My work at the:

Center for Interdisciplinary Research and Education, India

In silico approach for peptide vaccine design against emerging pathogens using alignment-free sequence descriptors (AFSDs)

Supervisors:

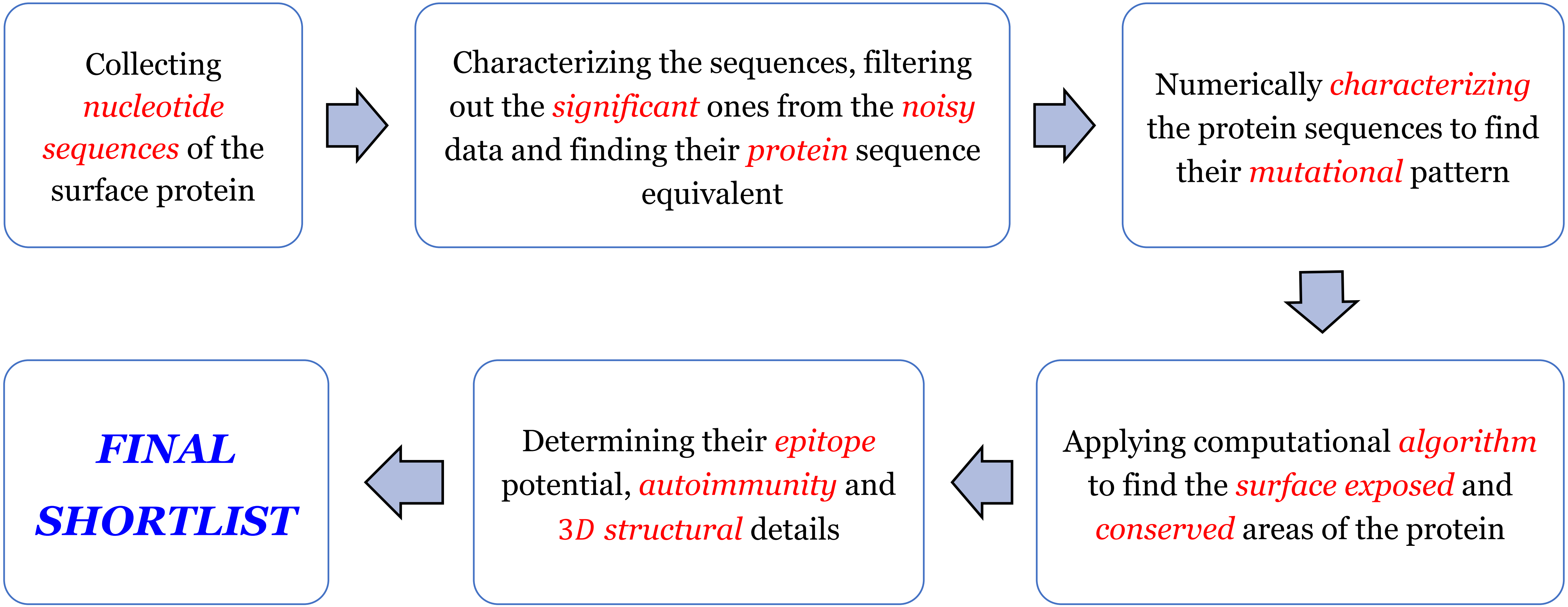
Dr. Ashesh Nandy

Center for Interdisciplinary Research and Education, India

Prof. Subhash C Basak

University of Minnesota Duluth

Overview of the protocol for vaccine design



Collecting *nucleotide sequences* of the surface protein

Characterizing the sequences, filtering out the *significant* ones from the *noisy* data and finding their *protein* sequence equivalent

Numerically *characterizing* the protein sequences to find their *mutational* pattern

Applying computational *algorithm* to find the *surface exposed* and *conserved* areas of the protein

Determining their *epitope* potential, *autoimmunity* and *3D structural* details

FINAL SHORTLIST

Graphical and numerical characterization of nucleotide sequences

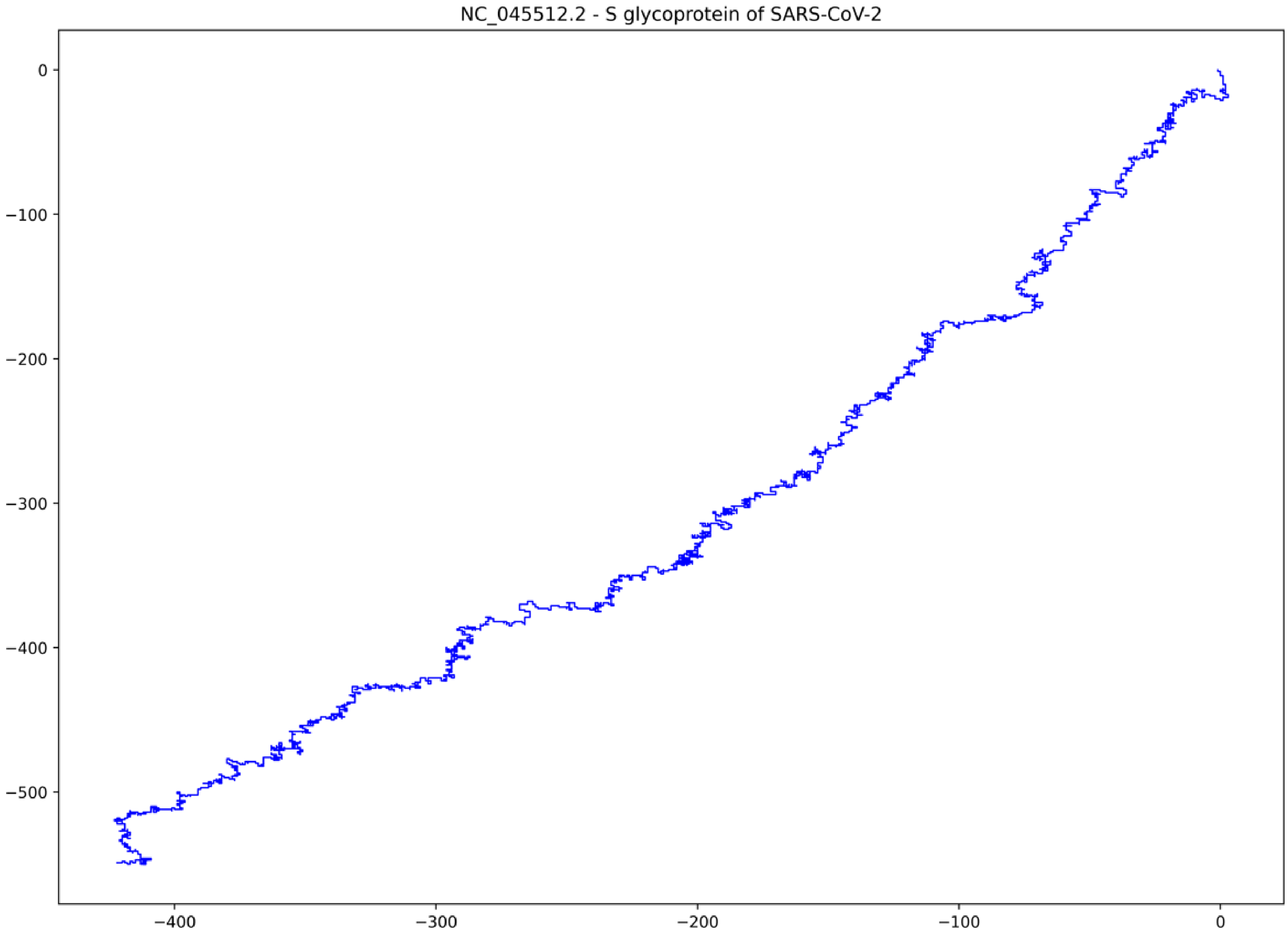
- ❑ Assign A, G, C, and T/U to $-X$, $+X$, $+Y$ and $-Y$ axes in a $2D$ Cartesian coordinate system respectively and start plotting from (0,0).

- ❑ Read the sequences base-by-base, move by +1 unit along the respective direction and keep plotting. The 1st order moments of the points along X and Y directions are:

$$\mu_x = \frac{\sum x_i}{N}, \qquad \mu_y = \frac{\sum y_i}{N}, \qquad where, N = number\ of\ plotted\ points$$

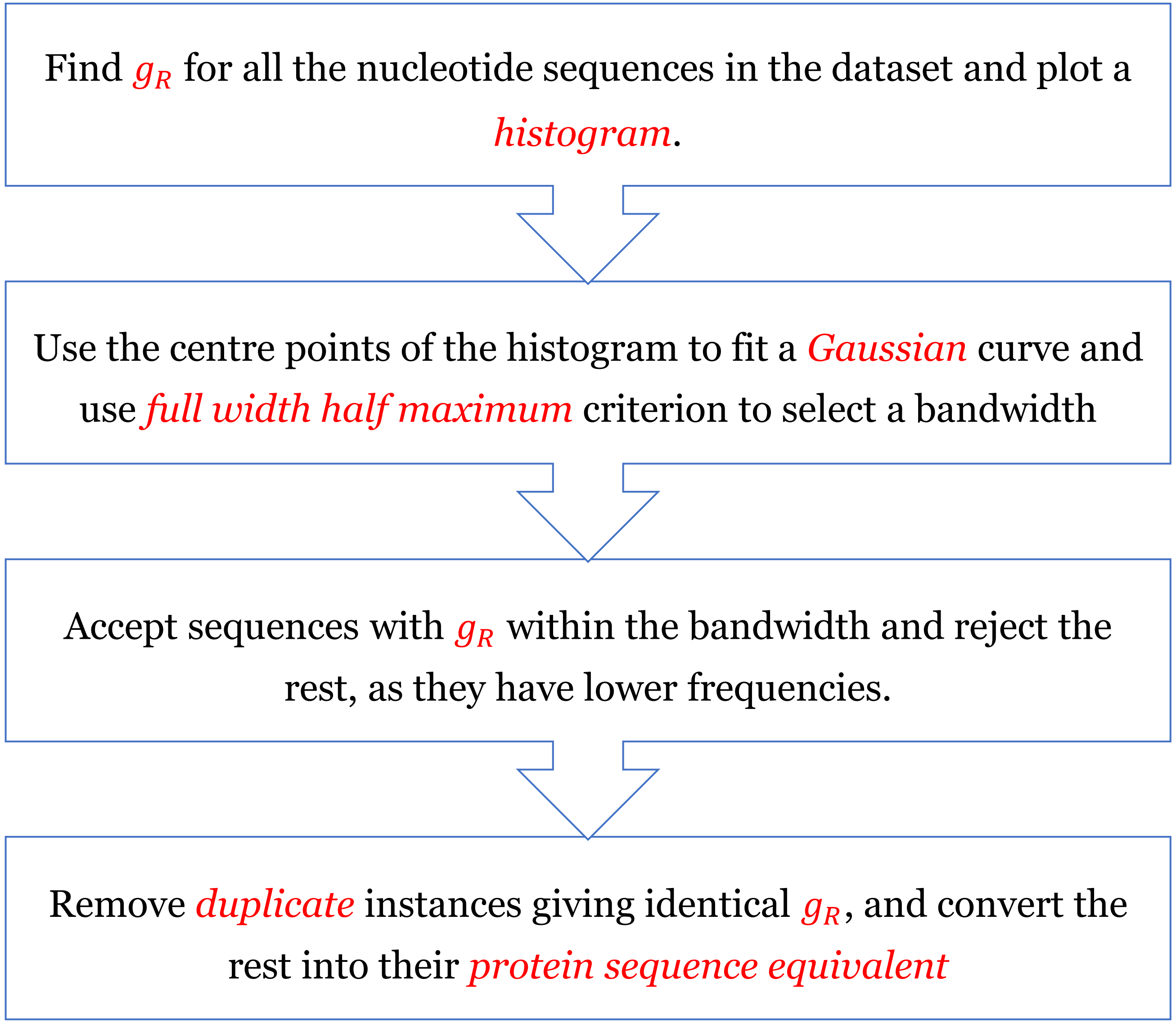
- ❑ Characterize the sequence with a quantity called “*Graph Radius* (g_R)” given as:

$$g_R = \sqrt{{\mu_x}^2 + {\mu_y}^2}$$



Plot for the nucleotide sequence of surface glycoprotein of SARS-CoV-2

Filtering the data

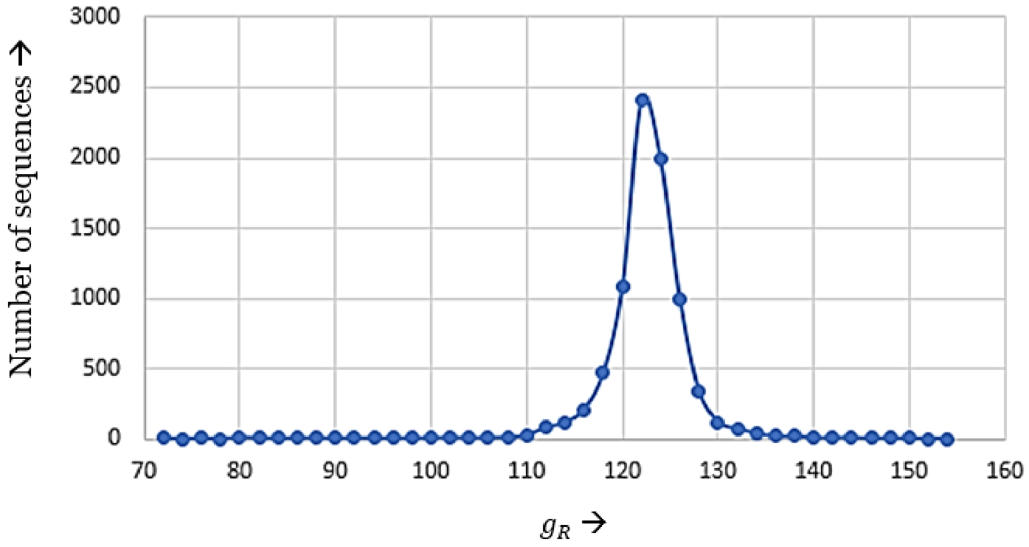


Find g_R for all the nucleotide sequences in the dataset and plot a *histogram*.

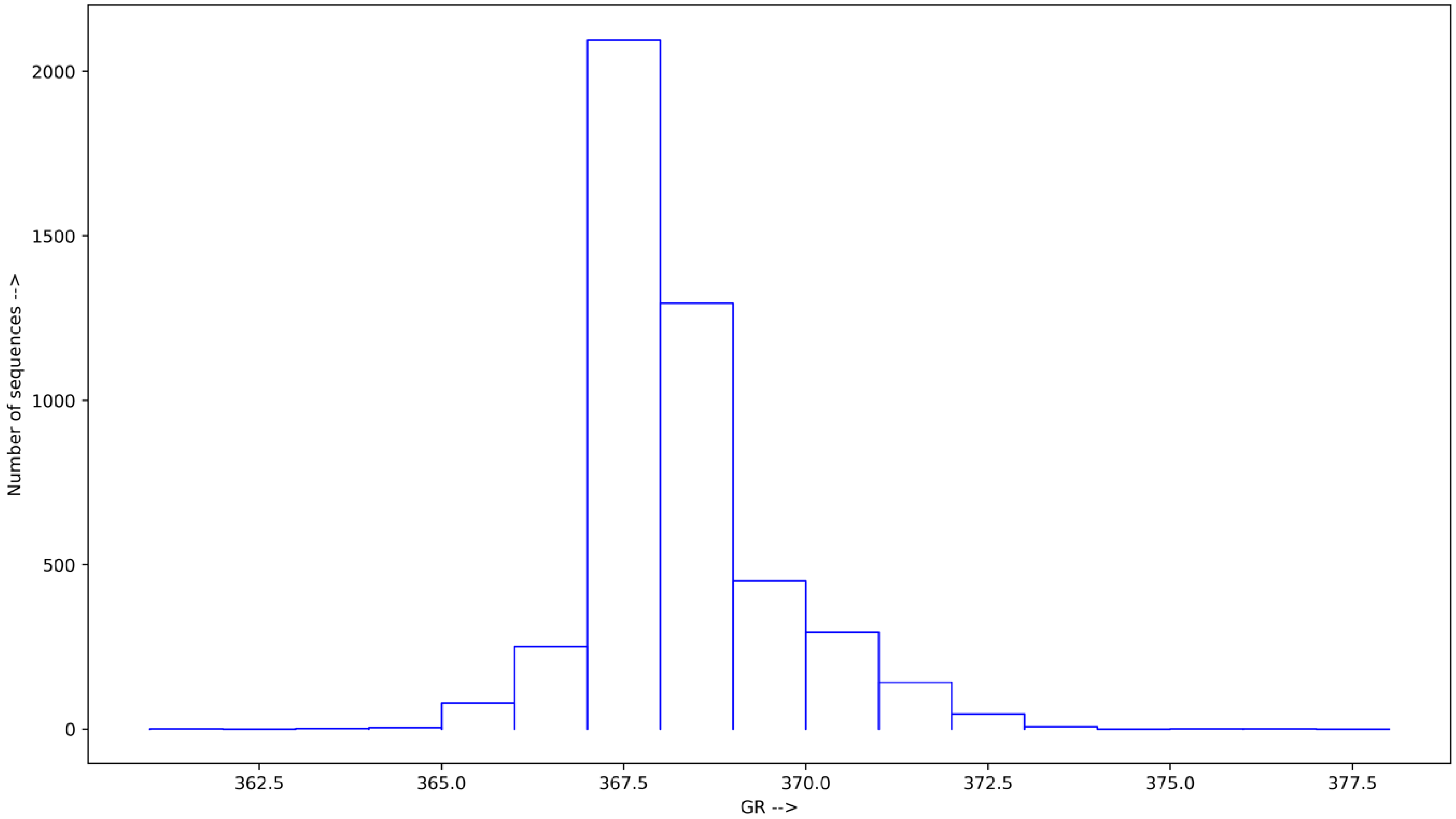
Use the centre points of the histogram to fit a *Gaussian* curve and use *full width half maximum* criterion to select a bandwidth

Accept sequences with g_R within the bandwidth and reject the rest, as they have lower frequencies.

Remove *duplicate* instances giving identical g_R, and convert the rest into their *protein sequence equivalent*



Gaussian curve for the surface-situated *Hemagglutinin (HA)* protein of H1N1, (sequences collected from 1910 to 2018)



The above histogram was obtained for *spike protein* of SARS-CoV-2.

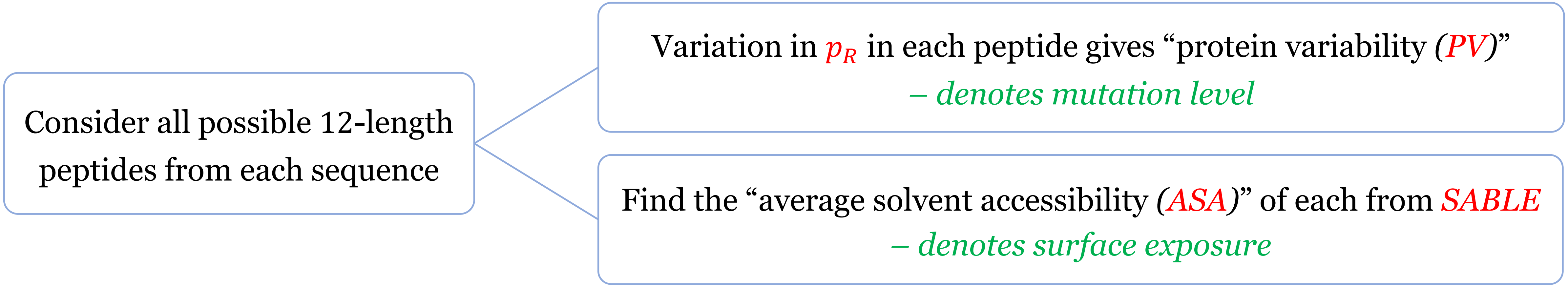
The bandwidth approximately ranged from 367 to 369.

Characterizing the protein sequences and quantifying their properties

- Imagine a *hypothetical* $20D$ system with each axis denoting an amino acid and apply a similar method for plotting the filtered protein sequences - the 1^{st} order moments will be:

$$\mu_1 = \frac{\sum x_{1_i}}{N}, \qquad \mu_2 = \frac{\sum x_{2_i}}{N}, \ldots\ldots\ldots\ldots\ldots, \mu_{20} = \frac{\sum x_{20_i}}{N}$$

- Characterize the sequence using a quantity, "*protein radius (p_R)*", defined as: $p_R = \sqrt{{\mu_1}^2 + {\mu_2}^2 + \cdots + {\mu_{20}}^2}$



Consider all possible 12-length peptides from each sequence

Variation in p_R in each peptide gives "protein variability (*PV*)"
– *denotes mutation level*

Find the "average solvent accessibility (*ASA*)" of each from *SABLE*
– *denotes surface exposure*

Finding the surface-exposed and conserved protein regions

Step I: w parameter

- ❑ For a 12-length peptide, define its *w parameter* as:

$$w = (ASA)_n + (1/PV)_n, \qquad where\ n\ indicates\ normalization$$

- ❑ If PV *indicates mutation chance*, then $1/PV$ gives how a peptide is *conserved* against mutation.

- ❑ We want the peptides to be *least mutated and highly surface exposed*. So,

$$(ASA)_n\ and\ (1/PV)_n\ must\ be\ high, and\ hence\ w\ parameter\ must\ be\ high$$

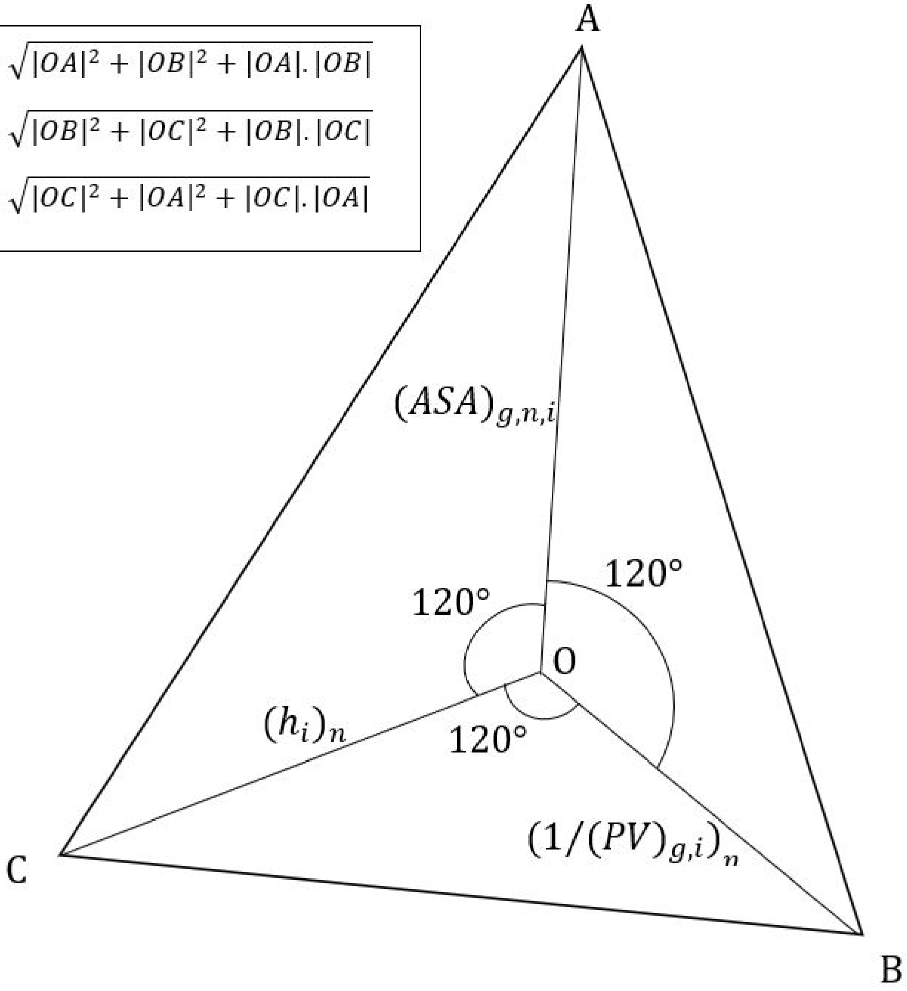
- ❑ Thus, using this *w parameter* value, we *rank* the 12-length peptides, consider top 50 – 100 ranks and group them into "*peptide zones*".

Step II: 2D Polygon Representation

$$|AB| = \sqrt{|OA|^2 + |OB|^2 + |OA|.|OB|}$$

$$|BC| = \sqrt{|OB|^2 + |OC|^2 + |OB|.|OC|}$$

$$|CA| = \sqrt{|OC|^2 + |OA|^2 + |OC|.|OA|}$$



□ For each zone, three properties are considered:

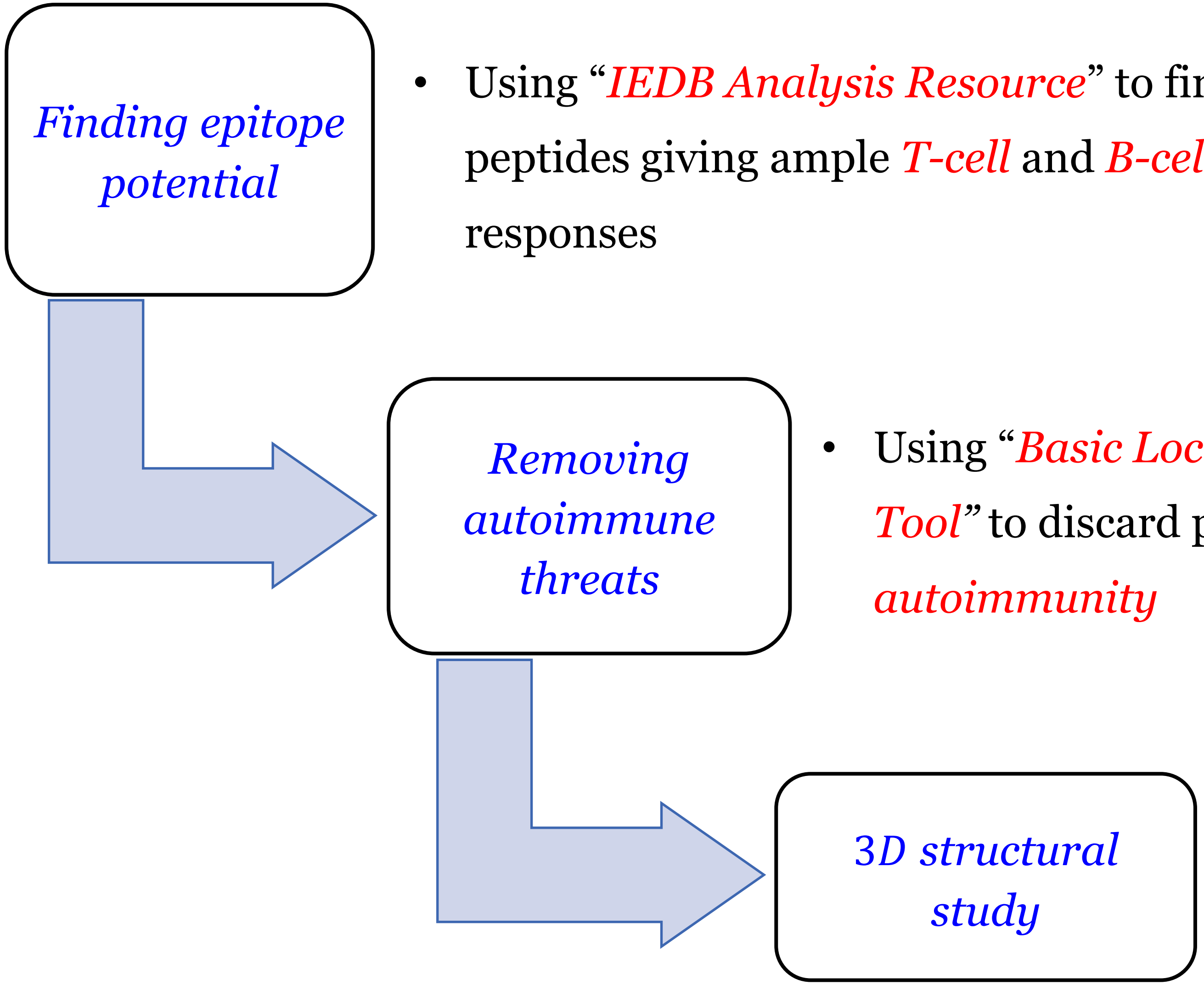
- Surface exposure (ASA)
- Conservativeness $(1/PV)$
- Area *across which the zone is spread* (n)

□ Arms OA, OB and OC represent these 3 values as shown.

□ Characterize the zone using the *area of the triangle* formed.

□ *The more the area, more suitable is the zone.*

Further steps



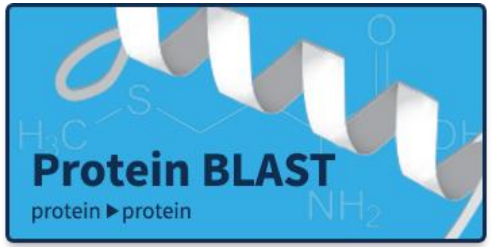
Finding epitope potential

- Using "*IEDB Analysis Resource*" to find peptides giving ample *T-cell* and *B-cell* responses

IMMUNE EPITOPE DATABASE AND ANALYSIS RESOURCE

Removing autoimmune threats

- Using "*Basic Local Alignment Search Tool*" to discard peptides causing *autoimmunity*



3D structural study

- Using platforms like "*Cn3D*", "*PyMOL*" to *visualize* the shortlisted candidates

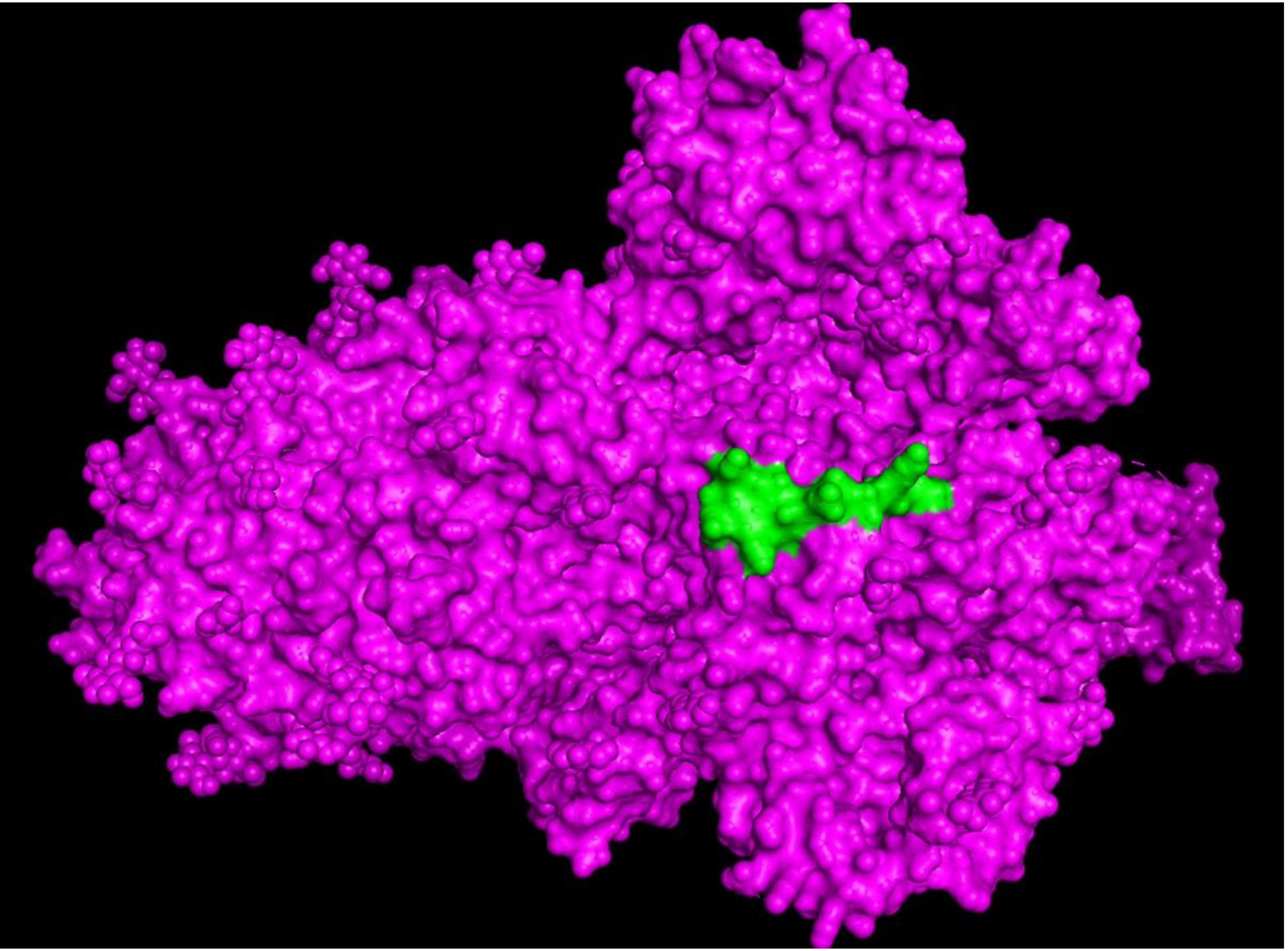
Final shortlist of vaccine targets

- ❑ Form a final shortlist of the desired vaccine candidates satisfying all these criteria

- ❑ Here, we have listed below the candidates found for *SARS-CoV-2* using its *surface glycoprotein.*

Sl. No.	Position in the spike protein sequence	Peptide region
1	527 – 541	PKKSTNLVKNKCVNF
2	696 – 710	TMSLGAENSVAYSNN
3	1132 – 1146	IVNNTVYDPLQPELD

The regions are *free from the mutations* in the past variants of *concern* and *interest* of SARS-CoV-2 as well as the currently surging "*Omicron*" variant.



Region 527 – 541 on a 3*D* model of surface protein of SARS-CoV-2

3*D* model created from: PyMOL

Future Avenues

- ❑ Comparing the efficacy of the design with the similar other vaccines under trial, like *EPV-CoV*19 (by EpiVax Inc.) and *CoVAC*-1

- ❑ The regions can be utilised by scientists working in wet labs to check how they are working against the target pathogen

Link for the application

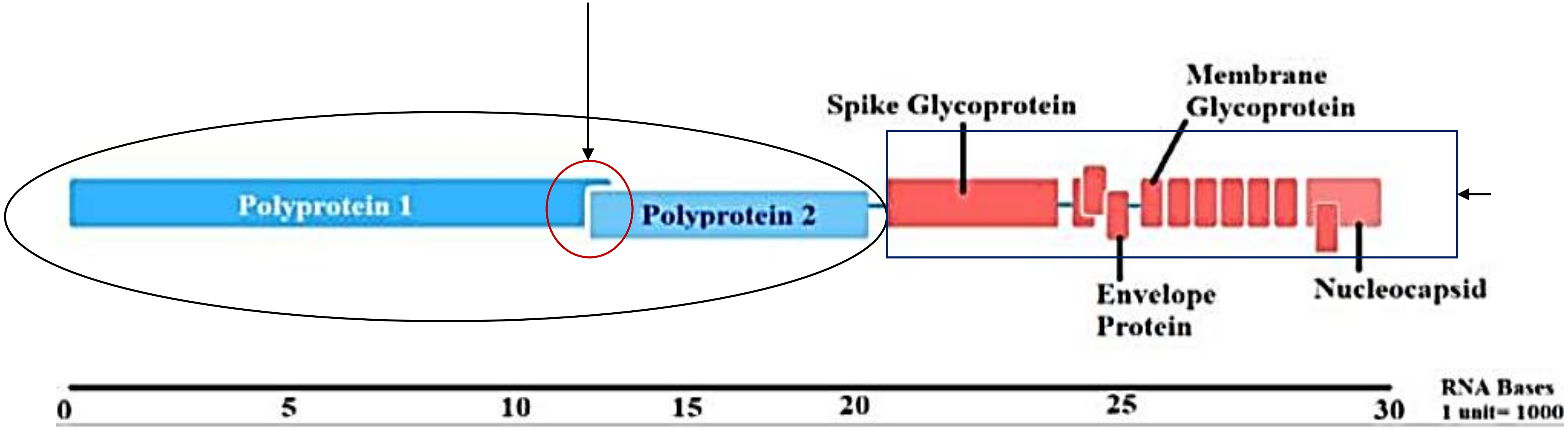
https://github.com/SubhamoyBiswas/Installation-Setup-for-Peptide-Vaccine-Analysis-Tool-PVAT

- ❑ Developed using *Python* programming (libraries used: *Tkinter*, *Matplotlib*, *NumPy*, *SciPy*)

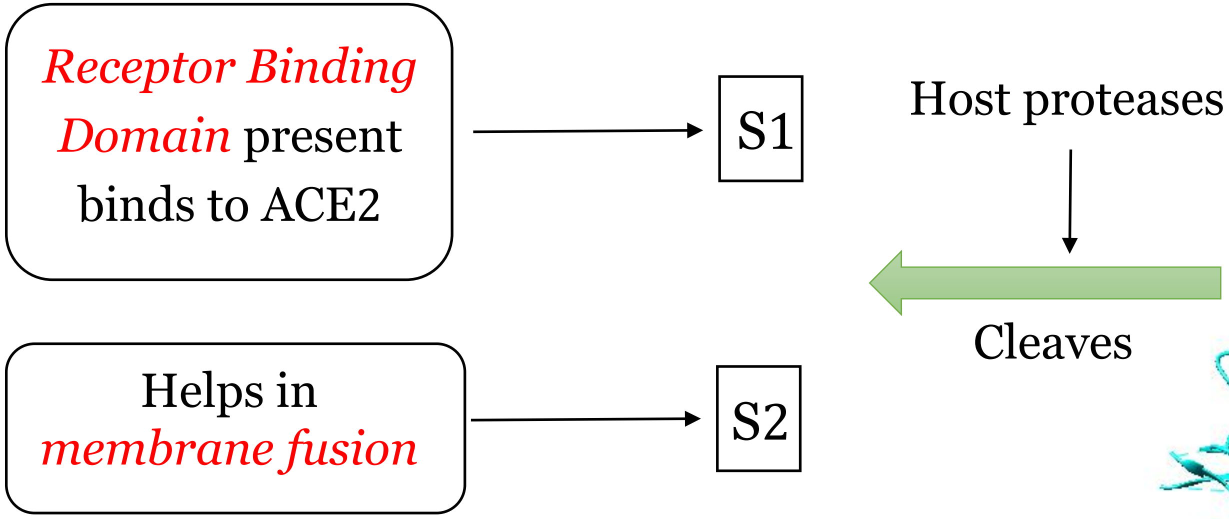
Viral Mutations

The bane of anti-viral vaccines is the high mutation rate in viral sequences that render any fixed target vaccines unsuitable after some mutations. Our protocol builds in a level of protection against such mutational weakness and sets a limit of how much mutational changes can be anticipated and protected against. We show this with special reference to the omicron variant of concern as per WHO.

The Biology of SARS-CoV-2



The Spike Protein



Receptor Binding Domain present binds to ACE2

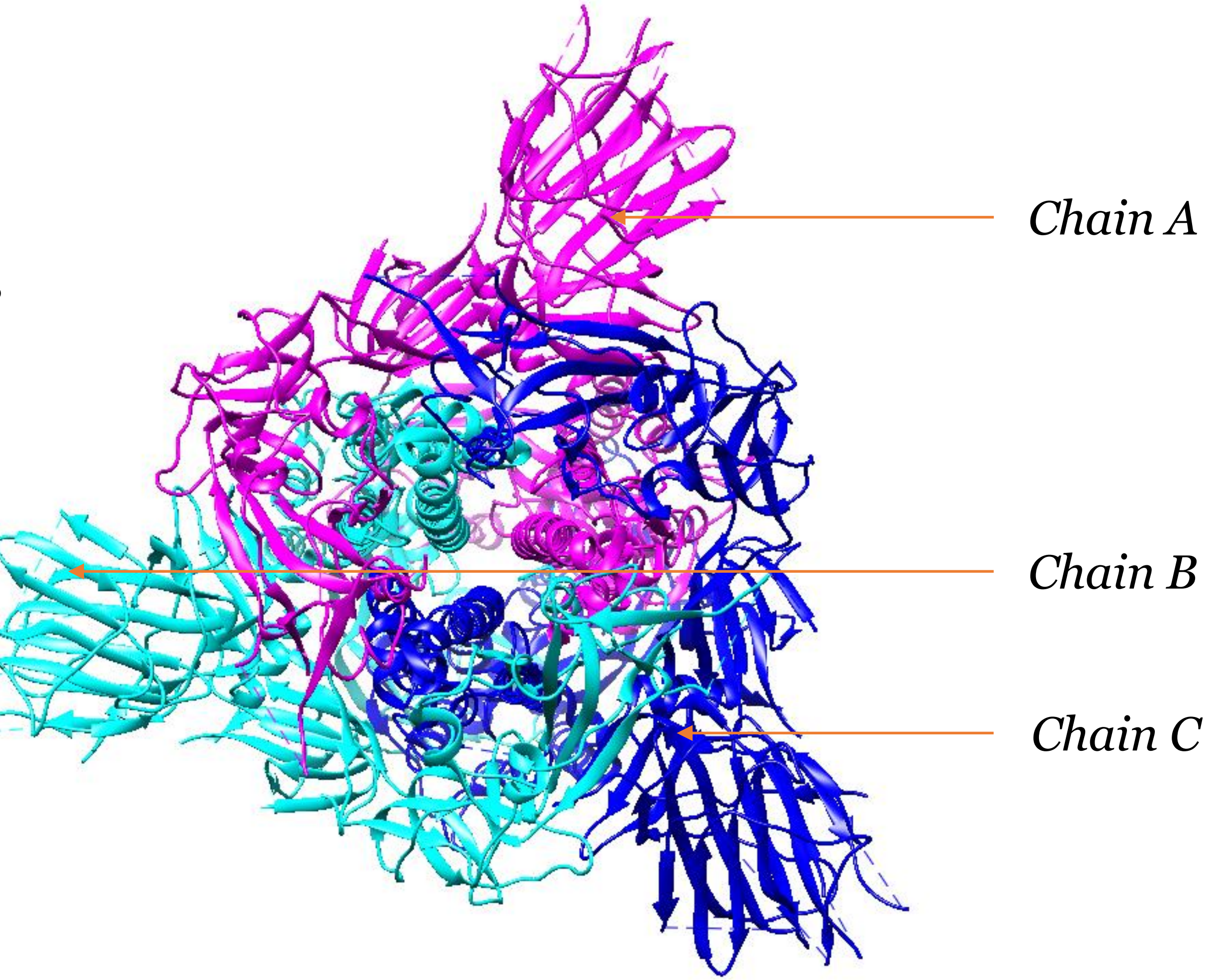
S1

Host proteases

Cleaves

Helps in *membrane fusion*

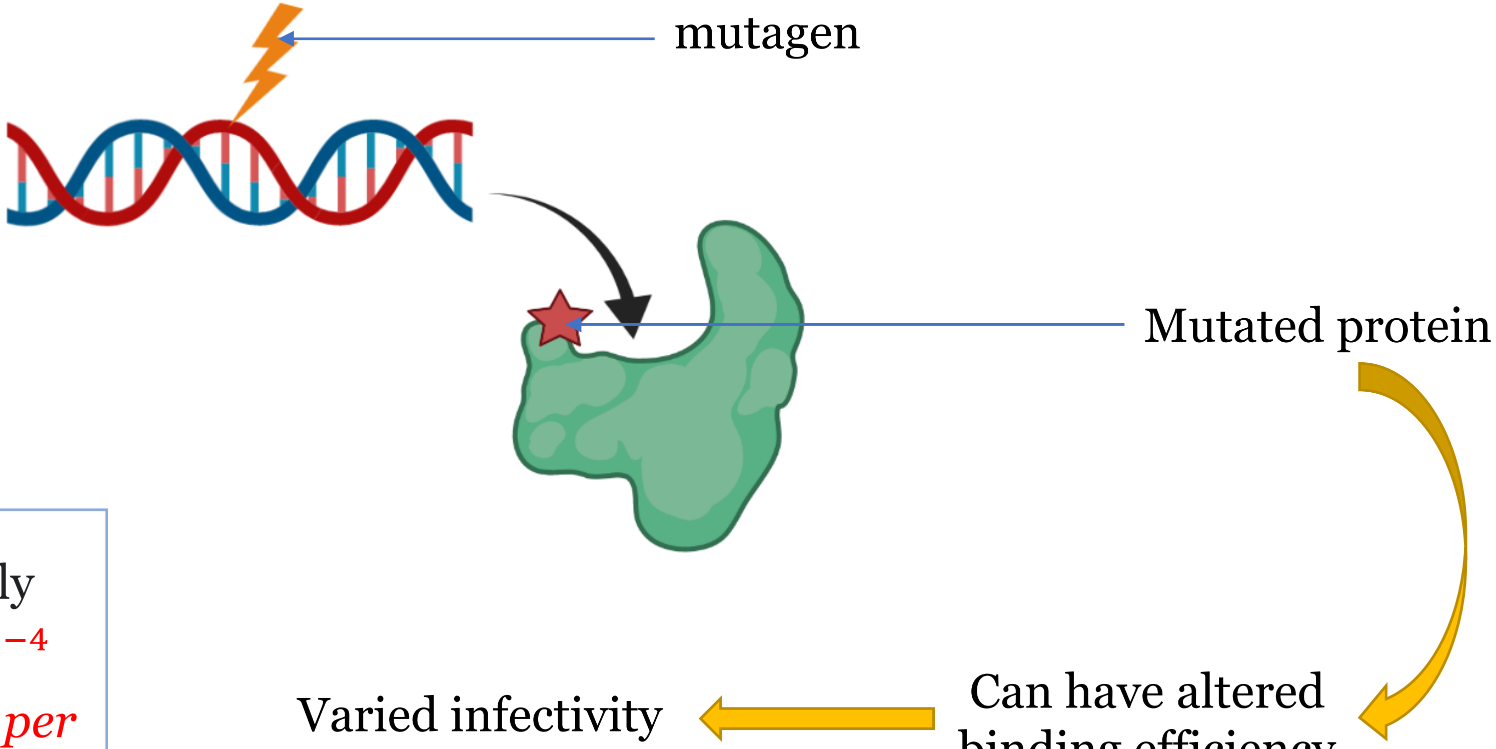
S2



Molecular Graphics performed in UCSF Chimera

Mutations

When a gene is damaged or modified in such a way that now the genetic message carried by that gene is altered, it is considered a mutation.



mutagen

Viral DNA or RNA

Mutated protein

Can have altered binding efficiency

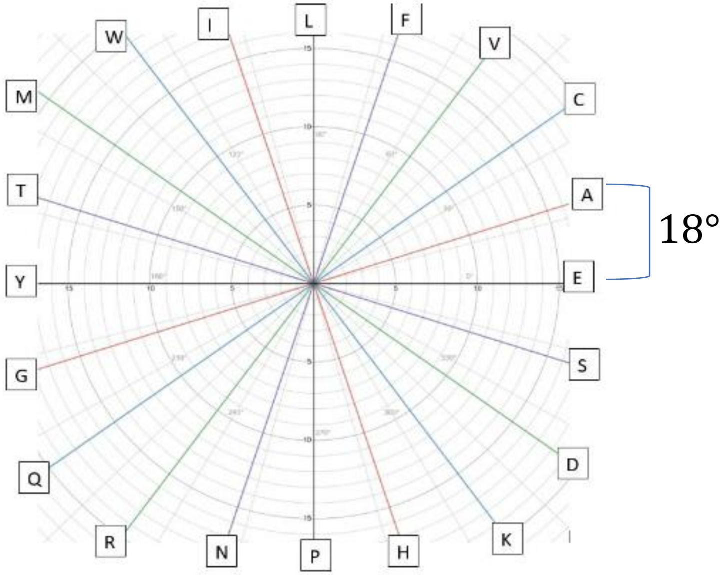
Varied infectivity

viral mutation rates roughly range between 10^{-8} and 10^{-4} substitutions *per nucleotide per replication*

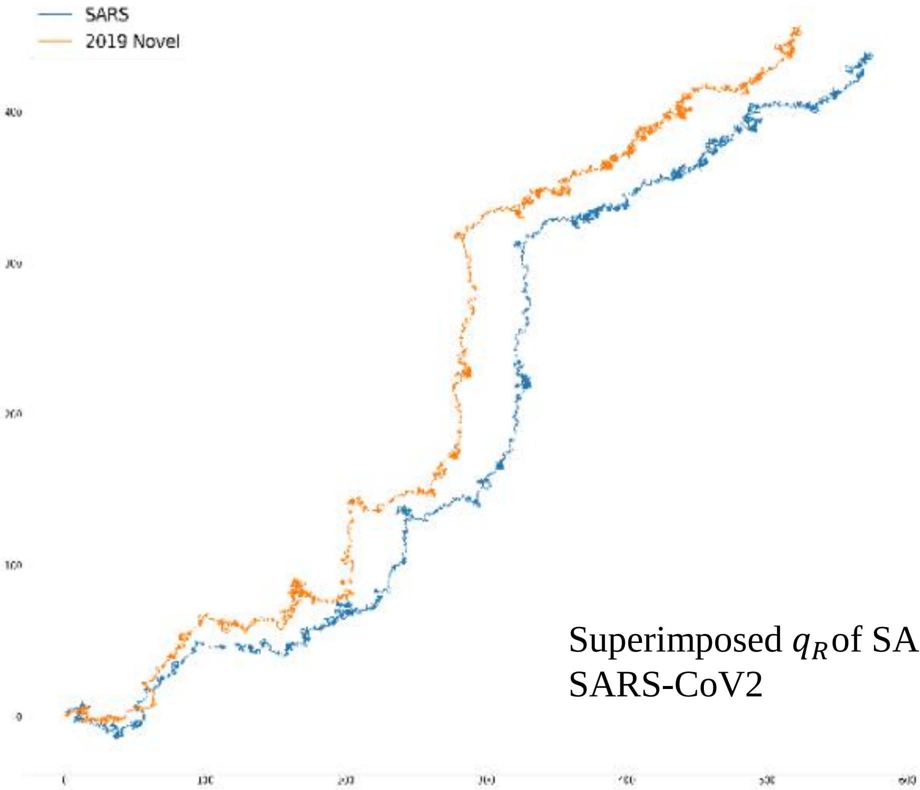
Figure created with help of BioRender (https://biorender.com/)

q_R *Method*

q_R is a technique to represent protein sequences in the $2D$ Cartesian plane based on their *hydrophobicity indices.*



Assigning the amino acids in the polar coordinate



Superimposed $q_R$of SARS and SARS-CoV2

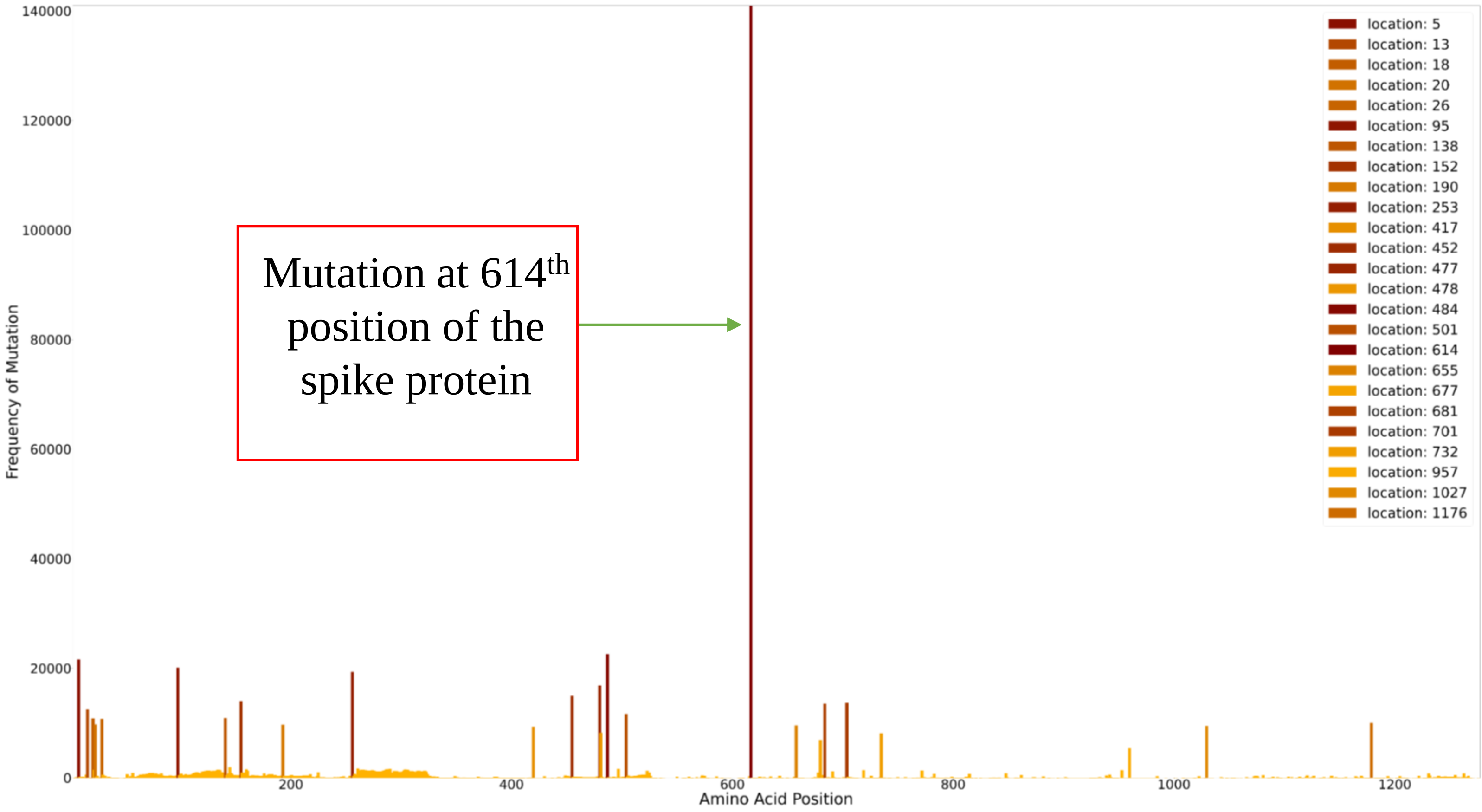
Mathematical descriptor of peptide sequences (weighted center of mass of the graph plot)

$$q_R = \sqrt{\mu_x^2 + \mu_y^2}$$

Visual representation of peptide sequences based on their intrinsic property of interaction with water (solvent)

Dey et al, MOL2NET, 2020

Mutational Hotspots



Mutation at 614th position of the spike protein

Mutational Hotspot Analysis of Spike Protein of SARS-CoV-2

Understanding the positions of amino acids inside the protein sequences where mutations have occurred frequently

↑

Try to understand why that particular mutation(s) is more accepted, i.e., try to find out the evolutionary significance

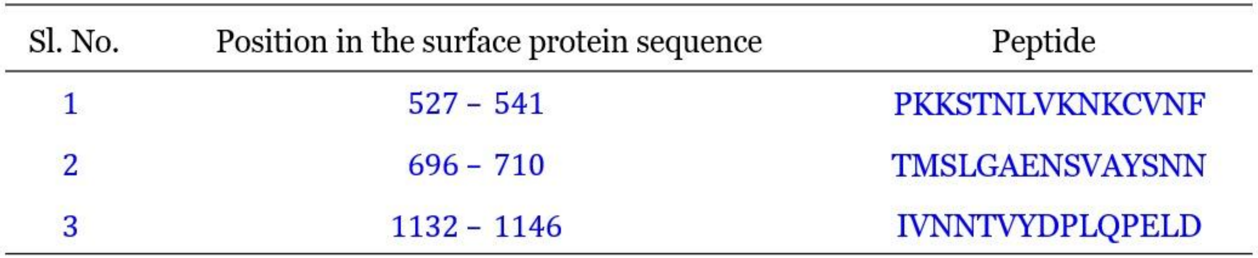
↑

Is that particular mutation more infectious?

Variants of Concern and potential peptide vaccine candidates

Sequence	Mutations in Spike Protein	
Delta Variant (B.1.617.2)	• $\boldsymbol{S}: T19R$ • $\boldsymbol{S}: E156 - (deletion)$ • $\boldsymbol{S}: F157 - (deletion)$ • $\boldsymbol{S}: R158G$ • $\boldsymbol{S}: L452R$	• $\boldsymbol{S}: T478K$ • $\boldsymbol{S}: D614G$ • $\boldsymbol{S}: P681R$ • $\boldsymbol{S}: D950N$
Beta Variant (B.1.351)	• $\boldsymbol{S}: D80A$ • $\boldsymbol{S}: D215G$ • $\boldsymbol{S}: L241 - (deletion)$ • $\boldsymbol{S}: L242 - (deletion)$ • $\boldsymbol{S}: A243 - (deletion)$	• $\boldsymbol{S}: K417N$ • $\boldsymbol{S}: E484K$ • $\boldsymbol{S}: N501Y$ • $\boldsymbol{S}: D614G$ • $\boldsymbol{S}: A701V$
Alpha Variant (B.1.1.7)	• $\boldsymbol{S}: H69 - (deletion)$ • $\boldsymbol{S}: V70 - (deletion)$ • $\boldsymbol{S}: Y144 - (deletion)$ • $\boldsymbol{S}: N501Y$ • $\boldsymbol{S}: A570D$	• $\boldsymbol{S}: D614G$ • $\boldsymbol{S}: P681H$ • $\boldsymbol{S}: T716I$ • $\boldsymbol{S}: S982A$ • $\boldsymbol{S}: D1118H$

Conserved peptides



Sl. No.	Position in the surface protein sequence	Peptide
1	527 – 541	PKKSTNLVKNKCVNF
2	696 – 710	TMSLGAENSVAYSNN
3	1132 – 1146	IVNNTVYDPLQPELD

No mutations found in the variants of concern were observed inside the potential peptide vaccine candidates obtained from our methods.

Data collected from https://covariants.org/

The Omicron Variant (B.1.1.529)

- **S:**A67V
- **S:**H69- (deletion)
- **S:**V70- (deletion)
- **S:**T95I
- **S:**G142- (deletion)
- **S:**V143- (deletion)
- **S:**Y144- (deletion)
- **S:**Y145D
- **S:**N211- (deletion)
- **S:**L212I
- **S:**G339D
- **S:**S371L
- **S:**S373P
- **S:**S375F
- **S:**K417N
- **S:**N440K
- **S:**G446S
- **S:**S477N
- **S:**T478K
- **S:**E484A
- **S:**Q493R
- **S:**G496S
- **S:**Q498R
- **S:**N501Y
- **S:**Y505H
- **S:**T547K
- **S:**D614G
- **S:**H655Y
- **S:**N679K
- **S:**P681H
- **S:**N764K
- **S:**D796Y
- **S:**N856K
- **S:**Q954H
- **S:**N969K
- **S:**L981F

This deletion was also observed in Alpha and Iota variants

An insertion, EPE was observed at position 214

Mutation helps in immuno-escape

Invitro studies show mutation at this site causes increased binding efficiency with ACE2 Receptor

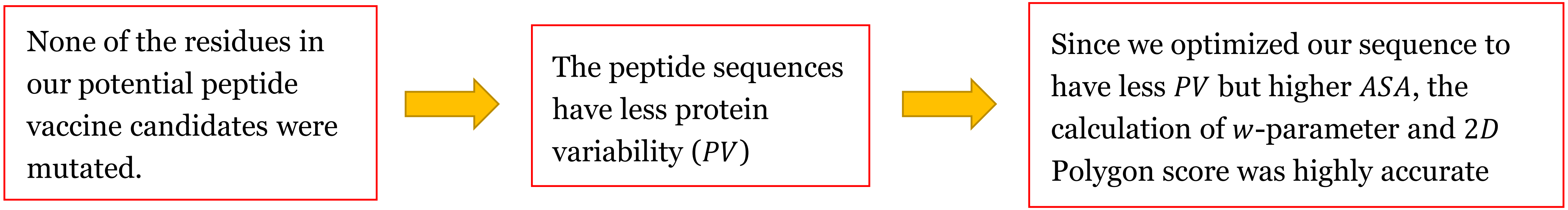
Mutation present in the S1-S2 Furin Cleavage Site

Might be a reason for high transmissibility

The mutations found in the spike protein of Omicron variants were not observed to be falling inside or in the vicinity of the peptide sequence of the potential peptide vaccine candidates obtained from our methods.

Data collected from https://covariants.org/

What does this signify…??



None of the residues in our potential peptide vaccine candidates were mutated.

The peptide sequences have less protein variability (PV)

Since we optimized our sequence to have less PV but higher ASA, the calculation of w-parameter and $2D$ Polygon score was highly accurate

- □ The protocol can be used to analyze sequences and obtain potential vaccine candidates against emerging pathogens with very high accuracy.

- □ The *optimal balancing* between Protein Variability and Amino Acid Solvent Accessibility helped us to obtain peptide regions highly conserved and accessible by solvents, hence, allowing antibodies to find and bind to these regions easily.

Conclusions

So, from this exercise we can draw the following conclusions:

❑ Our method allows a lot of mutational variability in the viral sequences. Using an algorithmic approach, we are able to find the common regions in the sequences that are *least variable*.

❑ The Omicron variety falls in the same class as the other variants and our prescription for peptide vaccine would be able to deal with it like the cases for the other variants.

References

- ❑ Biswas S, Manna S, Nandy A, Basak SC. (2021) “New Computational Approach for Peptide Vaccine Design Against SARS-COV-2”. *International Journal of Peptide Research and Therapeutics*, **27,** 2257–2273.

- ❑ Dey T, Chatterjee S, Manna S, Nandy A, Basak SC. (2021) “Identification and computational analysis of mutations in SARS-CoV-2”. *Computers in biology and medicine*, *129*, p.104166.

- ❑ Ghosh A and Nandy A. (2011) “Graphical representation and mathematical characterization of protein sequences and applications to viral proteins”. Advances in Protein Chemistry and Structural Biology, 83, pp.1-42.

- ❑ Raychaudhury C and Nandy A. (1999) “Indexing scheme and similarity measures for macromolecular sequences”. *Journal of chemical information and computer sciences*, *39*(2), pp.243-247.

- ❑ Biswas S, Dey T, Chatterjee S, Manna S, Nandy A, Das S, Nandy P, Basak SC. (2020) “*Novel Algorithms for In Silico Peptide Vaccine Design with Reference to Ebola Virus*”. IEEE International Conference on Computer, Electrical & Communication Engineering, doi: 10.1109/ICCECE48148.2020.9223075