Question 1

a) What is the interatomic distance at which $\boldsymbol{V = 0}$?

When $r = \sigma$

b) Express $\boldsymbol{r_m}$ as a function of appropriate parameters defined above

$$r_m = \sigma\left(\sqrt[n-m]{\frac{n}{m}}\right)$$

Obtain the depth of the potential, i.e. $\boldsymbol{V(r_m)}$

$$V(r_m) = \left[\frac{n}{n-m}\left(\frac{n}{m}\right)^{\frac{m}{n-m}}\right]\epsilon\left[\left(\frac{\sigma}{\sigma\left(\sqrt[n-m]{\frac{n}{m}}\right)}\right)^{n} - \left(\frac{\sigma}{\sigma\left(\sqrt[n-m]{\frac{n}{m}}\right)}\right)^{m}\right]$$

and then simplify the expression for $\boldsymbol{V(r_m)}$ as much as possible

$$\mathrm{V}(r_m) = \left[\frac{n}{n-m}\left(\frac{n}{m}\right)^{\frac{m}{n-m}}\right]\epsilon\left[\frac{\sigma^{\mathrm{n}}}{\left(\sqrt[n-m]{\frac{n}{m}}\right)^{n}\sigma^{n}} - \frac{\sigma^{\mathrm{m}}}{\left(\sqrt[n-m]{\frac{n}{m}}\right)^{m}\sigma^{m}}\right]$$

$$= \left[\frac{n}{n-m}\left(\frac{n}{m}\right)^{\frac{m}{n-m}}\right]\epsilon\left[\frac{1}{\left(\frac{n}{m}\right)^{\frac{n}{n-m}}} - \frac{1}{\left(\frac{n}{m}\right)^{\frac{m}{n-m}}}\right]$$

$$= \frac{\epsilon\frac{n}{n-m}\left(\frac{n}{m}\right)^{\frac{m}{n-m}}}{\left(\frac{n}{m}\right)^{\frac{n}{n-m}}} - \frac{\epsilon\frac{n}{n-m}\left(\frac{n}{m}\right)^{\frac{m}{n-m}}}{\left(\frac{n}{m}\right)^{\frac{m}{n-m}}}$$

$$= \frac{\epsilon\mathrm{n}}{n-m}\left(\frac{n}{m}\right)^{\frac{m-n}{n-m}} - \frac{\epsilon\mathrm{n}}{n-m}$$

$$= \frac{\epsilon n}{n-m}\frac{m}{n} - \frac{\epsilon\mathrm{n}}{n-m}$$

$$= \frac{\epsilon mn - \epsilon n^2}{n^2 - mn}$$

$$= \frac{-\epsilon(n^2 - mn)}{n^2 - mn}$$

$$= -\epsilon$$

c) Express the interatomic potential $\boldsymbol{V(r)}$ in terms of $\boldsymbol{r_m}$ and instead of $\boldsymbol{\sigma}$

Rearrange to give σ:

$$\sigma = \frac{r_m}{\sqrt[n-m]{\frac{n}{m}}}$$

Substituting into $V(r)$:

$$V(r) = \left[\frac{n}{n-m}\left(\frac{n}{m}\right)^{\frac{m}{n-m}}\right]\epsilon\left[\left(\frac{\frac{r_m}{\sqrt[n-m]{\frac{n}{m}}}}{r}\right)^{n} - \left(\frac{\frac{r_m}{\sqrt[n-m]{\frac{n}{m}}}}{r}\right)^{m}\right]$$

Simplify the resulting expression for $\boldsymbol{V(r)}$ as much as possible

$$V(r) = \left[\frac{n}{n-m}\left(\frac{n}{m}\right)^{\frac{m}{n-m}}\right]\epsilon\left[\left(\frac{\mathrm{r_m m}\frac{1}{\mathrm{n-m}}}{\mathrm{r n}^{\frac{1}{\mathrm{n-m}}}}\right)^{n} - \left(\frac{\mathrm{r_m m}\frac{1}{\mathrm{n-m}}}{\mathrm{r n}^{\frac{1}{\mathrm{n-m}}}}\right)^{m}\right]$$

$$= \frac{\epsilon\, m(r_m)^n}{r^n(n-m)} - \frac{\epsilon\, n(r_m)^m}{r^m(n-m)}$$

$$= \frac{\epsilon m(r_m)^n r^m - \epsilon n(r_m)^m r^n}{r^{m+n}(n-m)}$$

Write down $\boldsymbol{V(r)}$ in this form for the Lennard-Jones potential

$$\mathrm{V(r)} = \epsilon\frac{(r_m)^6 - 2r^6}{r^{12}}$$

d) Investigate the following function $\boldsymbol{f}: [\mathbf{0},\mathbf{1}] \mapsto [\mathbf{0},\mathbf{1}], \boldsymbol{f(t)} = -\boldsymbol{t^2(2t-3)}$ by finding $\boldsymbol{t_m} = \arg\max_{\boldsymbol{t}}[\boldsymbol{f(t)}]$ and the values $\boldsymbol{f(0)}$, $\boldsymbol{f(1)}$, $\boldsymbol{f'(0)}$ and $\boldsymbol{f'(1)}$

$$f(t) = -t^2(2t-3)$$
$$f'(t) = -6t(t-1)$$

$$t_m = 1$$
$$f(0) = 0$$
$$f(1) = 1$$
$$f'(0) = 0$$
$$f'(1) = 0$$

e) Utilise $\boldsymbol{f}$ defined in (d) to define a proper $\boldsymbol{S(r)}$

$$\boldsymbol{S(r)} = \begin{cases} \mathbf{0} & \boldsymbol{r \geq r_c} \\ \tilde{\boldsymbol{f}}\left(\dfrac{\boldsymbol{r_c - r}}{\boldsymbol{r_c - r_s}}\right) & \boldsymbol{r_s < r < r_c} \\ \mathbf{1} & \boldsymbol{r \leq r_s} \end{cases}$$

Where

$$\tilde{\mathrm{f}}(t) = \begin{cases} \mathbf{0} & \boldsymbol{t \leq 0} \\ -\mathbf{t^2(t-3)} & \boldsymbol{0 < t < 1} \\ \mathbf{1} & \boldsymbol{t \geq 1} \end{cases}$$

When $r \geq r_c$, $\tilde{\phi}(r) = \phi(r)S(r) = \phi(r) * 0 = 0$, achieving goal 2.

When $r \leq r_s$, $\tilde{\phi}(r) = \phi(r)S(r) = \phi(r) * 1 = \phi(r)$, achieving goal 3.

To check if $\tilde{\phi}(r)$ is continuous, we know that $\phi(r)$ is continuous so we only need to check $S(r)$. We first check $\tilde{f}(t)$. For $\tilde{f}(t)$ to be continuous, its three piece-functions must be continuous, and the two boundary points must meet. $f(t)$ is continuous (which can be shown through graphing it), and so are the two constant functions for 0 and 1. We showed in (d) that $f(0) = 0$ and $f(1) = 1$, thus $\tilde{f}(t)$ is also continuous.

The three piece-functions of $S(r)$ are therefore all continuous, so we now only need to check that the boundary points meet. We calculate $S(r_c) = \tilde{f}(0) = 0$ and $S(r_s) = \tilde{f}(1) = 1$. Thus the boundary points at r_c and r_s match between the piecewise functions, so $S(r)$ is continuous, so $\phi(r)S(r)$ is also continuous.

To show that $\tilde{\phi}'(r)$ is continuous, we first calculate that:

$$\boldsymbol{S'(r) = \begin{cases} 0 & r > r_c \\ \tilde{f}'\left(\dfrac{r_c - r}{r_c - r_s}\right) & r_s \le r \le r_c \\ 0 & r < r_s \end{cases}}$$

and

$$\tilde{f}'(t) = \begin{cases} 0 & t \le 0 \\ -6t(t-1) & 0 < t < 1 \\ 0 & t \ge 1 \end{cases}$$

Again, as noted in (d), $f'(0) = 0$ and $f'(1) = 0$. Since the three piece-functions are continuous and the boundary points meet, $\tilde{f}'(t)$ is continuous. Likewise, since $S'(r_c) = \tilde{f}'(0) = 0$ and $S'(r_s) = \tilde{f}'(1) = 0$, the three piece-functions meet at the boundary points (and the functions are all continuous), so $\tilde{\phi}'(r)$ is continuous. This achieves goal 1.

f) Considering the interatomic pair potentials $\tilde{\phi}_f(r)$ and $\tilde{\phi}_g(r)$ which one would be suitable for use given that one would like to meet the following conditions

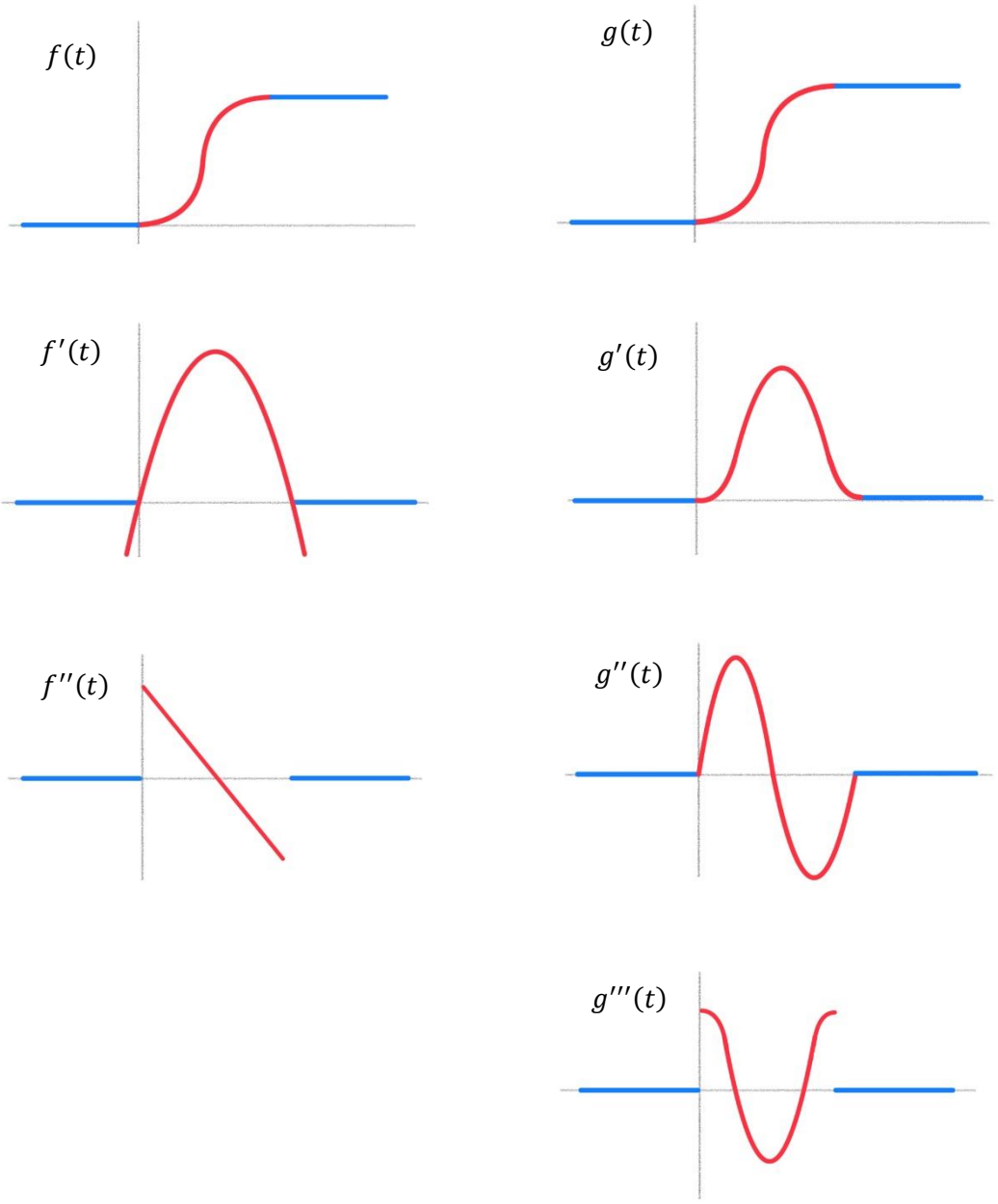
We first note the following equations:

$$f(t) = -t^2(2t-3)$$
$$f'(t) = -6t(t-1)$$
$$f''(t) = 6 - 12t$$

and:

$$g(t) = t - \frac{\sin(2\pi t)}{2\pi}$$
$$g'(t) = 1 - \cos(2\pi t)$$
$$g''(t) = 2\pi \sin(2\pi t)$$
$$g'''(t) = 4\pi^2 \cos(2\pi t)$$

The following are graphical representations of these functions:



We use the same reasoning as in (e) to calculate the continuousness of the functions. More formally, we can show that, for example, $f(t)$ is continuous by splitting it into three functions, $f_1(t) = 0$, $f_2(t) = -t^2(2t - 3)$ and $f_3(t) = 1$ (corresponding to the blue, red, and blue region in the $f(t)$ graph). We then note that at the first boundary point:

1. $f_1(0) = 0$
2. $f_2(0) = 0$
3. $\lim_{t \to 0^-} f_1(t) = 0$
4. $\lim_{t \to 0^+} f_2(t) = 0$

And at the second boundary point:

1. $f_2(1) = 1$
2. $f_3(1) = 1$
3. $\lim_{t \to 1^-} f_2(t) = 1$

4. $\lim_{t \to 1^+} f_3(t) = 1$

Thus at each boundary point, the values of the two adjacent functions and the limits as each function tends toward the boundary point are all equal. $f_1(t)$, $f_2(t)$ and $f_3(t)$ are continuous, so the whole function $f(t)$ is therefore continuous.

We apply these calculations to all seven functions to show that (by calculating continuity for all functions and then using the continuity of the derivative to find whether the function is smooth):

Function	Continuous?	Smooth?
$f(t)$	Yes	Yes
$f'(t)$	Yes	No
$f''(t)$	No	No
$g(t)$	Yes	Yes
$g'(t)$	Yes	Yes
$g''(t)$	Yes	No
$g'''(t)$	No	No

i) All interatomic pair potentials and their first derivatives (for calculating forces) are continuous functions

By comparing the functions f and g, we can see that both f and g would be suitable candidates, as both functions and their derivates are continuous. However, f has the benefit of not needing to use trigonometric functions, so if $\tilde{\phi}(r)$ is computationally intensive and requires a large number of evaluations of its constituent function (e.g. if it is used for calculating forces), f would be computationally easier to calculate

ii) All interatomic potentials and their first derivates are smooth functions [a smooth function is a function that is not only continuous but also has a continuous first derivative (or continuous derivatives up to an even higher order).]

As we can see from the above table, only g is suitable for this, as both g and g' are smooth (since g, g' and g'' are all continuous). Therefore only $\tilde{\phi}_g(r)$ is viable.

Question 2: generative modelling for (de novo) protein design

a) Briefly describe a computational biology topic of your choice, its brief history and provide a motivation for choosing it (0.5 pages, 5 marks)

Overview of generative modelling of de novo protein design

Generative modelling of de novo protein design is a field that aims to produce novel proteins that do not exist in nature. De novo protein design can be viewed as the reverse problem of the protein structure prediction problem [1] (which aims to calculate a protein's three-dimensional structure given its amino acid sequence [2]: given a three-dimensional structure, generate an amino acid sequence that will fold to form that structure. There are two types of de novo protein design: constrained and unconstrained. Constrained protein design involves generating a protein according to functional or structural requirements [3], such as a protein that contains a specific structural motif. Unconstrained protein design places no restrictions on the protein generation, and involves using generative modelling techniques to freely explore the protein search space and discover new proteins [4].

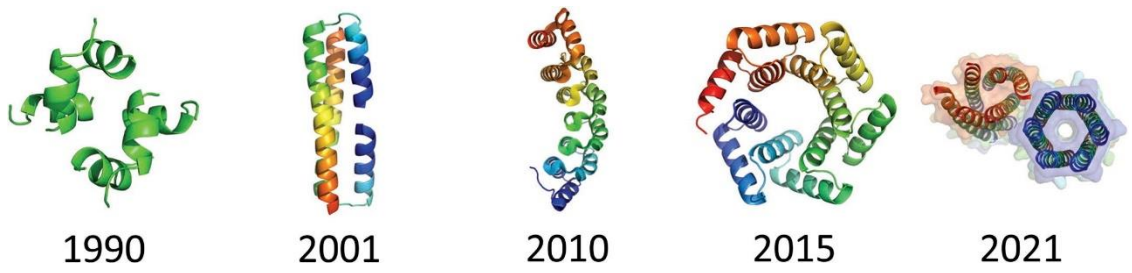
History of generative modelling of de novo protein design

The structure prediction problem and the protein design problem are equivalent, so advances in structure prediction methods lead to advances in protein design method. This is done by using statistical methods to search the sequence space by iteratively predicting the structure of a starting sequence and then improving the sequence in the desired direction towards the target structure. Such techniques include [5] stochastic algorithms (popular ones are Monte Carlo techniques and simulated annealing), dynamic programming, graph search, evolution-guided techniques, and dead-end elimination. These algorithms are still widely used to efficiently search the sequence space.

Historically, protein design mainly involved modifying *existing* proteins, using techniques like homology modelling [6, 7] (sometimes called comparative modelling), which predicts protein structure based on similarity with homologous proteins, and threading [8], which predicts structure based on the fingerprints of families of proteins. The recent development of AlphaFold [2] in 2021, which can predict protein structure to subatomic accuracy, has spurred a new wave of developments in de novo protein design.

Motivation for choosing this topic

There are many applications for the design of novel proteins, from nanotechnology [9] to personalised medicine [10], and I think this area has lots of potential for exciting future innovations.



This image from [11] neatly summarises the evolution of de novo protein design over the last few decades.

b) State some of the computational goal(s) or problem(s) most relevant to the topic. Aim to describe the computational goal(s) or formulate the problem in a way understandable for computer scientists (0.5 pages, 10 marks)

The computational goal of generative modelling of de novo protein design

The central goal of generative modelling of de novo proteins is the generation of an amino acid sequence that folds into a (partially, wholly or un-) specified structure. There are two challenges that complicate this process.

Problem 1

The first problem in de novo protein design is that the search space is exponentially large [12]. Proteins have a large number of degrees of freedom (each peptide bond has 2 degrees of freedom [13], and each amino acid can have over 100 possible rotamer states [5]), yet can fold into their native structure in a matter of nanoseconds. If the folding process sampled each possible structure at random, it would take a vast amount of time to find the native structure: this is known as the Levinthal paradox [14]. Thus the first challenge is to create computational techniques that can efficiently search this sequence space in a non-astronomical length of time.

Problem 2

The second problem in de novo protein design is understanding the complex biological and chemical processes that mediate protein folding. The major forces governing protein folding are hydrophobic forces [15], in which non-polar hydrophobic regions of the protein fold inwards, away from surrounding water molecules. Other forces such as hydrogen bonds, van der Waals interactions, disulfide bonds, electrophilic and electrostatic attraction also contribute to the protein folding. More and more refined models are being created, but these become computationally more difficult to calculate [16].

c) Describe how molecular biology would benefit from achieving the computational goal(s) or solving the stated problem(s) (0.5 pages, 5 marks)

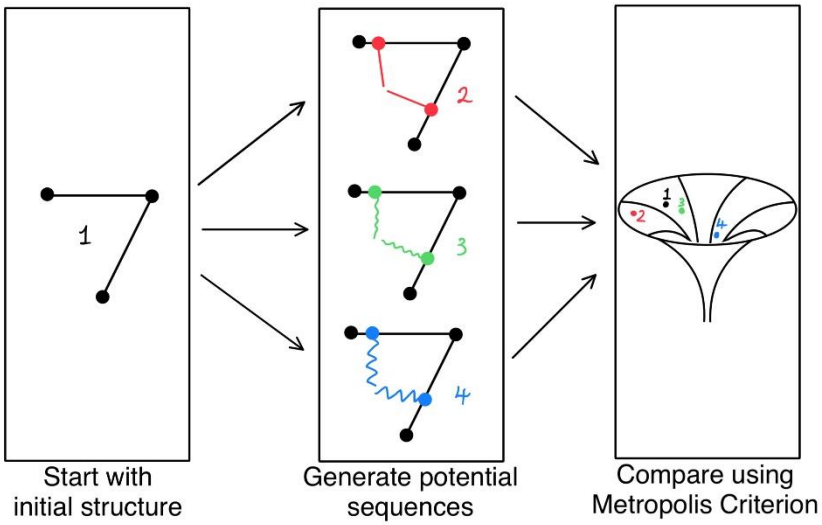
Molecular biology has already benefited enormously from de novo protein generation, with de novo proteins having already been developed for: wastewater treatment [17], nanotechnology [18], antimicrobial pesticides [19], luminescent biotechnology [20], COVID-19 biosensors [21], and RSV vaccine design [22]. Further advancements in the field would likely bring even more peptide successes.

d) Characterise some existing solutions (to date) to this problem available in the literature and clearly cite references

i) Describe the underlying ideas or algorithms associated with these solutions using schemes, pseudocodes, equations, model architectures (as applicable) to demonstrate your understanding of the technical background related to these solutions (2 pages, 20 marks)

Rosetta model

A simplified model I have created (inspired by Figure 1 of [23]) of a step in Rosetta's algorithm is:

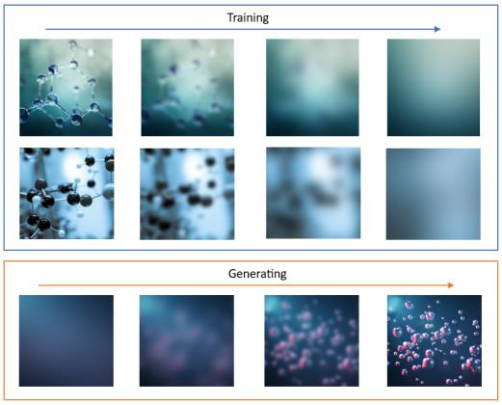


1. For the first step, either a randomly-generated backbone is used (if the goal is unconditional generation) or a short section of desired backbone (if the goal is conditional generation: to build a protein with a specific motif or binding site). For subsequent steps, use the structure from the previous step
2. Potential sequences are usually generated with a 'Mover' (as described in [24]). This involves either generating new residues, or generating new rotamers for an existing residue.

   A scoring function is then used to calculate the energy of each sequence (where a lower energy score means better stability). The most recent scoring function that has been developed for Rosetta is ref2015, which is a linear weighted combination of 19 different energy terms [25], such as:
   a. *hbond_sr_bb*, *hbond_lr_bb*, *hbond_bb_sc*, and *hbond_sc* which all measure the hydrogen bonding used to bury polar atoms away from the solvent
   b. *dslf_fa13* which measures 'disulfide geometry potential', i.e. the strength of disulfide bonds between cysteines
   c. *fa_atr* and *fa_rep* which measure the Lennard-Jones attraction and repulsion respectively
3. Instead of deterministically selecting the sequence with the lowest energy (which can result in getting trapped in local energy minima), Rosetta uses a nondeterministic stochastic technique, specifically a Monte Carlo method [26] with a Metropolis criterion [24], which checks whether a sequence has an improved energy score (i.e. is strictly lower than the current), in which case it accepts, and if not, it selects it with probability (at step t): $e^{-(Energy_{new}-Energy_{current})/t}$

RFdiffusion model

RFdiffusion is a state-of-the-art protein generation model, standing for RoseTTAFold Diffusion [27]. It incorporates the biological power of Rosetta with the computational power of diffusion models to generate de novo proteins both unconditionally and conditionally (it can be conditioned on topology, symmetry, motif and active site scaffolding). Diffusion models work by steadily corrupting target data (by adding noise, sampled from a Gaussian distribution), and training the model to learn the denoising process, in order to learn the distribution [28].



This diagram illustrates the guiding principle behind diffusion models (created using Microsoft stock images, and based on diagrams I found during research, such as [29]). The model learns to turn random noise into meaningful outputs, giving it the ability to create a diverse array of de novo proteins. In the forward diffusion process (i.e. training), noise is added at each step. In the reverse diffusion process (i.e. generating), the model iteratively subtracts noise at each step, according to its learned distribution, in order to generate a plausible structure from the random noise.

As described in [30], one possible set of equations determining the learning and generating process is a Markov chain:

$$p(x_T|x_0) = \prod_{t=1}^{T} p(x_t|x_{t-1})$$

Where p is the distribution of noise and x_i is the ith sample. Thus the reverse process for generating new examples is:

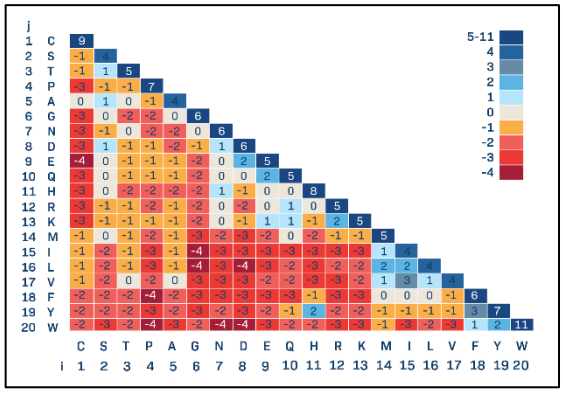
$$p_{\text{trained}}(x_0) = p(x_T) \prod_{t=1}^{T} p_{\text{trained}}\,(x_{t-1}|x_t)$$

Where p is now parameterised and trained. The model begins with pure noise (x_0) and iteratively denoises it over T steps.

EvoDiff model

EvoDiff is a very recent model (September 2023) developed by Microsoft researchers that incorporates evolutionary information to guide de novo protein design [31]. It incorporates Multiple Sequence Alignment (MSA) – which is a family of algorithms that aligns the sequences of proteins in the same family to identify areas that are frequently conserved or replaced [32].

The key features of this model are that it works on a discrete search space (amino acid sequences) rather than a continuous search space (rotamers and backbone angles), and that the corruption process it uses[1] is based on the frequency of amino acid mutations in evolution (rather than random noise sampled from a Gaussian distribution).



This is the BLOSUM62 matrix (taken directly from [33]) which was used in EvoDiff to create a transition matrix to base the mutation probabilities on. This matrix shows the likelihood of a base i mutating into another base j, based on observed frequencies. A higher number means substitution is more likely, while a lower (or negative) number means substitution is less likely. For example, serine (S) is more likely to change to threonine (T), with a score of +1, than to change into proline (P), with a score of -1.

[1] For simplicity of comparison I am only focusing on the discrete corruption process (D3PM), rather than the order-agnostic masked corruption process (OAPM) they also discuss in the paper

ii) Quantify and compare advantages and limitations of existing solutions using figures and tables (as applicable). If available, you may use data (e.g. on quantifying the performance of a given method) in the literature but you have to create your own informative (and carefully labelled) figures and tables (1 page, 10 marks)

Comparing diffusion models, VAEs and GANs

Due to the recency of EvoDiff and other models, there are limited data-based comparisons with other models (and, indeed, much of the recent literature seems to focus on *creating* models, with most data comparisons created prior to 2016), so I will focus on comparing the underlying models and algorithms.

I will compare three broad families of de novo protein design models: diffusion models, variational autoencoders (VAEs), and generative adversarial networks (GANs). VAEs work by encoding (and thus compressing) the input data in order to extract the key features [34]. GANs work by simultaneously training a generator model and a discriminator model: the former aims to create convincing 'fake' examples while the latter aims to distinguish between real and fake data, with the overall aim of improving the generator to the point where it generates convincingly real data [35].

Table comparing the advantages and limitations of de novo protein design models

Strengths/limitations	Diffusion model	VAE	GAN
Stability during training [36]	Stable	Stable	Unstable (subject to mode collapse)
Training data [37]	Need extensive datasets for training	Need less data	Need less data
Ease of training [38] [37]	Easy to train	Easy to train	Hard to train (hard to tell when the models have converged)
Ease of generating a sample (i.e. computational complexity) [36]	Difficult (the sequential steps outlined previously require multiple passes through the model)	Not difficult	Not difficult
Diversity of output [38]	Diverse	Diverse	Not diverse (the model is not incentivised to cover all data)
Fidelity of output [36-38]	High	Low (latent encodings can overlap leading to averaging)	High

Comparing RFdiffusion and EvoDiff

As suggested by the names, both RFdiffusion and EvoDiff use diffusion processes to generate proteins. As summarised above, diffusion-based models have the advantage of being flexible and able to accommodate high-dimensional data [39], at the cost of being computationally intensive [28]. Diffusion models have begun to outperform generative adversarial networks [40], and thus are currently one of the most successful models for de novo protein generation.

The key features of RFdiffusion [27] are that it can perform unconditional protein generation, symmetric unconditional generation (such as dihedral or cyclic symmetries) and conditional generation (such as designing binders or creating scaffolding to stabilise a known motif).

The key feature of EvoDiff [31] is that it can generate proteins that are 'inaccessible to structure-based models', such as proteins containing disordered regions, as summarised by its title 'sequence is all you need'. As a result of focusing on sequence alone, it uses a discrete diffusion framework whose search space is just amino acid residues (not rotamers or bond angles).

e) Provide critical assessment of all solutions you consider and discuss their potential (future) impact (1 page, 10 marks)

Critical assessment of diffusion models

Based on the comparison above, the (theoretically) best solution is a diffusion-based model. RFdiffusion, for example, achieved 'state of the art performance across 25 benchmark motif scaffolding problems'. However, diffusion models are computationally expensive: both to train, and (once trained) to subsequently *generate* de novo proteins. One problem that I propose might be encountered in the future is that of funding: high computational complexity means high computational resources, which in turn means a high cost of building the architecture, a high cost of sourcing hardware to run the software on, and a high cost of continuously running the software to generate outputs. This has the undesired effect that only financially-lucrative proteins are likely to be able to be generated using diffusion models (note, for example, that EvoDiff was uncoincidentally developed by Microsoft: one of the largest technology companies in the world). Thus the high computational complexity of these models means that 'pro bono' proteins (consider a hypothetical protein that reduces the side effects experienced during chemotherapy, but does not change the efficacy of the chemotherapy – this will likely be under-researched compared to a hypothetical alternative protein that has a reduced cost of production but increases side effects in patients).

Critical assessment of VAEs and GANs

Societal and ethical issues aside, if the computational complexity is considered a limiting factor, the next best option, in my opinion, is VAEs. This is because GANs are limited in diversity, and one of the reasons that de novo protein generation exploded in popularity is specifically because of its ability to create *novel* and diverse proteins [12, 41]. As noted in the EvoDiff article, one of the model's key abilities is to 'generate proteins inaccessible to structure-based models, such as those with disordered regions' [31], something that is inherently absent from GANs.

Potential future impacts of de novo protein design

In terms of positive impact, generative models (collectively) could enable breakthrough discoveries in medicine: in targeted vaccine development (by creating de novo ligands, e.g. [42]), in personalised medicine (such as for detecting predictive markers to estimate the likelihood of developing complications after contracting influenza [43]), in improving existing medicine (e.g. by increasing binding affinity or protein efficacy). As discussed at the beginning, many advances have already been made, and it is highly likely that we will continue to see advancements and discoveries made (both in healthcare and elsewhere, e.g. in nanotechnology) as computational efficiency and ability increases, and as de novo protein design continues to be studied.

f) Based on your understanding of the state-of-art in the field, propose two specific suggestions for future development to complement, replace or improve on existing solutions. Provide clear arguments to support your suggestions and describe what expected advances and limitations your proposed developments would bring when compared to existing methods available in the literature (1 page, 10 marks)

Expanding datasets

Currently, less than 2% of recorded DNA sequences have a corresponding structure identified in the Protein Databank (PDB) [44], since sequencing methods (like high-throughput sequencing) are much faster and more accessible than structure determination methods. This problem, called the protein sequence-structure gap [45], poses a particular problem for predicting the structure of larger or more dynamic proteins [46], and has been likened to 'dark matter' with the vast expanse of unexplored space [47].

Thus my suggestion for future development is to focus on expanding existing datasets of protein structure information with experimentally validated structures, in particular the structures of long, unstable or disordered proteins, as these are often underrepresented [48]. Current techniques include X-ray crystallography, nuclear magnetic resonance (NMR), and cryo-electron microscopy (cryo-EM) [49], but further research could be done to improve these techniques or develop new techniques in order to facilitate the rapid and cheap determination of structure.

Quantum computing

As outlined in part (d), diffusion models in particular are the most powerful and promising models at the moment, but are limited by computational complexity of training and generation. Thus any techniques that can improve computational speed will allow for rapid advancement of diffusion models.

Using quantum computing would allow more detailed searching of the sequence space. For example, if a very small energy 'funnel' existed (away from the main funnel), coarse-grained stochastic search methods may miss it, but fine-grained search methods with greater computational power may find it).

Using quantum computing to better search the so-called 'protein universe' [50] (i.e. the search space of protein structures) would also complement recent techniques aiming to characterise the protein universe, such as Google's research into annotating the protein universe using deep learning [51].

References

1. Khoury GA, Smadbeck J, Kieslich CA, Floudas CA. Protein folding and de novo protein design for biotechnological applications. Trends Biotechnol. 2014 Feb;32(2):99-109. PMID: 24268901. doi: 10.1016/j.tibtech.2013.10.008.
2. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021 2021/08/01;596(7873):583-9. doi: 10.1038/s41586-021-03819-2.
3. Bhardwaj G, Mulligan VK, Bahl CD, Gilmore JM, Harvey PJ, Cheneval O, et al. Accurate de novo design of hyperstable constrained peptides. Nature. 2016 2016/10/01;538(7625):329-35. doi: 10.1038/nature19791.
4. Coluzza I. Constrained versus unconstrained folding free-energy landscapes. Molecular Physics. 2015 2015/09/17;113(17-18):2905-12. doi: 10.1080/00268976.2015.1043031.
5. Samish I, MacDermaid CM, Perez-Aguilar JM, Saven JG. Theoretical and computational protein design. Annu Rev Phys Chem. 2011;62:129-49. PMID: 21128762. doi: 10.1146/annurev-physchem-032210-103509.
6. Muhammed MT, Aki-Yalcin E. Homology modeling in drug discovery: Overview, current applications, and future perspectives. Chem Biol Drug Des. 2019 Jan;93(1):12-20. PMID: 30187647. doi: 10.1111/cbdd.13388.
7. Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, et al. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res. 2018 Jul 2;46(W1):W296-w303. PMID: 29788355. doi: 10.1093/nar/gky427.
8. Brender JR, Shultis D, Khattak NA, Zhang Y. An Evolution-Based Approach to De Novo Protein Design. Methods Mol Biol. 2017;1529:243-64. PMID: 27914055. doi: 10.1007/978-1-4939-6637-0_12.
9. Yang Z, Kang S-g, Zhou R. Nanomedicine: de novo design of nanodrugs. Nanoscale. 2014;6(2):663-77. doi: 10.1039/C3NR04535H.
10. Suwinski P, Ong C, Ling MHT, Poh YM, Khan AM, Ong HS. Advancing Personalized Medicine Through the Application of Whole Exome Sequencing and Big Data Analytics. Frontiers in Genetics. 2019 2019-February-12;10. doi: 10.3389/fgene.2019.00049.
11. Woolfson DN. A Brief History of De Novo Protein Design: Minimal, Rational, and Computational. Journal of Molecular Biology. 2021 2021/10/01/;433(20):167160. doi: doi.org/10.1016/j.jmb.2021.167160.
12. Huang P-S, Boyken SE, Baker D. The coming of age of de novo protein design. Nature. 2016 2016/09/01;537(7620):320-7. doi: 10.1038/nature19946.
13. Rose GD, Fleming PJ, Banavar JR, Maritan A. A backbone-based theory of protein folding. Proceedings of the National Academy of Sciences. 2006;103(45):16623-33. doi: doi:10.1073/pnas.0606843103.
14. Zwanzig R, Szabo A, Bagchi B. Levinthal's paradox. Proc Natl Acad Sci U S A. 1992 Jan 1;89(1):20-2. PMID: 1729690. doi: 10.1073/pnas.89.1.20.
15. Newberry RW, Raines RT. Secondary Forces in Protein Folding. ACS Chem Biol. 2019 Aug 16;14(8):1677-86. PMID: 31243961. doi: 10.1021/acschembio.9b00339.
16. Bryson JW, Betz SF, Lu HS, Suich DJ, Zhou HX, O'Neil KT, et al. Protein Design: A Hierarchic Approach. Science. 1995;270(5238):935-41. doi: doi:10.1126/science.270.5238.935.
17. Ahring BK, Christiansen N, Mathrani I, Hendriksen HV, Macario AJ, Conway de Macario E. Introduction of a de novo bioremediation ability, aryl reductive dechlorination, into anaerobic granular sludge by inoculation of sludge with Desulfomonile tiedjei. Appl Environ Microbiol. 1992 Nov;58(11):3677-82. PMID: 1482188. doi: 10.1128/aem.58.11.3677-3682.1992.
18. Shimizu K, Mijiddorj B, Usami M, Mizoguchi I, Yoshida S, Akayama S, et al. De novo design of a nanopore for single-molecule detection that incorporates a β-hairpin peptide. Nature Nanotechnology. 2022 2022/01/01;17(1):67-75. doi: 10.1038/s41565-021-01008-w.
19. Zeitler B, Herrera Diaz A, Dangel A, Thellmann M, Meyer H, Sattler M, et al. De-Novo Design of Antimicrobial Peptides for Plant Protection. PLoS One. 2013;8(8):e71687. doi: 10.1371/journal.pone.0071687.
20. Yeh AH-W, Norn C, Kipnis Y, Tischer D, Pellock SJ, Evans D, et al. De novo design of luciferases using deep learning. Nature. 2023 2023/02/01;614(7949):774-80. doi: 10.1038/s41586-023-05696-3.

21. Quijano-Rubio A, Yeh H-W, Park J, Lee H, Langan RA, Boyken SE, et al. De novo design of modular and tunable protein biosensors. Nature. 2021 2021/03/01;591(7850):482-7. doi: 10.1038/s41586-021-03258-z.
22. Sesterhenn F, Yang C, Bonet J, Cramer JT, Wen X, Wang Y, et al. De novo protein design enables the precise induction of RSV-neutralizing antibodies. Science. 2020;368(6492):eaay5051. doi: doi:10.1126/science.aay5051.
23. Kuhlman B. Designing protein structures and complexes with the molecular modeling program Rosetta. J Biol Chem. 2019 Dec 13;294(50):19436-43. PMID: 31699898. doi: 10.1074/jbc.AW119.008144.
24. Leman JK, Weitzner BD, Lewis SM, Adolf-Bryfogle J, Alam N, Alford RF, et al. Macromolecular modeling and design in Rosetta: recent methods and frameworks. Nature Methods. 2020 2020/07/01;17(7):665-80. doi: 10.1038/s41592-020-0848-2.
25. Alford RF, Leaver-Fay A, Jeliazkov JR, O'Meara MJ, DiMaio FP, Park H, et al. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. J Chem Theory Comput. 2017 Jun 13;13(6):3031-48. PMID: 28430426. doi: 10.1021/acs.jctc.7b00125.
26. Rohl CA, Strauss CEM, Misura KMS, Baker D. Protein Structure Prediction Using Rosetta. Methods in Enzymology: Academic Press; 2004. p. 66-93.
27. Watson JL, Juergens D, Bennett NR, Trippe BL, Yim J, Eisenach HE, et al. De novo design of protein structure and function with RFdiffusion. Nature. 2023 2023/08/01;620(7976):1089-100. doi: 10.1038/s41586-023-06415-8.
28. Guo Z, Liu J, Wang Y, Chen M, Wang D, Xu D, et al. Diffusion models in bioinformatics and computational biology. Nature Reviews Bioengineering. 2023 2023/10/27. doi: 10.1038/s44222-023-00114-9.
29. Luo C. Understanding diffusion models: A unified perspective. arXiv preprint arXiv:220811970. 2022.
30. Chang Z, Koulieris GA, Shum HP. On the Design Fundamentals of Diffusion Models: A Survey. arXiv preprint arXiv:230604542. 2023.
31. Alamdari S, Thakkar N, Berg Rvd, Lu AX, Fusi N, Amini AP, et al. Protein generation with evolutionary diffusion: sequence is all you need. bioRxiv. 2023:2023.09.11.556673. doi: 10.1101/2023.09.11.556673.
32. Thompson JD, Linard B, Lecompte O, Poch O. A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives. PLoS One. 2011 Mar 31;6(3):e18093. PMID: 21483869. doi: 10.1371/journal.pone.0018093.
33. LabXchange. BLOSUM62 Substitution Matrix. 2021; Available from: https://www.labxchange.org/library/items/lb:LabXchange:24d0ec21:lx_image:1.
34. Pandi A, Adam D, Zare A, Trinh VT, Schaefer SL, Burt M, et al. Cell-free biosynthesis combined with deep learning accelerates de novo-development of antimicrobial peptides. Nature Communications. 2023 2023/11/08;14(1):7197. doi: 10.1038/s41467-023-42434-9.
35. Gupta A, Zou J. Feedback GAN for DNA optimizes protein functions. Nature Machine Intelligence. 2019 2019/02/01;1(2):105-11. doi: 10.1038/s42256-019-0017-4.
36. Dhaduk H. The Art and Science Behind Diffusion Models: How Businesses Can Benefit. SimForm; 2023.
37. Comparison between Diffusion Models vs GANs (Generative Adversarial Networks). MLK - Machine Learning Knowledge; 2023.
38. Gainetdinov A. Diffusion Models vs. GANs vs. VAEs: Comparison of Deep Generative Models. Towards AI; 2023.
39. Strokach A, Kim PM. Deep generative modeling for protein design. Current Opinion in Structural Biology. 2022 2022/02/01/;72:226-36. doi: doi.org/10.1016/j.sbi.2021.11.008.
40. Yang L, Zhang Z, Song Y, Hong S, Xu R, Zhao Y, et al. Diffusion Models: A Comprehensive Survey of Methods and Applications. ACM Comput Surv. 2023;56(4):Article 105. doi: 10.1145/3626235.
41. Yang C, Sesterhenn F, Bonet J, van Aalen EA, Scheller L, Abriata LA, et al. Bottom-up de novo design of functional proteins with complex structural features. Nature Chemical Biology. 2021 2021/04/01;17(4):492-500. doi: 10.1038/s41589-020-00699-x.
42. Costa CFS, Barbosa AJM, Dias AMGC, Roque ACA. Native, engineered and de novo designed ligands targeting the SARS-CoV-2 spike protein. Biotechnology Advances. 2022 2022/10/01/;59:107986. doi: doi.org/10.1016/j.biotechadv.2022.107986.

43. Valenzuela-Sánchez F, Valenzuela-Méndez B, Rodríguez-Gutiérrez JF, Rello J. Personalized medicine in severe influenza. Eur J Clin Microbiol Infect Dis. 2016 Jun;35(6):893-7. PMID: 26936615. doi: 10.1007/s10096-016-2611-2.
44. Sheehan D, O'Sullivan S. Online homology modeling as a means of bridging the sequence-structure gap. Bioengineered bugs. 2011 11/01;2:299-305. doi: 10.4161/bbug.2.6.16116.
45. Rost B, Sander C. Bridging the protein sequence-structure gap by structure predictions. Annu Rev Biophys Biomol Struct. 1996;25:113-36. PMID: 8800466. doi: 10.1146/annurev.bb.25.060196.000553.
46. Schwede T. Protein Modeling: What Happened to the “Protein Structure Gap”? Structure. 2013 2013/09/03/;21(9):1531-40. doi: doi.org/10.1016/j.str.2013.08.007.
47. Taylor WR, Chelliah V, Hollup SM, MacDonald JT, Jonassen I. Probing the "dark matter" of protein fold space. Structure. 2009 Sep 9;17(9):1244-52. PMID: 19748345. doi: 10.1016/j.str.2009.07.012.
48. Slabinski L, Jaroszewski L, Rodrigues AP, Rychlewski L, Wilson IA, Lesley SA, et al. The challenge of protein structure determination--lessons from structural genomics. Protein Sci. 2007 Nov;16(11):2472-82. PMID: 17962404. doi: 10.1110/ps.073037907.
49. Dokholyan NV. Experimentally-driven protein structure modeling. J Proteomics. 2020 May 30;220:103777. PMID: 32268219. doi: 10.1016/j.jprot.2020.103777.
50. Kolodny R, Pereyaslavets L, Samson AO, Levitt M. On the Universe of Protein Folds. Annual Review of Biophysics. 2013;42(1):559-82. PMID: 23527781. doi: 10.1146/annurev-biophys-083012-130432.
51. Bileschi ML, Belanger D, Bryant DH, Sanderson T, Carter B, Sculley D, et al. Using deep learning to annotate the protein universe. Nature Biotechnology. 2022 2022/06/01;40(6):932-7. doi: 10.1038/s41587-021-01179-w.