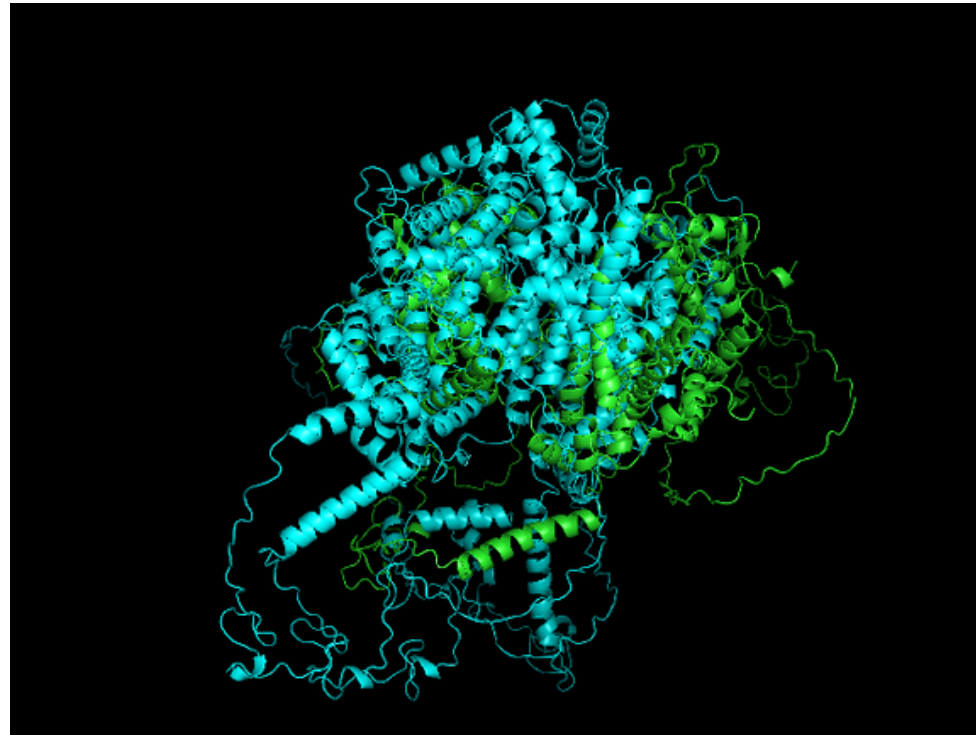
# PYMOL- 3D STRUCTURES

- Used PYMOL to superimpose different 3-d structures and see the key structural differences in them.
- It helps in understanding how different hiv strains from different regions differ in their structures and how it might impact their resistance towards medications.
- Such insights help in identifying mutations potentially linked to drug resistance, altered binding affinity, or immune escape, thereby informing more region-specific treatment strategies.

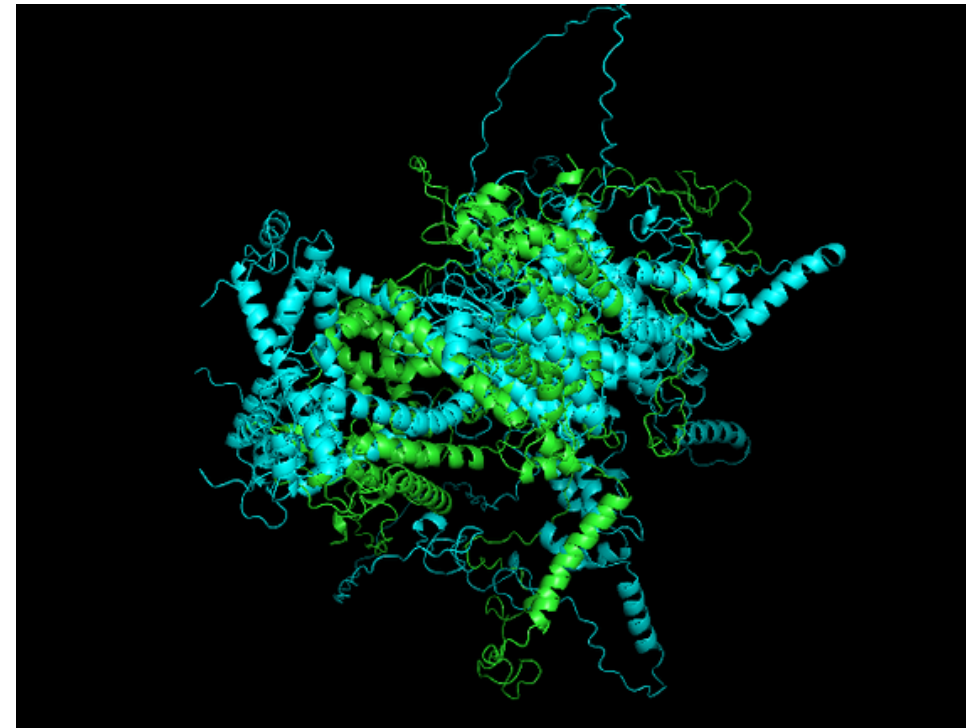


**MULTIPLE  
ALIGNMENT**

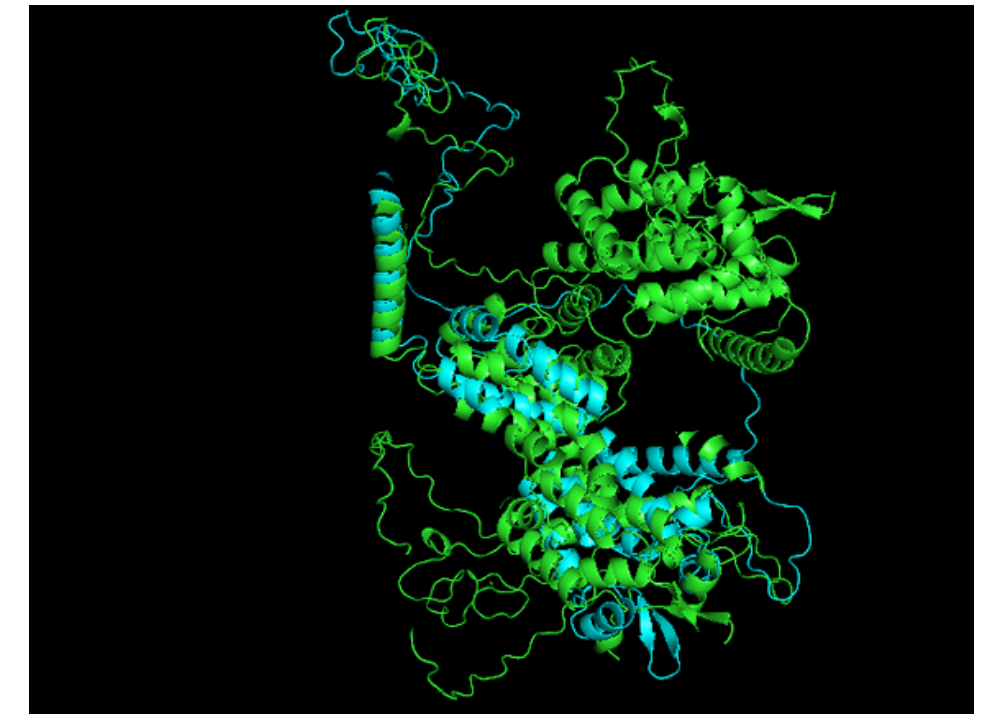




**INDIA-NIGERIA**



**INDIA - SA**

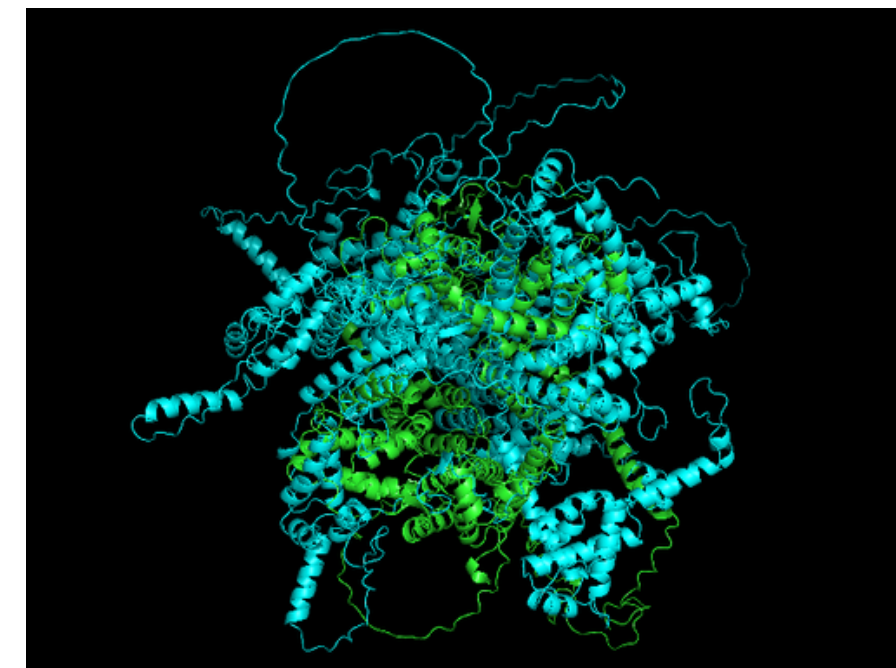


**INDIA - BRAZIL**

### **STRUCTURAL SIMILARITY (RMSD VALUES)**

- Low RMSD ( $< 2 \text{ \AA}$ ) between strains indicates high structural conservation, even if mutations are present.
- High RMSD ( $> 2\text{-}3 \text{ \AA}$ ) suggests conformational change due to mutations.

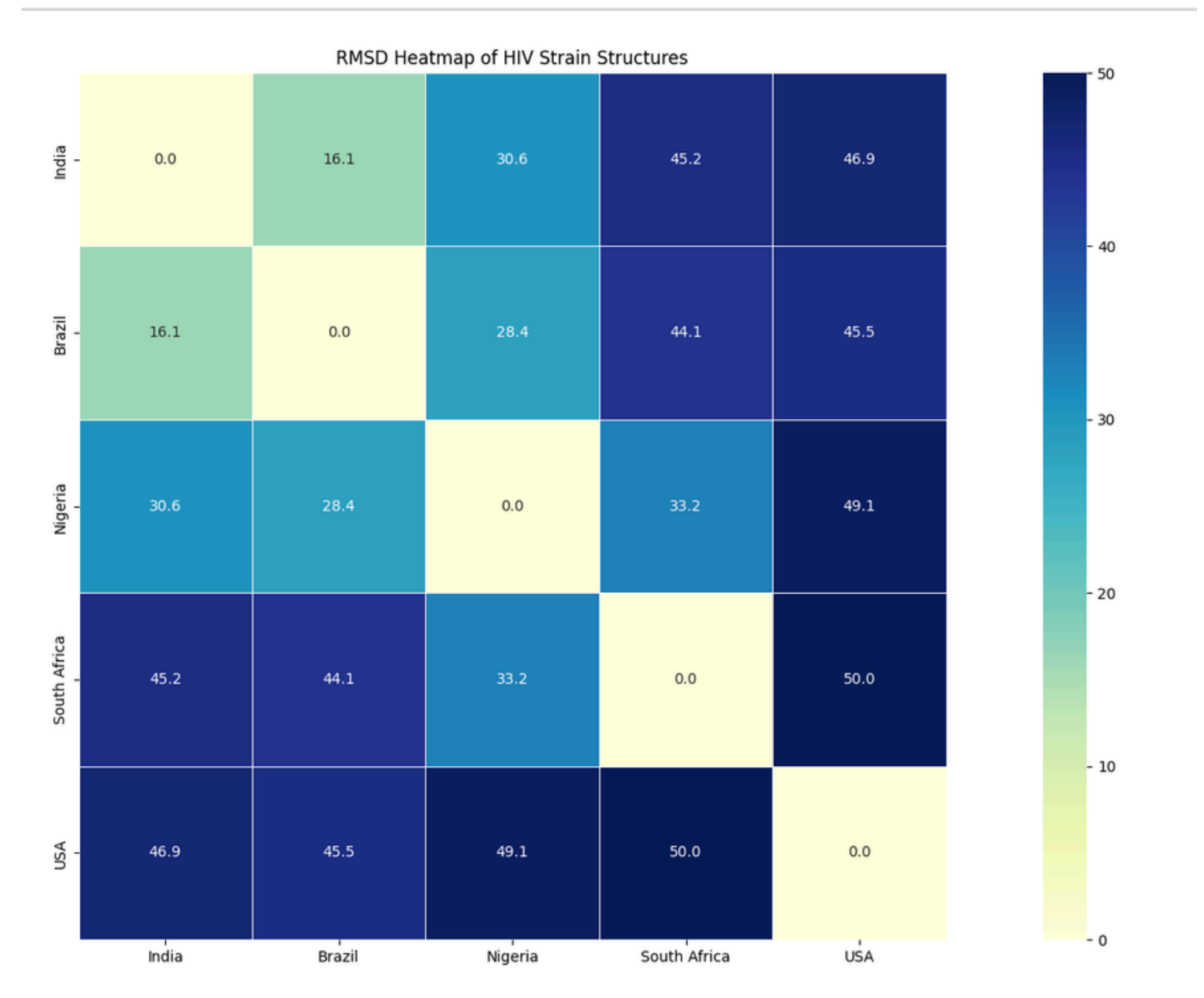
Conclusion: Mutations have little to no effect on structure (likely neutral) or significantly alter folding, which may impact function or drug interaction.



**INDIA - USA**

# RMSD VALUES

- India and Brazil are both structurally and evolutionarily closer, supporting co-evolution or shared ancestry.
- High RMSD despite phylogenetic proximity (e.g., South Africa or Nigeria–USA) might suggest post-translational or conformational factors that aren't evident from sequence alone.
- These differences can be biologically meaningful:
- Structural variations in gag-pol can affect viral replication efficiency, drug resistance, or immune evasion.
- Clades with high structural RMSD may respond differently to treatments.



# DRUG IMPACT

- As observed in the analysis, the HIV strains from Brazil and India exhibit significant similarities.
- Notably, these strains appear closely positioned in the phylogenetic tree, indicating a shared evolutionary lineage. This is further supported by their low RMSD (Root Mean Square Deviation) values when comparing their 3D structures, suggesting high structural similarity.
- Additionally, superimposition of their protein models in PyMOL revealed minimal conformational differences, particularly in the integrase region of the viral genome.

Integrase Strand Transfer Inhibitors	
bictegravir (BIC)	Susceptible
cabotegravir (CAB)	Susceptible
dolutegravir (DTG)	Susceptible
3://hivdb.stanford.edu/hivdb/by-sequences/report/?name=us	
25, 11:06 PM	
Sequence Analysis Report: user	
elvitegravir (EVG)	Susceptible
raltegravir (RAL)	Susceptible

BRAZIL

# DRUG IMPACT

Due to these structural and evolutionary similarities, it is likely that the same set of integrase strand transfer inhibitors (INSTIs) is effective against both strains. This reinforces the hypothesis that the integrase domain, which plays a critical role in HIV replication, remains conserved between these two regional variants, making them similarly susceptible to specific antiretroviral therapies.

Integrase Strand Transfer Inhibitors	
bictegravir (BIC)	Potential Low-Level Resistance
cabotegravir (CAB)	Low-Level Resistance
dolutegravir (DTG)	Potential Low-Level Resistance
elvitegravir (EVG)	Low-Level Resistance
raltegravir (RAL)	Low-Level Resistance

INDIA

# RESOURCES

## Sequence Data

- NCBI GenBank – for downloading HIV-1 gag-pol nucleotide sequences.
- <https://www.ncbi.nlm.nih.gov/genbank/>

## Phylogenetic Tree Construction

- MEGA X – software for building phylogenetic trees and evolutionary analysis.
- <https://www.megasoftware.net/>

## Mutation Analysis

- SIFT (Sorting Intolerant From Tolerant) – for predicting the functional effect of amino acid substitutions.
- <https://sift.bii.a-star.edu.sg/>

## Protein Structure Prediction

- AlphaFold Protein Structure Database (EMBL-EBI) – to generate and visualize 3D structures.
- <https://alphafold.ebi.ac.uk/>
- PyMOL – for visualizing RMSD and structural alignments.

## Stanford Drug Resistance Database

- <https://hivdb.stanford.edu/>