Generating Immune-aware SARS-CoV-2 Spike Proteins for Universal Vaccine Design

Dominic Phillips [*1] **Hans-Christof Gasser** [*1] **Sebestyén Kamp** [*1] **Aleksander Pałkowski** [2] **Lukasz Rabalski** [3]
Diego A. Oyarzún [1 4 5 6] **Ajitha Rajan** [1 6] **Javier Antonio Alfaro** [2 6]

Abstract

Dozens of SARS-CoV-2 vaccines have been approved for public use, yet there remains a risk that the virus evolves to escape vaccine protection. This motivates the development of universal vaccines capable of protecting against current and potentially new strains of the virus. A key challenge is the lack of computational tools to design new viral proteins capable of vaccine escape, which could serve as good targets for the development of universal vaccines.

Here, we designed Variational Autoencoder (VAE) capable of generating SARS-CoV-2 spike proteins with variable immune visibility to the cell-mediated immune response. We compared our model with two simpler generative models; a random-mutator and an 11-gram language model. All three models can generate stable, structurally valid sequences, yet only the VAE model can generate low immunogenicity sequences that interpolate smoothly along the principal variance directions of known natural sequences. This model provides an effective computational tool for the generation of spike protein sequences useful for universal vaccine design. We provide its source code at https://github.com/hcgasser/SpikeVAE.

1. Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is a beta coronavirus first identified in December 2019 (Wang et al., 2020). The global pandemic it caused spurred vaccine development at a record-breaking rate (Petri, 2020). The first approval of a vaccine for widespread use in the UK came in January 2021 - just 11 months after the first viral sequence was released (Ledford et al., 2020). As of May 2022, there are a total of 30 vaccines approved for public use globally, utilising a wide range of technologies (Craven, 2022).

Despite this success in vaccine development, there is strong evidence that the Omicron variant can partially escape mainstream vaccines (Cao et al., 2022; Cele et al., 2022; Flemming, 2022). This motivates the development of *universal vaccines* that are broadly effective against both current and potential future strains (Nachbagauer & Krammer, 2017).

Much of the research into universal SARS-CoV-2 vaccines has been inspired by similar efforts for influenza - which has a similar evolutionary rate (Beans, 2022; Callaway, 2021). In particular, the main focus of efforts has been on identifying conserved *viral epitopes*: preserved regions of viral antigens recognised by the immune system (Sanchez-Trincado et al., 2017; El-Manzalawy & Honavar, 2010).

There are several computational methods for automatically identifying conserved epitopes (Qiu et al., 2019; Malone et al., 2020). They are, however, limited in that they cannot extrapolate beyond existing known sequences, which restricts their use for developing universal vaccines. In addition, a reliance on conserved regions misses the opportunity to leverage the commonalities in a comprehensive map of epitopes of the virus. A promising alternative approach could be generative models (Strokach & Kim, 2022), which have found success for *in silico* generation of proteins with particular biological function, such as improved ligand binding or antibiotic resistance properties (Madani et al., 2020; Greener et al., 2018; Chhibbar & Joshi, 2019).

In this paper, we demonstrate how VAE, a popular type of generative model, can be employed to generate novel protein sequences. We designed and evaluated an immune-aware

[*] Equal contribution [1] School of Informatics, University of Edinburgh, Edinburgh, United Kingdom [2] International Centre for Cancer Vaccine Science, University of Gdańsk, Gdańsk, Poland [3] Laboratory of Recombinant Vaccines, Intercollegiate Faculty of Biotechnology of University of Gdansk and Medical University of Gdańsk, Gdańsk, Poland [4] School of Biological Sciences, University of Edinburgh [5] The Alan Turing Institute, London [6] Joint advisors. Correspondence to: Dominic Phillips <Dominic.Phillips@ed.ac.uk>, Hans-Christof Gasser <h.gasser@sms.ed.ac.uk>, Sebestyén Kamp <s.kamp@sms.ed.ac.uk>.

Proceedings of the 1^{st} Workshop on Healthcare AI and COVID-19, ICML 2022.

VAE that selectively generates synthetic SARS-CoV-2 spike proteins with variable levels of immune visibility. We evaluate the validity and stability of the generated sequences and their tertiary structures with respect to natural spike proteins. Generated sequences with low immune visibility might pose the highest risk of vaccine escape, and therefore should be considered when designing forward-looking vaccines.

The structure of the paper is as follows. Section 2 covers the necessary background in immunology, SARS-CoV-2 biology and generative models for protein design. Section 3 details our methods. Results are presented and discussed in Section 4. Conclusions and further work can be found in Section 5.

2. Background

2.1. Vaccine-induced immune response

Vaccines enable fast pathogen detection by the adaptive immune system. Detected pathogen parts (antigens) leading to an immune response are called immunogenic. Antibodies produced by B-Cells are part of the humoral response, while the cellular response is mainly T-Cell based. Cytotoxic T-lymphocytes (CTL) deal with viruses inside cells. In nucleated cells the MHC class I (MHC-I) mechanism facilitates the detection of infected cells (Rock et al., 2016). The MHC-I protein presents fragments of all proteins present in a cell (called peptides, typically 8–10 amino acids (AAs) long) on the cell surface (Rock et al., 2016). These are produced by proteasomes and then transferred into the endoplasmic reticulum (ER) where they encounter MHC-I proteins. Some peptides will bind to the MHC-I's binding grooves. These peptide-MHC protein complexes (pMHCs) then migrate to the cell surface, where they are presented to the extracellular environment. They can act as epitopes to the T-cell receptor (TCR) of CTL. If these CTLs are activated and the epitope is considered by them to be non-self, then the cell will be destroyed (Rock et al., 2016). Models exist that predict which peptides of a protein will get presented (see Section 2.4).

2.2. Principles of viral mutation

Considering potential mutations in vaccine design is always necessary, particularly for highly mutagenic viruses. Mutations can be caused by replication errors, damage to the nucleic acids, host proteins changing the virus's genetic material, but also by so-called diversity-generating retroelements (DGRs) (Benler et al., 2018). Recombinations of coinfecting viruses exchanging genetic information can also lead to major genetic changes (Fleischmann, 1996). Mutation rates vary across viruses. In general ribonucleic acid (RNA) viruses (like SARS-CoV-2 and HIV) display higher mutation rates than DNA based viruses (like HPV) (Sanjuán & Domingo-Calap, 2016). This is mainly due to the RNA virus's lack of proofreading functionality (Fleischmann, 1996). Single-stranded viruses (like SARS-CoV-2) are also renown to have higher mutation rates than double-stranded ones (Sanjuán & Domingo-Calap, 2016).

Positive-stranded RNA coronaviruses display a comparatively high mutation rate (Piepoli et al., 2020). An estimate for SARS-CoV-2's overall mutation rate is in the order of 10^{-6} $\text{nt}^{-1}\text{cycle}^{-1}$ (per nucleotide per infection cycle) (Borges et al., 2021), whilst the spike protein estimate is even an order of magnitude higher at 10^{-5} $\text{nt}^{-1}\text{cycle}^{-1}$ (Borges et al., 2021). Due to its exposure to the external environment (making it accessible to antibodies) the spike protein was the focus of vaccine development and is also the focus of this work. Typically, there are 50 to 100 of them on a single virus (Piepoli et al., 2020).

2.3. Generative models in Protein Design

Autoregressive (AR) models were early generative models for AA sequences. For example, Shin et al. have used a LSTM recurrent neural network (RNN) to generate single domain antibodies. In recent years, RNN models are increasingly being replaced by transformers (Vaswani et al., 2017) - also in protein design. Wu et al. used them to generate signal peptides (control protein secretion in cells). Also, recent advances in image generative models have been transferred to protein generation - for example Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) and VAEs. Repecka et al. have developed ProteinGAN which they use to generate new malate dehydrogenase (MDH) variations and were able to experimentally verify that 24% of these displayed MDH's catalytic activity *in vitro*.

We use the second popular generative model - VAEs. These are based on the idea of Autoencoders (AEs). Both consist of an encoder and a decoder. For example, a AA sequence is fed into the encoder outputting a lower-dimensional sequence representation - the latent variable. This is then fed into the decoder to reconstruct the original input. The loss represents how well the original input was reconstructed. Unfortunately AEs tend not to generate a diverse set of examples outside the dataset. Therefore Kingma & Welling introduced VAEs. Here the encoder's output is used to produce an expected value and variance for the latent variable. A normally distributed random variable with these specifications is then sampled and fed into the decoder. This enables the model to attribute useful values to a continuum of latent variables. The distribution of the latent variables is controlled by a loss term that includes the KL-divergence between the actual latent variable distribution and an idealized standard normal (Kingma & Welling, 2014) in addition to the reconstruction loss.

Early work using VAE for protein design was done by Sinai et al.. They use encoder and decoder networks with three dense layers of 250 units and dropout. They used five latent variables for prediction and two for visualization. They find that the lower-dimensional representation had only a slight weakening effect on predictive power. In comparison, on almost 70,000 luciferase-like oxidoreductases Hawkins-Hooker et al. trained a fully connected encoder and decoder on aligned sequences (MSA VAE), and on unaligned sequences they trained a convolutional neural network encoder in combination with an auto-regressive decoder (AR VAE). Their MSA VAE model was better able to capture long-distance dependencies reflecting the 3D structure of the folded protein. In total, 6 of their 12 generated AR VAE variants and 9 of their 11 generated MSA VAE variants demonstrated measurable luminescence. They used conditional VAEs (Sohn et al., 2015) to control the variant solubility - we use this to control for level of immune visibility (see Section 2.4).

2.4. Antigenicity prediction

Not all viral protein parts are presented on the cell's surface (Section 2.1). With antigenicity we mean a peptides propensity to be displayed on the cell's surface. The higher, the easier detection should be for the immune system. Several models predict presentation - for example NetMHCpan (Reynisson et al., 2020), MHCflurry (O'Donnell et al., 2020) and ImmunoBERT (Gasser et al., 2021).

3. Methods

3.1. Data

We acquired a large set of SARS-CoV-2 spike proteins from the Global Initiative on Sharing All Influenza Data (GISAID) database (Shu & McCauley, 2017). Its entries are sourced from high-quality, deeply sequenced genomics data from laboratories across the globe. As of January 2022, the database consists of 13,817,026 individual SARS-CoV-2 spike protein sequences.

Notably, some database sequences are only fragments. Others have undetermined amino-acids at various residue positions. We therefore filtered the database by discarding any sequence of length less than 1200 (the Wuhan reference variant is 1273 AA long (Huang et al., 2020)) or any sequence containing an undetermined residue. After this, 4,438,573 sequences remained - 167,133 distinct spike protein sequences. We selected the top 65,000 most common ones [1] and sequence aligned them using MUSCLE v.3.8 (Edgar, 2004). To make this tractable, we split and align the sequences in batches of 5000 and then hierarchically merged the aligned lists using MUSCLE's merge function. This produced aligned sequences of length 1,299 AAs.

For variant annotation we retrieved variant consensus sequences from Expasy (Duvaud et al., 2021) (Wuhan, Alpha, Beta, Gamma, Delta, Epsilon, Omicron BA.1). Sequences were then annotated with a variant based on the shortest edit distance using BioPython (Cock et al., 2009) pairwise alignment.

3.2. Embedding visualisations

In our report, we include visualisations of natural and synthetic sequence datasets. The visuals are constructed in three stages, described below.

Stage 1: Sequence encoding
Sequences, being text data, are challenging to directly visualise. Rather, it is convenient to first encode each sequence as a high-dimensional numerical vector. A range of methods has been developed for this purpose (Jing et al., 2020). At the broadest level, they are divided into position-independent encodings (which assign the same numerical encoding to each AA regardless of its location) and position-dependent encodings (which encode residues based on their contextual environment, and so may assign a different encoding to the same AA appearing at different locations).

Due to their simplicity of implementation, we settled on using position-independent encodings. We used BLOSUM62 encoding(Henikoff & Henikoff, 1992), an encoding derived from AA evolutionary substitution data, since it has the best performance in protein fold recognition tasks (Jing et al., 2020); see also Appendix 7.1.

Stage 2: Sequence masking (optional)
In universal vaccine design, it is germane to map out conserved epitope space. This space is different for B cells than it is for T cells. For B cells, the epitope space is related to those regions of the protein that are exposed on the surface (Sanchez-Trincado et al., 2017). To visualise B-cell epitope space, we apply an 'epitope mask' to each sequence. The epitope mask effectively weights the importance of residues by how proximal they are to the surface. We obtain and apply the mask as follows: (1) Apply the DSSP program (Touw et al., 2015) (Kabsch & Sander, 1983) to the PDB of the original Wuhan variant (Cai et al., 2021) to obtain a normalised (between 0 and 1) residue by residue surface solvent accessibility (SAV) score. (2) Pairwise align each sequence with the FASTA sequence of the original Wuhan variant using BioPython (Cock et al., 2009) and use this to compute an aligned SAV for each sequence[2]. (3) Apply the

[1] In practice, this means that each of these sequences had been detected in at least two individuals. The most common sequence had been detected in over half a million individuals.

[2] To make this alignment possible, we make the simplifying assumption that AA mutations did not modify solvent accessibility and that deletion mutations do not affect solvent accessibilities at other residue locations. We also assume that the solvent accessi-

mask by multiplying a sequence's encoding vector by the solvent accessibility value of each residue. The mask has the effect of biasing the visualisation towards displaying only the variation in those residues that are proximal to the protein surface.

Stage 3: Sequence embedding
There are many algorithms for embedding high-dimensional data to lower dimensions for ease of visualisation (Van Der Maaten et al., 2009). For simplicity we use just two methods: PCA (Principal Component Analysis), a linear dimensionality reduction algorithm famed for its simplicity(Jolliffe & Cadima, 2016), and t-SNE (t-distributed Stochastic Neighbor Embedding), a non-linear dimensionality reduction algorithm that has previously been extensively used in protein visualisations (Maaten & Hinton, 2008). For t-SNE, we pre-process by dimensionality-reducing down to 50 dimensions with PCA and then further reduce to two dimensions with t-SNE using a cosine metric, perplexity 30, a learning rate of 200, and running for 1000 iteration steps. We use these same parameters for all t-SNE visualisations.

3.3. Generative models

We investigate three classes of generative models. In order of increasing complexity these are: (1) a random-mutator model, (2) an N-gram language model, (3) a VAE model with antigenicity constraints. Models (1) and (2) are used as baselines when evaluating the sequences generated by the more complex VAE model. Each model generates sequences of the same length (1299aa).

Random-mutator model
This uses knowledge of positional AA variation in a dataset to generate novel sequences that differ from a natural sequence by several, statistically independent single-point mutations. Specifically, given a dataset of protein sequences, the residue by residue AA probability distributions are calculated and a 'mode sequence' is constructed by taking the most common AA at each residue. The number of positional differences between each sequence and the mode sequence are then calculated. For the SARS-CoV-2 dataset, we find that the number of mutational differences is approximately exponentially distributed with a mean of 8.81. This leads naturally to the following generative model:

1. Randomly select a natural sequence from the dataset.
2. Sample exponential variable N (mean 8.81, rounded).
3. Resample a random residue (from its AA distribution).
4. Repeat step 3 until exactly N positions changed.

N-gram language model
An N-gram language model is trained on the dataset defined in Section 3.1. New sequences are generated based on an $N-1$ residue initialising sequence (we use the first $N-1$ AA residues of the dataset's modal sequence) followed by the repeated sampling of the next AA in the sequence form the language model's probability distribution conditioned on the preceding $N-1$ residue tokens.

We used the NLTK library to experiment with N-gram models for $N = 3, 5, 7, 9, 11, 13, 15, 17$. The 11-gram model produced sequences with a positional entropy distribution most similar to the natural sequences[3].

VAE model with antigenicity conditioning
Our VAE model is based on fixed-length multiple sequence alignment (MSA) (1,299 positions) data, justifying the usage of a fully-connected neural network architecture similar to Hawkins-Hooker et al.. We use one-hot encoding. In addition to the 27,279 variables (1,299 positions times 20 amino acids + missing token) representing the sequence the encoder receives an additional 3 variables encoding the sequence antigenicity (low, medium, high).

A caveat of VAEs is that they can suffer from KL vanishing - the loss function's KL-divergence term gets optimized to close to zero, while the regeneration loss stays high (Fu et al., 2019). In extrema, the generator only outputs a single example - similar to GANs' mode collapse. We tried several ways to counter this - for example slowly increasing the weight of the KL divergence term in the loss function, cyclical annealing (Fu et al., 2019) and the ControlVAE (Shao et al., 2020) approach. The last one gave us the best control over the training process and how to manage the trade-off between reconstruction loss and generating normally distributed latent representations. It adapts the well-known PID control process (J & H, 2006) to control the KL divergence's weight in the loss to keep it stable at a predefined level.

Antigenicity calculation: We used NetMHCpan-4.1 to assess a spike protein's antigenicity. It requires a peptide as well as a MHC-I protein as input. We used its predicted eluted ligand (EL) rank[4] to assess antigenicity (presentation). Their online tool attributes weak antigenicity to the top 2% of scores and strong antigenicity to the top 0.5% (Nielsen, 2020).

We use the following algorithm to calculate a proxy (we call antigenicity score (AS)) for how many peptides within a sequence will be presented: For a sliding 9-AA window over the sequence we calculate the NetMHCpan EL rank for each of 12 common MHC alleles [5]. If this EL rank is below 2.0 (NetMHCpan's standard value for weak-binding

bility of an insertion mutation can be estimated as the mean of the solvent accessibility of its immediately adjacent residues.

[3] Section 4.3 has details on the positional entropy method.

[4] The rank of the predicted score in comparison to random naturally occurring peptides.

[5] HLA-A01:01, HLA-A02:01, HLA-A03:01, HLA-A24:02, HLA-A26:01, HLA-B07:02, HLA-B08:01, HLA-B27:05, HLA-B39:01, HLA-B40:01, HLA-B58:01, HLA-B15:01

peptides), then we count it as a hit. Our AS is the average (over the 12 HLA alleles) count of these hits across the spike protein. So, if there are 49 hits across the protein in 4 HLA alleles and 50 in another 4 alleles and 51 in the remaining ones, the AS of the sequence would be 50 (their average). A sequence is categorised as low, medium, or high antigenic, dependent on its rank in the ASs of all sequences in our database. Sequences with an AS within the first quartile (AS ≤ 49.833) are considered low antigenic, whilst sequences with scores in the last quartile (AS ≥ 50.333) are considered high antigenic and all others are considered medium antigenic. This categorization is then fed into the encoder and decoder.

There are 5 blocks in the encoder. The first encoder block reduces the 27,282-dimensional representation (27,279 for sequence, 3 for antigenicity) to a 512-dimensional one. Each block we used has the same structure. It begins with a linear layer followed by batch normalization and a leaky ReLU (negative slope of 0.1). All but the last block end with a dropout layer. Each block after the first one halves the output dimension. So the last block outputs 32 dimensions, which are then mapped via two linear layers to 30-dim values for the expected latent variable and log variance.

Using those specifications, normally distributed latent variables are sampled and fed into the generator. It also constitutes of 5 blocks with the same structure as the encoder blocks. Only here they perform up-sampling. The first one takes gets the 30-dimensional latent variable and the 3-dimensional antigenicity variable. The first block's output has 64 dimensions. Each block except for the last (which outputs 27,279 dimensions) doubles the dimensions. Details on the training and hyperparameter selection procedures of our model are provided in the Appendix.

Sampling: We generated 30-dimensional multivariate Gaussian distributed random variables (covariance matrix estimated from training sequences' latent vectors) with appended one-hot encoded antigenicity to input into the generator. This delivers a probability distribution over the AAs for each position. The sampled sequence is the maximum likelihood estimate for this distribution.

3.4. Evaluation

Generative models for protein sequences, unlike those for natural language, are difficult to evaluate through visual inspection alone. Therefore, we have created an evaluation pipeline to check that our approach is valid.

Firstly, we verify that the distribution of generated sequences is realistic, i.e. that they closely match the distribution of a representative sample of natural sequences. We do this through PCA embedding visualisations and positional entropy comparisons of generated sequences with natural sequences[6].

Secondly, we evaluate whether individual generated sequences are valid, energetically stable, and have tertiary structures similar to natural spike proteins. We do this with three increasingly rigorous checks, at each stage rejecting sequences that fail a check:

1. Conserved regions check From the sequence aligned dataset we identify 77 residue positions that are conserved amongst all natural sequences. Mutations in these regions are therefore highly unlikely to occur and we reject any generated sequence having any of these 'forbidden' mutations.

2. Sequence stability check We align the remaining generated sequences with each natural sequence in turn, and identify the natural sequence that has the highest alignment score. We then compute the point mutational differences between these two sequences. There are many bioinformatics tools, such as MAESTRO (Laimer et al., 2015), Rosetta, and DDGun (Pancotti et al., 2022), that use point mutational differences to predict a thermodynamic stability difference between two protein variants. We opted to use DDGun (Montanucci et al., 2019), as it is the only tool that can operate on just sequence information and account for multiple point mutations (Sanavia et al., 2020).

We supply DDGun with the FASTA file of the natural sequence as well as the point mutation list. The algorithm then estimates the change in free energy of unfolding, also known as $\Delta\Delta G$ (Fang, 2012), upon independently making each mutation. A positive $\Delta\Delta G$ indicates a destabilizing effect, a value close to zero value indicates a neutral mutation, whereas a negative value indicates a stabilizing effect. We make the simplifying assumption that the thermodynamic effects of the point mutations are independent, and estimate the combined $\Delta\Delta G$ value to be the linear sum of the independent contributions. Applying this procedure to each sequence enables us to rank generated sequences in order of decreasing predicted stability (i.e. from most negative $\Delta\Delta G$ to most positive).

3. Verifying tertiary structure We then use AlphaFold2 to predict the tertiary structure of the top 10 most stable sequences from each type of generative model. AlphaFold2 is a state-of-the-art, high accuracy protein folding algorithm (Kwon et al., 2021). From sequence input, the algorithm

[6] The normalised positional entropy S_n at a residue position n is defined as

$$S_n = \frac{\sum_{k=1}^{N} \frac{c_k(n)}{C} \ln \frac{c_k(n)}{C}}{\ln \frac{1}{N}},$$

where N is the number of possible tokens at each residue position (here $N = 21$; 20 possible naturally occurring AAs and one alignment token, "-"), $c_k(n)$ is the number of times the k^{th} token appeared at the n^{th} residue position in the dataset and $C = \sum_{k=1}^{N} c_k(n)$ is the total number of sequences in the dataset.

outputs plausible protein tertiary structures, ranked in order of decreasing likelihood. We take the highest ranking structure and use PyMol (Yuan et al., 2017) to compute the root-mean-square deviation (RMSD) between the coordinate positions of this structure and two different reference SARS-CoV-2 structures: (1) an Alpha-Folded structure of the reference Wuhan sequence (Duvaud et al., 2021), and (2) an experimental SARS-CoV-2 PDB structure, 7NIU (Cai et al., 2021). This particular structure was chosen since its defining sequence has the highest similarity with the Wuhan reference amongst all PDB structures.

To verify whether these RMSD values are reasonable, we compare them to the expected range of RMSDs amongst natural sequences: we compute the maximum pairwise RMSD value between the AlphaFolded structures of the consensus sequences (found to be 0.94Å, Appendix Table 5). Additionally, we compute the maximum RMSD value between these consensus structures and the experimental 7N1U structure (found to be 2.17Å, Appendix table 6). If the AlphaFolded structure of a generated sequence has its RMSD values with the SARS-CoV-2 reference structures to be smaller than the above defined values (0.94Å and 2.17Å respectively), then we consider it to be a plausible, stable spike protein.

4. Results

4.1. SARS-CoV-2 dataset visualisations

In Figure 1 we present t-SNE visualisations of the population of SARS-CoV-2 spike proteins, with and without epitope masking (Figure 1 (a) and (b) respectively).

Figure 1 (a) summarises several well-documented features of SARS-CoV-2 viral diversity, namely (i) there are seven main variants, but several of these, most notably delta, have dozens of subvariants, (ii) each subvariant is dominated by one sequence that is significantly more widespread, (iii) for each subvariant, there may be several hundred, or even thousands, of closely related ‘satellite sequences’ that are circulating at much lower levels in the general population.

Also note that when we apply an epitope mask (Figure 1 (b)) , these satellite sequences spread out more, forming prominent comet-like tails. This may reflect the presence of selective pressure for spike protein mutations to occur more frequently in regions visible to B-cell immunity, i.e. surface epitopes. Alternatively, it may be a visual artifact of the simplifying assumptions made when determining a sequence’s solvent accessibility vector (see footnote 5, Subsection 3.2)

4.2. Generated sequences

We generated 8,880 random sequence samples using the random-mutator and language model. For the VAE model, we generated 50,000 samples of each of low, medium, and high antigenicity sequences by sampling from the latent space as per the method in Section 3.3. In Table 1 below we report the fraction of these sampled sequences that are distinct (i.e. non-degenerate) and novel (i.e. not appearing even once in the training data).

For the VAE-generated sequences, we additionally ran NetMHCpan on each generated sequence to verify that antigenicity distributions were indeed skewed towards the desired category. This confirmed that the conditioned VAE was therefore working as expected (see Figure 3 (a)).

Overall, amongst the VAE-generated sequences, we identified 78 novel 9-mers that did not appear in the training data sequences (see Appendix 7.3 for the list of 9-mers).

4.3. Evaluating and comparing the generative models

We compare the t-SNE and PCA embeddings of generated sequences with natural sequences (Figure 3). We also follow standard practice (see e.g. (Repecka et al., 2021)) and compare the normalised positional entropy (Figure 2).

Although the VAE model performs the worst on entropy similarity (the RMSD difference between the natural entropy distribution and the VAE entropy distribution is nearly double the respective RMSD values for the language model or random mutator model, see Figure 2), an inspection of the embedding visualisations clearly shows that it outperforms the simpler generative models. This is perhaps most apparent in the PCA plot in Figure 3 (b) where the VAE-generated sequences nicely interpolate along the two principal axes delineated by the natural sequences, whereas both the random-mutator and language model tend to ‘fill in the square’ and produce a large number of off-distribution sequences.

Table 1. Count statistics of sequences generated from various generative models.

Model	Samples	Distinct Sequences	Novel Sequences	Novel (%)
VAE (low)	50,000	932	875	94
VAE (medium)	50,000	1,308	1,110	85
VAE (high)	50,000	1,298	1,082	83
11-gram	8,880	8,686	8,446	97
Random Mutator	8,800	4,225	1,830	43

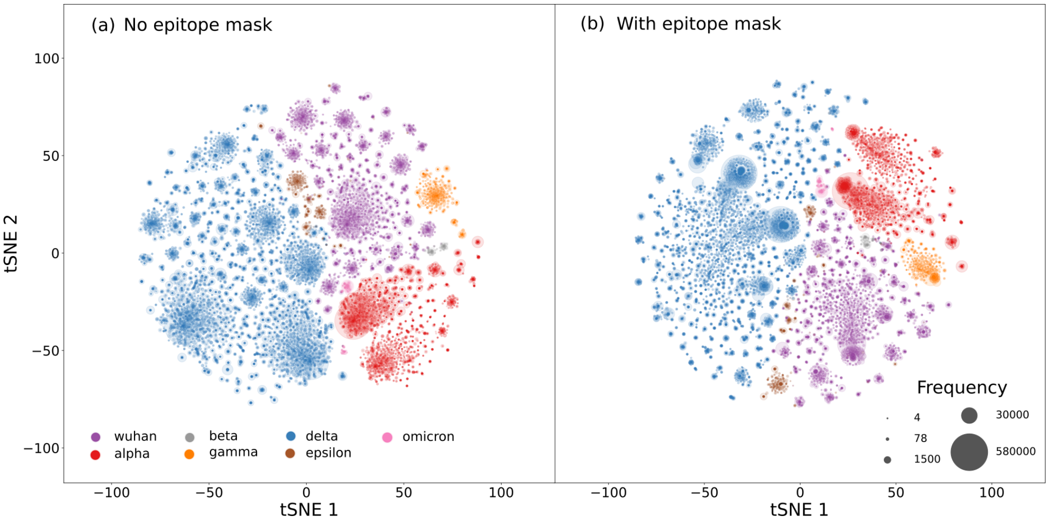

Figure 1. t-SNE visualisation of the 25,000 most common SARS-CoV-2 spike protein sequences. (a) shows the t-SNE embedding if no mask is applied to the sequences, and (b) shows the t-SNE embedding if an epitope mask is applied.

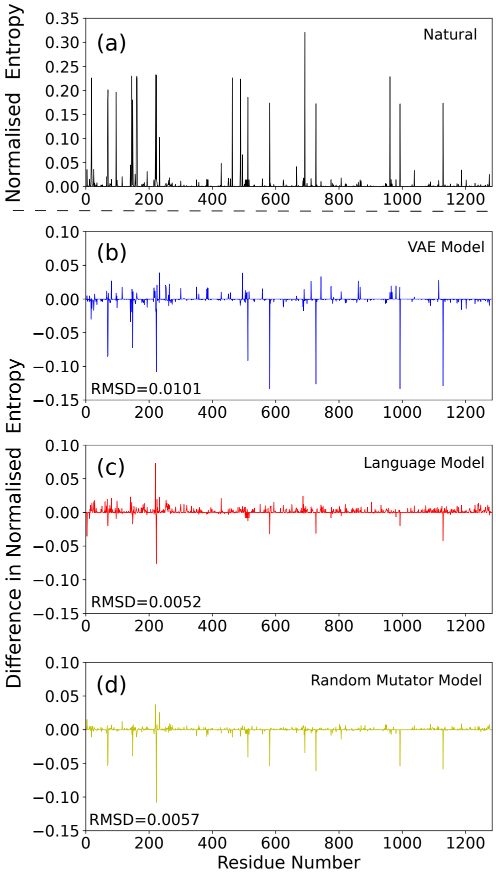

Figure 2. Normalised entropy comparisons of natural and generated sequences. (a) Residue-wise normalised entropy of the natural sequences. (b), (c), and (d) show the difference in residue-wise normalise entropy vs the natural sequences for the three generative models

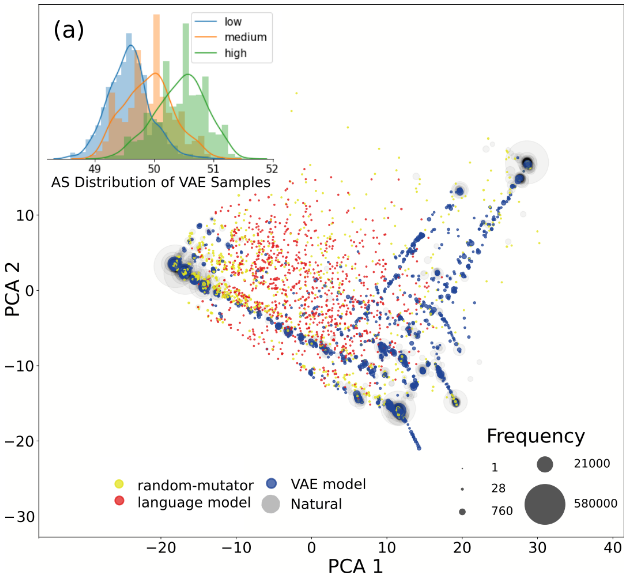

Figure 3. PCA visualisation of the 1000 most commonly generated SARS-CoV-2 spike protein sequences from each of the three generative models: random-mutator (yellow), language model (red), and VAE model (blue). Shown in grey are the 3000 most common naturally occurring sequences for comparison. Subfigure (a) shows the density distributions of antigenicity scores (AS) of low, medium and high antigenicity sequences sampled from the VAE. Note how the VAE sequences cluster faithfully around the natural sequences, whereas random-mutator and language model sequences generate many 'off-distribution' sequences.

Table 2. Mean values for RMSD compared to the Wuhan AlphaFold, RMSD compared to 7N1U, $\Delta\Delta G$ and AS — AS for each model and the consensus sequences.

MODEL	RMSD WA (Å)	RMSD 7N1U (Å)	$\Delta\Delta G$ (KCAL/MOL)	AS
VAE	0.48 ± 0.28	2.15 ± 0.07	-5.17 ± 0.51	49.93 ± 0.34
RANDOM	0.32 ± 0.23	2.13 ± 0.03	-2.51 ± 0.31	50.96 ± 0.67
11GRAM	0.41 ± 0.26	2.11 ± 0.03	-3.37 ± 1.82	50.15 ± 0.59

This illustrates that positional entropy RMSD calculations, one of the classical measures for evaluating sequence generative models, can be misleading. The reason being that generative models can often achieve very low positional entropy RMSD values without faithfully recreating the underlying distribution. This disparity is most apparent for simple models, such as N-gram language models or random-mutator models, that are effectively trained by construction to just mimic the global statistics of training data.

4.4. From Sequences to Structure

We follow the evaluation pipeline (Section 3.4) to investigate the generated sequences' tertiary structure: First, we took the 1000 most common generated sequences from all three generative models and checked that each sequence had no mutations in conserved regions. This led to three sequences from the low-antigenicity VAE model being rejected. Then, of the remaining sequences, we took the top 600 most common and calculated their $\Delta\Delta G$ with DDGun. The ten sequences with the lowest $\Delta\Delta G$ from each model were folded with AlphaFold2 and their RMSDs against reference SARS-CoV-2 structures were calculated (Table 2).

Notably, all three generative models produced stable structures with RMSD values falling in the expected natural range (i.e. RMSD WA$<$ 0.94Åand RMSD 7N1U$<$ 2.17Å, see Section 3.4). However, the VAE generated sequences that were ranked most stable by DDGun (-5.17kcal/mol compared to -2.51kcal/mol and -3.37kcal/mol for the random and language models respectively) whilst also having the lowest antigenicity of the three models, albeit marginally (see Table 2). This indicates that the VAE model approach is superior at generating stable, novel, structurally plausible sequences in addition to being superior at sampling sequences from a known natural distribution (Section 4.3). In Figure 4, Appendix we contrast the AlphaFolded structure of one of the most stably-ranked VAE proteins aligned with the corresponding chain of the SARS-CoV-2 Spike protein.

5. Conclusions and Further Work

We designed and evaluated a conditional Variational Autoencoder (VAE) capable of selectively generating novel SARS-CoV-2 spike proteins with low immune visibility. We discover that the VAE model can generate stable, structurally valid sequences that are smoothly distributed along the principal variance directions of natural sequences.

As in most vaccine efforts, we focused on the spike protein due to its high accessibility to antibodies. However, the spike protein is one of the most rapidly mutating regions of the genome. In contrast, the T-cell response that relies on the MHC-I pathway can also incorporate internal proteins of the virus; proteins that might be more evolutionarily stable. One way to develop this project further would be to incorporate more stable regions in the genome in a generative model, possibly leading to the identification of better-conserved peptides. Another direction of research would be to train a model that incorporates data on spike proteins from multiple viruses. This would increase generated sequence diversity as well as the variance in their AS.

At this stage it is not clear how comprehensively we are sampling the evolutionary space of SARS-CoV-2 spike proteins. To better understand this future works may apply a k-fold cross-validation by training on independent subsets of sequences and evaluating what fraction of the novel peptide 9-mers in unseen data is predicted by our model. Individual structures could then be experimentally investigated *in vitro*.

Finally, we note that the principles that underlie our conditional VAE architecture are not specific to SARS-CoV-2 but could also be easily extended to other rapidly mutating viruses such as MERS, influenza, and HIV.

Acknowledgements

This work was supported by the United Kingdom Research and Innovation (grant EP/S02431X/1), UKRI CDT in Biomedical AI at the University of Edinburgh, School of Informatics. We thank the PL-Grid and CI-TASK Infrastructure, Poland, for providing their resources. The work was supported by the project 'International Centre for Cancer Vaccine Science', International Agendas Programme of the Foundation for Polish Science co-financed by the European Union under the European Regional Development Fund. LR was supported by the Ministry of Science and Higher Education Decision No. 54/WFSN/2020. "Co-infections with SARS-CoV-2, Database of COVID-19 Accompanying Infections".

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. Optuna: A next-generation hyperparameter optimization framework. 2019. URL http://arxiv.org/abs/1907.10902.

Beans, C. Researchers getting closer to a "universal" flu vaccine. 119(5), 2022. doi: 10.1073/pnas.2123477119. URL https://www.pnas.org/doi/10.1073/pnas.2123477119. Publisher: Proceedings of the National Academy of Sciences.

Benler, S., Cobián-Güemes, A. G., McNair, K., Hung, S.-H., Levi, K., Edwards, R., and Rohwer, F. A diversity-generating retroelement encoded by a globally ubiquitous bacteroides phage. 6(1):191, 2018. ISSN 2049-2618. doi: 10.1186/s40168-018-0573-6. URL https://doi.org/10.1186/s40168-018-0573-6.

Borges, V., Alves, M. J., Amicone, M., Isidro, J., Zé-Zé, L., Duarte, S., Vieira, L., Guiomar, R., Gomes, J. P., and Gordo, I. Mutation rate of SARS-CoV-2 and emergence of mutators during experimental evolution, 2021. URL https://www.biorxiv.org/content/10.1101/2021.05.19.444774v1. Section: New Results Type: article.

Cai, Y., Zhang, J., Xiao, T., Lavine, C. L., Rawson, S., Peng, H., Zhu, H., Anand, K., Tong, P., Gautam, A., Lu, S., Sterling, S. M., Walsh, R. M., Rits-Volloch, S., Lu, J., Wesemann, D. R., Yang, W., Seaman, M. S., and Chen, B. Structural basis for enhanced infectivity and immune evasion of SARS-CoV-2 variants. 373(6555):642–648, 2021. ISSN 1095-9203. doi: 10.1126/science.abi9745.

Callaway, E. Beyond omicron: what's next for COVID's viral evolution. 600(7888):204–207, 2021. doi: 10.1038/d41586-021-03619-8. URL https://www.nature.com/articles/d41586-021-03619-8. Bandiera_abtest: a Cg_type: News Feature Number: 7888 Publisher: Nature Publishing Group Subject_term: SARS-CoV-2, Virology, Evolution.

Cao, Y., Wang, J., Jian, F., Xiao, T., Song, W., Yisimayi, A., Huang, W., Li, Q., Wang, P., An, R., Wang, J., Wang, Y., Niu, X., Yang, S., Liang, H., Sun, H., Li, T., Yu, Y., Cui, Q., Liu, S., Yang, X., Du, S., Zhang, Z., Hao, X., Shao, F., Jin, R., Wang, X., Xiao, J., Wang, Y., and Xie, X. S. Omicron escapes the majority of existing SARS-CoV-2 neutralizing antibodies. 602(7898):657–663, 2022. ISSN 1476-4687. doi: 10.1038/s41586-021-04385-3. URL https://www.nature.com/articles/s41586-021-04385-3. Number: 7898 Publisher: Nature Publishing Group.

Cele, S., Jackson, L., Khoury, D. S., Khan, K., Moyo-Gwete, T., Tegally, H., San, J. E., Cromer, D., Scheepers, C., Amoako, D. G., Karim, F., Bernstein, M., Lustig, G., Archary, D., Smith, M., Ganga, Y., Jule, Z., Reedoy, K., Hwa, S.-H., Giandhari, J., Blackburn, J. M., Gosnell, B. I., Abdool Karim, S. S., Hanekom, W., von Gottberg, A., Bhiman, J. N., Lessells, R. J., Moosa, M.-Y. S., Davenport, M. P., de Oliveira, T., Moore, P. L., and Sigal, A. Omicron extensively but incompletely escapes pfizer BNT162b2 neutralization. 602(7898):654–656, 2022. ISSN 1476-4687. doi: 10.1038/s41586-021-04387-1. URL https://www.nature.com/articles/s41586-021-04387-1. Number: 7898 Publisher: Nature Publishing Group.

Chhibbar, P. and Joshi, A. Generating protein sequences from antibiotic resistance genes data using generative adversarial networks. 2019. URL http://arxiv.org/abs/1904.13240.

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and de Hoon, M. J. L. Biopython: freely available python tools for computational molecular biology and bioinformatics. 25(11): 1422–1423, 2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp163. URL https://doi.org/10.1093/bioinformatics/btp163.

Craven. COVID-19 vaccine tracker, 2022. URL https://www.raps.org/news-and-articles/news-articles/2020/3/covid-19-vaccine-tracker.

Duvaud, S., Gabella, C., Lisacek, F., Stockinger, H., Ioannidis, V., and Durinx, C. Expasy, the swiss bioinformatics resource portal, as designed by its users. 49:W216–W227, 2021. ISSN 0305-1048. doi: 10.1093/nar/gkab225. URL https://doi.org/10.1093/nar/gkab225.

Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. 5(1):113, 2004. ISSN 1471-2105. doi: 10.1186/1471-2105-5-113. URL https://doi.org/10.1186/1471-2105-5-113.

El-Manzalawy, Y. and Honavar, V. Recent advances in b-cell epitope prediction methods. 6 Suppl 2:S2, 2010. ISSN 1745-7580. doi: 10.1186/1745-7580-6-S2-S2.

Fang, Y. Protein folding: The gibbs free energy. 2012. URL http://arxiv.org/abs/1202.1358.

Fleischmann, W. R. Viral genetics. In Baron, S. (ed.), *Medical Microbiology*. University of Texas Medical Branch at Galveston, 4th edition, 1996. ISBN 978-0-9631172-1-2. URL http://www.ncbi.nlm.nih.gov/books/NBK8439/.

Flemming, A. Omicron, the great escape artist. 22 (2):75–75, 2022. ISSN 1474-1741. doi: 10.1038/s41577-022-00676-6. URL https://www.nature.com/articles/s41577-022-00676-6. Number: 2 Publisher: Nature Publishing Group.

Fu, H., Li, C., Liu, X., Gao, J., Celikyilmaz, A., and Carin, L. Cyclical annealing schedule: A simple approach to mitigating KL vanishing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 240–250. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1021. URL https://aclanthology.org/N19-1021.

Gasser, H.-C., Bedran, G., Ren, B., Goodlett, D., Alfaro, J., and Rajan, A. Interpreting BERT architecture predictions for peptide presentation by MHC class i proteins. 2021. URL http://arxiv.org/abs/2111.07137.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html.

Greener, J. G., Moffat, L., and Jones, D. T. Design of metalloproteins and novel protein folds using variational autoencoders. 8(1):16189, 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-34533-1. URL https://www.nature.com/articles/s41598-018-34533-1. Number: 1 Publisher: Nature Publishing Group.

Hawkins-Hooker, A., Depardieu, F., Baur, S., Couairon, G., Chen, A., and Bikard, D. Generating functional protein variants with variational autoencoders. 17(2):e1008736, 2021. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1008736. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008736. Publisher: Public Library of Science.

Henikoff, S. and Henikoff, J. G. Amino acid substitution matrices from protein blocks. 89(22):10915–10919, 1992. ISSN 0027-8424. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC50453/.

Huang, Y., Yang, C., Xu, X.-f., Xu, W., and Liu, S.-w. Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. 41(9):1141–1149, 2020. ISSN 1745-7254. doi: 10.1038/s41401-020-0485-4. URL https://www.nature.com/articles/s41401-020-0485-4. Number: 9 Publisher: Nature Publishing Group.

J, K. and H, T. *Advanced PID control*. ISA-The Instrumentation, Systems, and Automation Society, 2006. ISBN 978-1-55617-942-6. OCLC: 60557376.

Jing, X., Dong, Q., Hong, D., and Lu, R. Amino acid encoding methods for protein sequences: A comprehensive review and assessment. 17(6):1918–1931, 2020. ISSN 1557-9964. doi: 10.1109/TCBB.2019.2911677.

Jolliffe, I. T. and Cadima, J. Principal component analysis: a review and recent developments. 374(2065): 20150202, 2016. doi: 10.1098/rsta.2015.0202. URL https://royalsocietypublishing.org/doi/10.1098/rsta.2015.0202. Publisher: Royal Society.

Kabsch, W. and Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. 22(12):2577–2637, 1983. ISSN 0006-3525. doi: 10.1002/bip.360221211.

Kingma, D. P. and Ba, J. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*, January 2017. URL http://arxiv.org/abs/1412.6980.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. In *arXiv:1312.6114 [cs, stat]*, 2014. URL http://arxiv.org/abs/1312.6114.

Kwon, S., Won, J., Kryshtafovych, A., and Seok, C. Assessment of protein model structure accuracy estimation in CASP14: Old and new challenges. *Proteins: Structure, Function, and Bioinformatics*, 89(12):1940–1948, 2021. ISSN 1097-0134. doi: 10.1002/prot.26192. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.26192.

Laimer, J., Hofer, H., Fritz, M., Wegenkittl, S., and Lackner, P. MAESTRO - multi agent stability prediction upon point mutations. 16(1):116, 2015. ISSN 1471-2105. doi: 10.1186/s12859-015-0548-6. URL https://doi.org/10.1186/s12859-015-0548-6.

Ledford, H., Cyranoski, D., and Van Noorden, R. The UK has approved a COVID vaccine — here's what scientists now want to know. 588(7837): 205–206, 2020. doi: 10.1038/d41586-020-03441-8. URL https://www.nature.com/articles/d41586-020-03441-8. Bandiera_abtest: a Cg_type: News Explainer Number: 7837 Publisher: Nature Publishing Group Subject_term: Vaccines, SARS-CoV-2.

Maaten, L. v. d. and Hinton, G. Visualizing data using t-SNE. 9(86):2579–2605, 2008. ISSN 1533-7928. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

Madani, A., McCann, B., Naik, N., Keskar, N. S., Anand, N., Eguchi, R. R., Huang, P.-S., and Socher, R. ProGen: Language modeling for protein generation, 2020. URL https://www.biorxiv.org/content/10.1101/2020.03.07.982272v1. Section: New Results Type: article.

Malone, B., Simovski, B., Moliné, C., Cheng, J., Gheorghe, M., Fontenelle, H., Vardaxis, I., Tennøe, S., Malmberg, J.-A., Stratford, R., and Clancy, T. Artificial intelligence predicts the immunogenic landscape of SARS-CoV-2 leading to universal blueprints for vaccine designs. 10(1):22375, 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-78758-5. URL https://www.nature.com/articles/s41598-020-78758-5. Number: 1 Publisher: Nature Publishing Group.

Montanucci, L., Capriotti, E., Frank, Y., Ben-Tal, N., and Fariselli, P. DDGun: an untrained method for the prediction of protein stability changes upon single and multiple point variations. 20(14):335, 2019. ISSN 1471-2105. doi: 10.1186/s12859-019-2923-1. URL https://doi.org/10.1186/s12859-019-2923-1.

Nachbagauer, R. and Krammer, F. Universal influenza virus vaccines and therapeutic antibodies. 23(4):222–228, 2017. ISSN 1198-743X. doi: 10.1016/j.cmi.2017.02.009. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5389886/.

Nielsen. NetMHCpan-4.1 online tool, 2020. URL https://services.healthtech.dtu.dk.

O'Donnell, T. J., Rubinsteyn, A., and Laserson, U. MHCflurry 2.0: Improved pan-allele prediction of MHC class i-presented peptides by incorporating antigen processing. 11(1):42–48.e7, 2020. ISSN 2405-4712. doi: 10.1016/j.cels.2020.06.010. URL https://www.sciencedirect.com/science/article/pii/S2405471220302398.

Pancotti, C., Benevenuta, S., Birolo, G., Alberini, V., Repetto, V., Sanavia, T., Capriotti, E., and Fariselli, P. Predicting protein stability changes upon single-point mutation: a thorough comparison of the available tools on a new dataset. pp. bbab555, 2022. ISSN 1477-4054. doi: 10.1093/bib/bbab555. URL https://doi.org/10.1093/bib/bbab555.

Petri, W. COVID-19 vaccines were developed in record time – but are these game-changers safe?, 2020.

Piepoli, S., Shamloo, B., Bİrcan, A., Adebali, O., and Erman, B. Molecular biology of sars-cov-2. 8:73–88, 2020. doi: 10.25002/tji.2020.1293.

Qiu, X., Duvvuri, V. R., and Bahl, J. Computational approaches and challenges to developing universal influenza vaccines. 7(2):E45, 2019. ISSN 2076-393X. doi: 10.3390/vaccines7020045.

Repecka, D., Jauniskis, V., Karpus, L., Rembeza, E., Rokaitis, I., Zrimec, J., Poviloniene, S., Laurynenas, A., Viknander, S., Abuajwa, W., Savolainen, O., Meskys, R., Engqvist, M. K. M., and Zelezniak, A. Expanding functional protein sequence spaces using generative adversarial networks. 3(4): 324–333, 2021. ISSN 2522-5839. doi: 10.1038/s42256-021-00310-5. URL https://www.nature.com/articles/s42256-021-00310-5. Number: 4 Publisher: Nature Publishing Group.

Reynisson, B., Alvarez, B., Paul, S., Peters, B., and Nielsen, M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. 48:W449–W454, 2020. ISSN 0305-1048. doi: 10.1093/nar/gkaa379. URL https://doi.org/10.1093/nar/gkaa379.

Rock, K. L., Reits, E., and Neefjes, J. Present yourself! by MHC class i and MHC class II molecules. 37(11):724–737, 2016. ISSN 1471-4906. doi: 10.1016/j.it.2016.08.010. URL https://www.sciencedirect.com/science/article/pii/S1471490616301004.

Sanavia, T., Birolo, G., Montanucci, L., Turina, P., Capriotti, E., and Fariselli, P. Limitations and challenges in protein stability prediction upon genome variations: towards future applications in precision medicine. 18:1968–1979, 2020. ISSN 2001-0370. doi: 10.1016/j.csbj.2020.07.011. URL https://www.sciencedirect.com/science/article/pii/S2001037020303433.

Sanchez-Trincado, J. L., Gomez-Perosanz, M., and Reche, P. A. Fundamentals and methods for t- and b-cell epitope prediction. 2017, 2017. ISSN 2314-8861. doi: 10.1155/2017/2680160. URL https://www.hindawi.com/journals/jir/2017/2680160/. Publisher: Hindawi.

Sanjuán, R. and Domingo-Calap, P. Mechanisms of viral mutation. 73(23):4433–4448, 2016. ISSN 1420-682X. doi: 10.1007/s00018-016-2299-6. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5075021/.

Shao, H., Yao, S., Sun, D., Zhang, A., Liu, S., Liu, D., Wang, J., and Abdelzaher, T. ControlVAE: Controllable variational autoencoder. 2020. URL http://arxiv.org/abs/2004.05988.

Shin, J.-E., Riesselman, A. J., Kollasch, A. W., McMahon, C., Simon, E., Sander, C., Manglik, A., Kruse, A. C., and Marks, D. S. Protein design and variant prediction using autoregressive generative models, 2021. URL https://www.biorxiv.org/content/10.1101/757252v2. Section: New Results Type: article.

Shu, Y. and McCauley, J. GISAID: Global initiative on sharing all influenza data – from vision to reality. 22(13): 30494, 2017. ISSN 1025-496X. doi: 10.2807/1560-7917.ES.2017.22.13.30494. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5388101/.

Sinai, S., Kelsic, E., Church, G. M., and Nowak, M. A. Variational auto-encoding of protein sequences. 2018. URL http://arxiv.org/abs/1712.03346.

Sohn, K., Lee, H., and Yan, X. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/hash/8d55a249e6baa5c06772297520da2051-Abstract.html.

Strokach, A. and Kim, P. M. Deep generative modeling for protein design. 72:226–236, 2022. ISSN 0959-440X. doi: 10.1016/j.sbi.2021.11.008. URL https://www.sciencedirect.com/science/article/pii/S0959440X21001573.

Touw, W. G., Baakman, C., Black, J., te Beek, T. A. H., Krieger, E., Joosten, R. P., and Vriend, G. A series of PDB-related databanks for everyday needs. 43:D364–368, 2015. ISSN 1362-4962. doi: 10.1093/nar/gku1028.

Van Der Maaten, L., Postma, E., and Van den Herik, J. Dimensionality reduction: a comparative review. 10: 66–71, 2009.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. 2017. URL http://arxiv.org/abs/1706.03762.

Wang, C., Horby, P. W., Hayden, F. G., and Gao, G. F. A novel coronavirus outbreak of global health concern. 395(10223):470–473, 2020. ISSN 0140-6736, 1474-547X. doi: 10.1016/S0140-6736(20)30185-9. URL https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)30185-9/fulltext. Publisher: Elsevier.

Wu, Z., Yang, K. K., Liszka, M. J., Lee, A., Batzilla, A., Wernick, D., Weiner, D. P., and Arnold, F. H. Signal peptides generated by attention-based neural networks. 9 (8):2154–2161, 2020. doi: 10.1021/acssynbio.0c00219. URL https://doi.org/10.1021/acssynbio.0c00219. Publisher: American Chemical Society.

Yuan, S., Chan, H. S., and Hu, Z. Using PyMOL as a platform for computational drug design. 7(2): e1298, 2017. ISSN 1759-0884. doi: 10.1002/wcms.1298. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/wcms.1298.

6. Acronyms

AA amino acid. 2–5, 13

AE Autoencoder. 2

AR autoregressive. 2

AS antigenicity score. 4, 5, 8, 17

CTL Cytotoxic T-lymphocytes. 2

DGR diversity-generating retro-element. 2

EL eluted ligand. 4

ER endoplasmic reticulum. 2

GAN Generative Adversarial Network. 2, 4

GISAID Global Initiative on Sharing All Influenza Data. 3

LSTM Long short-term memory. 2

MDH malate dehydrogenase. 2

MHC-I MHC class I. 2, 4, 8

MSA multiple sequence alignment. 4

pMHC peptide-MHC protein complex. 2

RMSD root-mean-square deviation. 6, 8

RNA ribonucleic acid. 2

RNN recurrent neural network. 2

SARS-CoV-2 Severe acute respiratory syndrome coronavirus 2. 1–4, 6–8, 14, 15

TCR T-cell receptor. 2

VAE Variational Autoencoder. 1–4, 6, 8, 15

7. Appendix

7.1. Sequence encoding preliminary experiments

We experimented with 13 different types of position-independent AA encodings and quantitatively compared these based on the fraction of explained variance accounted for by the first two dimensions of their principal components analysis (Table 3). By this metric, Meiler parameters performed best whilst most other methods were comparable. However, visual inspection of t-SNE plots revealed that all methods, apart from Micheletti potentials (which performed noticeably worse), led to similar clustering quality and appearance.

Since there was no visual basis on which to select one method over another, we followed the advice of Jing et al. 2020 and used BLOSUM62; in tests, the authors showed that, amongst position-independent encodings, BLOSUM62 had the best performance in protein fold recognition tasks (relevant for epitope visualisation).

7.2. VAE training and hyperparameter optimization

Training: We used Adam (Kingma & Ba, 2017) ($\beta_1 = 0.9, \beta_2 = 0.999$, learning rate of $\alpha = 3 \times 10^{-4}$, weight decay 1×10^{-6}) to optimize the VAE loss described in Section 2.3 including the control mechanism by Shao et al. (using $K_p = 0.001, K_i = 0.0005, K_d = 0$ and a minimum beta of 0.0001 and a maximum beta of 1).

Table 3. Percentage of variation of SARS-CoV-2 spike protein data explainable by the first two dimensions of PCA when using different types of position-independent encodings. Higher is better. a) One hot, b) One hot 6 bit, c) Binary 5 bit, d) Hydrophobicity matrix, e) Meiler parameters, f) Acthely factors, g) PAM250, h) BLOSUM62, i) Miyazawa energies, j) Micheletti potentials, k) AESNN3, l) ANN4D, m) ProtVec.

	A)	B)	C)	D)	E)	F)	G)	H)	I)	J)	K)	L)	M)
WITHOUT EPITOPE MASK	53	54	56	75	78	55	54	57	61	34	52	67	51
WITH EPITOPE MASK	60	64	64	75	82	67	65	64	69	76	58	74	60

Hyperparameter optimization: We search for the hyper-parameters in Table 4 using the optuna framework (Akiba et al., 2019). This utilizes a tree-structured Parzen estimator mechanism to enhance the search in this high-dimensional space. The aim is to find hyper-parameters that result in the best generated sequences. As it would be unfeasible to perform all sequence assessment steps we did after the model training for each candidate hyper-parameter setting, used a ballpark assessment method. We decided to minimise the Euclidean distance between the position-wise entropy vector of our training dataset to the position-wise entropy vector of a set of 100 randomly generated low antigenicity sequences utilizing a candidate hyper-parameters setting. We ran 30 trials for 25 epochs each. The finally retrieved best parameters (column "selected" in Table 4 of the Appendix 7.2) were trained for a total of 100 epochs.

Table 4. Result of the hyperparameter search.

HYPERPARAMETER	OPTIONS	SELECTED
NUMBER OF BLOCKS	2 TO 7	5
FIRST HIDDEN DIMENSION	512, 1024 OR 2048	512
LATENT SPACE DIMENSION	2 TO 50	30
DROPOUT PERCENTAGE	5% TO 50%	0.394
KL TARGET	0.01 TO 1.00	0.232

7.3. New epitopes

The 78 novel 9-mers that were generated by the VAE model that did not appear in the training data sequences.

```
TPINLVDDL, FLPFFSNVW, TRFQLHRSY, LGHKNNKSW, TRFQHRSYL, DVHKNNKSW,
VEPDLPQGF, INITRFQTA, RSYDSSSGW, EPLPQGFSA, SCMESEFVY, HAINGTKRF,
TPIIVEPDL, FQTALHRSY, FLDVHKNNK, ALHDSSSGW, ITRFQHRSY, FFSNVWHAI,
EEPELPQGF, LHNSSGWTA, KSWMESERV, LVGGLPQGF, LPFFSNVWF, KSCMESEVY,
YLGDSSSGW, QTALHRSYL, KSWMESEFV, EPDELPQGF, IIVEPEEPL, PEEPLPQGF,
SWMESEFVY, SWMESERVY, SWMESESVY, VEPEEPEDL, IVEPEEPEL, FSNVITKRF,
SYGDSSSGW, HAISGNGTK, RFQLHRSYL, FHAISGNGT, ISGNGTKRF, HSGNGTKRF,
HRSDSSSGW, AISNGTKRF, NITRFQTAL, TRFQTALHR, PFFSNVKRF, HRSNSSSGW,
LALHNSSGW, HAISNGTKR, KSWMESESV, SWMESGFVY, CNDPFLDYY, LVDDLPQGF,
AISGNGTKR, LPGGSSSGW, NIVDLPQGF, EEPEEPEDL, TPINLVGGL, TPINLVDDL,
TRFQLHRSY, HTPINLVRL, LGHKNNKSW, TRFQHRSYL, DVHKNNKSW, INITRFQTA,
RSYDSSSGW, EPLPQGFSA, SCMESEFVY, FQTALHRSY, FLDVHKNNK, ALHDSSSGW,
ITRFQHRSY, EEPELPQGF, KSWMESERV, LVGGLPQGF, VRLPQGFSA, YLGDSSSGW,
NLVRLPQGF, QTALHRSYL, KSWMESEFV, IIVEPEEPL, PEEPLPQGF, SWMESGSVY,
SWMESEFVY, SWMESERVY, IVEDLPQGF, VEPEEPEDL, SWMESESVY, IVEPEEPEL,
SYGDSSSGW, RFQLHRSYL, HRSDSSSGW, AISNGTKRF, NITRFQTAL, TRFQTALHR,
FFSNVTKRF, HRSNSSSGW, HAISNGTKR, KSWMESESV, SWMESGFVY, CNDPFLDYY,
SCMESGFVY, VRDDLPQGF, NIVDLPQGF, LPGDSSSGW, EEPEEPEDL, TPINLVGGL,
VNFRNRTQL, TRFQLHRSY, HTPINLVRL, EPRDLPQGF, TPIIVERDL, LGHKNNKSW,
TRFQHRSYL, DVHKNNKSW, INITRFQTA, RSYDSSSGW, EPLPQGFSA, SCMESEFVY,
FQTALHRSY, FLDVHKNNK, ALHDSSSGW, ITRFQHRSY, EEPELPQGF, KSWMESERV,
VNRTNRTQL, VRLPQGFSA, NLVRLPQGF, QTALHRSYL, KSWMESEFV, IIVEPEEPL,
PEEPLPQGF, SWMESGSVY, SWMESEFVY, SWMESERVY, IVEDLPQGF, VEPEEPEDL,
IVEPEEPEL, SYGDSSSGW, KHTPIIVER, RFQLHRSYL, HRSDSSSGW, AISNGTKRF,
NITRFQTAL, VRDDLPQGF, YPGDSSSGW, TRFQTALHR, VERDLPQGF, VRGGLPQGF,
HAISNGTKR, SWMESGFVY, CNDPFLDYY, FSNVGTKRF, EPEPEPEDL, NIVDLPQGF,
LPGDSSSGW, EEPEEPEDL, LVDDLPQGF.
```

7.4. Evaluation outcomes

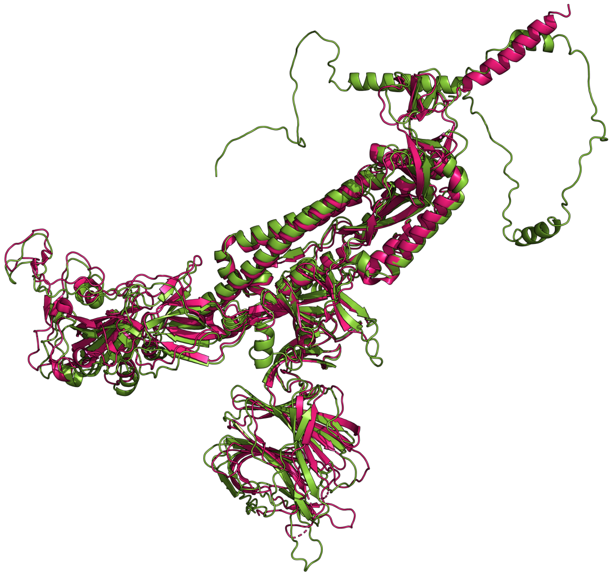

Figure 4. Aligned Spike proteins of SARS-CoV-2. Pink: 7N1U (chain A). Green: an example of a VAE model generated structure (VAE_9, Table 7). Note the striking similarity between the structures, especially in the regions far away from the flexible terminus of the protein.

Table 5. Root-mean-squared deviation (RMSD, Å) between all pairs of consensus sequence structures generated by AlphaFold.

SEQUENCE NAME	WUHAN	ALPHA	BETA	DELTA	EPSILON	GAMMA	OMICRON
WUHAN	0.0						
ALPHA	0.897	0.0					
BETA	0.941	0.225	0.0				
DELTA	0.274	0.799	0.830	0.0			
EPSILON	0.931	0.228	0.186	0.834	0.0		
GAMMA	0.408	0.840	0.903	0.227	0.898	0.0	
OMICRON	0.336	0.831	0.825	0.270	0.875	0.329	0.0

Table 6. Root-mean-squared deviation (RMSD, Å) between the AlphaFolded Wuhan structure and the 7N1U structure compared to the AlphaFolded structures of the consensus sequences.

SEQUENCE NAME	WUHAN ALPHAFOLDED	7N1U
WUHAN	-	2.135
ALPHA	0.897	2.164
BETA	0.941	2.168
DELTA	0.274	2.069
EPSILON	0.931	2.152
GAMMA	0.408	2.043
OMICRON	0.336	2.125

Table 7. Evaluating the top 10 predicted most-stable generated sequences (according to DDGun) for each of the three generative models: VAE, 11gram language model, and random mutator model. We present RMSD values of the generated sequence's AlphaFolded structure compared to the Wuhan AlphaFold and 7N1U. Also shown is the estimated change in stability ($\Delta\Delta G$), and the antigenicity score (AS) of each generated sequence. Table 2 (main text) shows the average of these results across each model.

Sequence Name	RMSD with WA (Å)	RMSD with 7N1U (Å)	$\Delta\Delta G$ (kcal/mol)	AS
VAE_1	0.915	2.173	-6.2	50.00
VAE_2	0.263	2.078	-5.7	49.83
VAE_3	0.167	2.120	-5.7	50.17
VAE_4	0.343	2.135	-5.3	49.75
VAE_5	1.008	2.135	-5.0	50.67
VAE_6	0.216	2.095	-4.9	50.08
VAE_7	0.635	2.302	-4.8	49.50
VAE_8	0.381	2.166	-4.8	49.67
VAE_9	0.636	2.217	-4.7	49.50
VAE_10	0.282	2.081	-4.6	50.08
11gram_1	0.974	2.108	-8.3	50.17
11gram_2	0.227	2.137	-5.1	50.67
11gram_3	0.868	2.116	-2.9	49.25
11gram_4	0.469	2.096	-2.7	50.75
11gram_5	0.281	2.112	-2.7	51.25
11gram_6	0.216	2.144	-2.5	50.33
11gram_7	0.260	2.074	-2.4	49.42
11gram_8	0.298	2.050	-2.4	49.83
11gram_9	0.271	2.119	-2.4	49.75
11gram_10	0.270	2.109	-2.3	50.08
random_mut_1	0.216	2.074	-3.0	51.08
random_mut_2	0.982	2.211	-2.8	51.00
random_mut_3	0.214	2.109	-2.8	51.33
random_mut_4	0.221	2.114	-2.5	49.17
random_mut_5	0.209	2.142	-2.5	51.08
random_mut_6	0.173	2.129	-2.5	51.25
random_mut_7	0.386	2.137	-2.5	51.33
random_mut_8	0.190	2.128	-2.5	51.58
random_mut_9	0.340	2.120	-2.1	50.42
random_mut_10	0.222	2.105	-1.9	51.33