AI PROTEIN DESIGN & DRUG DISCOVERY

Abhishek Singh

ABOUT ME

My name is Abhishek Singh, and I hold a Master's degree in Machine Learning and Computer Vision from Queen Mary University of London. Currently, I am a Trainee in AI Drug Discovery at Hummingbird Bioscience, where I am exploring the application of AI-based computational methods to bioinformatics and protein design.

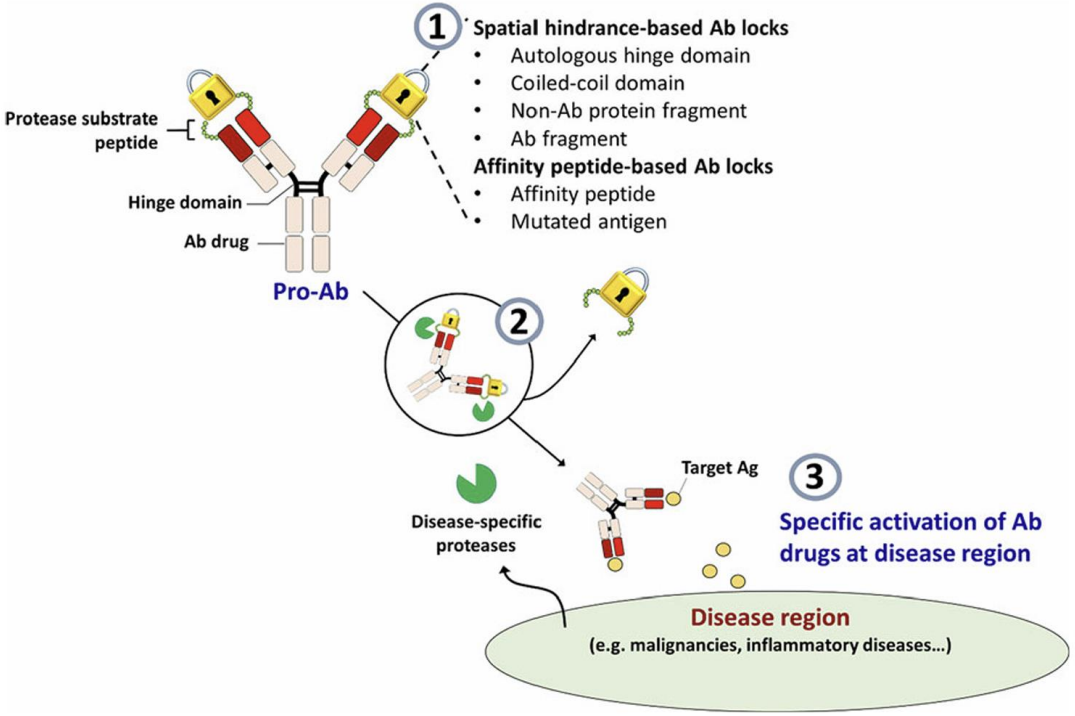
I bring a diverse background, with experience in domains such as the F&B industry, AI consultancy, and music informatics. While my journey in bioinformatics and drug discovery is just beginning, I am excited to leverage my expertise in machine learning, data analytics, and computer vision to contribute to advancements in this field.

RECENT WORK EXPERIENCE

At Hummingbird Bioscience, my work was focused on two key areas:

- Model exploration and set up different Machine Learning based models currently being developed for drug design and protein folding, which could potentially aid us in our antibody design and Binder/Mask Design projects.
- Secondly, after setting and deployment of these models in a computationally-efficient and user-friendly manner was to use apply these methods to generate de-novo peptide binders that would bind to our target of interests.

ANTIBODY LOCKS



Concept Introduction

Antibody locks are an innovative and effective approach designed to prevent/restrict the antibody activity until they are in close proximity to tumour cells, ensuring tumour-specific activation.

One common strategy is Spatial hindrance-based Ab locks, where a peptide lock is attached with a protease substrate linker to the antibody covering the antigen-binding sites.

The linker is cleavable in the presence of tumour-associated proteases to expose the binding site.

•**Mechanism**:

A common design incorporates a **helix coiled-coil structure**:

- **One helix** protrudes from the heavy chain.
- **Another helix** extends from the light chain.

These helices are connected via a **cleavable linker** sensitive to tumour-secreted peptidases.

PROBLEM STATEMENT

Hummingbird is currently advancing into nanobody development using phage display library to design different candidates from a stable clinically validated nanobody.

Unlike conventional antibodies, since nanobodies only have a single-chain, it makes the existing masking strategies, such as coiled-coil helix locks, developed for FAb regions, incompatible for them.

To tackle the above problem, we produced the strategy of Cleavable Linker-Based Locks attached to the N-terminus of our nanobody.

Since our candidates from phage library will have variability in the CDR loops, while retaining the framework region, we chose to target the constant framework regions with our masks to bring about conformational changes to the CDR loops while blocking them.

BINDER DESIGN PIPELINE

Backbone

RFDiffusion was used to generate mask designs targeting framework 4.

Sequence

ProteinMPNN trained on soluble proteins was used for inverse folding of RFDiff outputs.

Prediction Filtering

AF2 reduced MSA is used to predict the structures for the sequence output from MPNN.

The results are filtered based on RMSD and PAE Values

Liability Removal

Removal of designs with the potential risk of manufacturing liabilities.

Validation

MDsimulation of the shortlisted designs for final validation

RFDIFFUSION

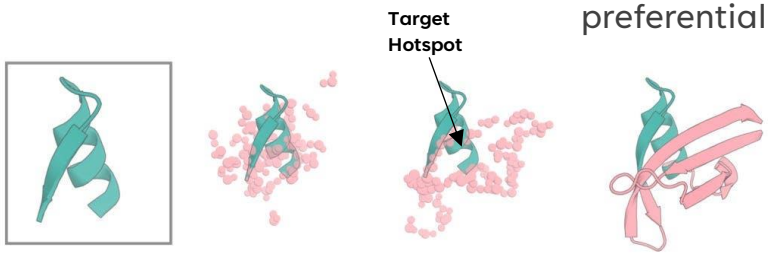
Approach

RFDiffusion is a powerful diffusion-based protein design model with various modes. In particular we experimented with the following modes:

1.Motif Scaffolding: This involves manipulating the contig map, which allows for precise control over the output by embedding specific structural motifs. It's highly effective for guiding protein design.

2.Fold Conditioning: This mode enables control over the structural conformation of the output by conditioning the model to generate proteins with desired folding patterns. When used alongside contig map manipulations, this is particularly useful for **Binder Design**—designing proteins or nanobodies with specific target-binding capabilities.

For our nanobody masking approach, we utilized **RFDiffusion's contig map** to generate peptides of varying lengths from the **N-terminus** of the nanobody. We instructed the model to ensure that the generated masks maintain continuous chains without introducing breaks. Additionally, we defined **target hotspots**, which act as guiding regions for the model, allowing the generated masks to preferentially bind to these target hotspots during the design process.



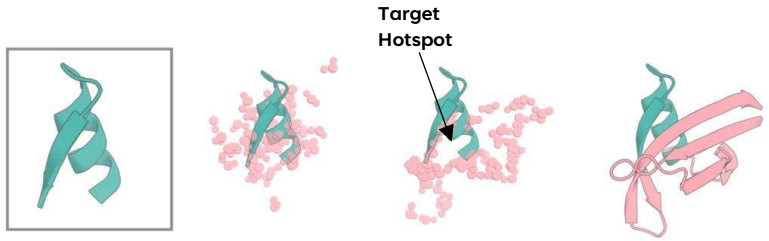
RFDIFFUSION

Approach

Since we plan to incorporate an unstructured cleavable linker connecting the mask and the nanobody's N-terminus, it is essential to identify the stable structural conformation of the **nanobody-linker complex** before proceeding with RFDiffusion generation. To achieve this, we opted to perform **molecular dynamics (MD) simulations** on the AF3 predicted nanobody-linker complex. This approach allows us to determine the most stable structural conformation of the linker, ensuring that it effectively guides the RFDiffusion model to generate masks targeting our designated hotspot and improves the confidence of the predicted models.

We took the MD simulated nanobody-linker complex as the input for RFDiffusion and generated >20,000 backbone designs of varying length between 20-80 residue length attached to the N-terminus of the nanobody-linker complex.

We used the base RFdiffusion model which often generates helical binders. These have high computational and experimental success rates when compared to the Beta model which gives other secondary structures in the outputs. From our observation, the Beta model outputs had low model confidence when compared to base model.



INVERSE FOLDING

Protein-MPNN

Since **RFDiffusion** generates only backbone structures as its outputs, an **inverse folding mechanism** is necessary to assign functional sequences to the designed binders. We selected **Soluble MPNN (Sol-MPNN)** for this purpose—a retrained version of Protein-MPNN optimized specifically for soluble proteins.

Sol-MPNN focuses on features such as hydrophilicity, solvent-accessible regions, charge distribution, and interactions with aqueous environments, ensuring that the designed proteins maintain high solubility upon expression. This minimizes aggregation risks and ensures functional stability in solution.

We passed all 20,000 RFDiffusion-generated designs through **Sol-MPNN**, and for each design, four sequences were generated at a sampling temperature of 0.1. To minimize risk, we adjusted the bias of the MPNN model to exclude **cysteines** from the predicted sequence outputs, as their presence increases the risk of disulfide bond formation and aggregation.

AF2 PREDICTION FILTERING

- We used reduced MSA Alphafold 2 algorithm by decreasing the MSA cluster size and generating structures with random seed. All the outputs generated by Sol-MPNN were passed through AF2. The num of recycles for binder design was set to 3, which gave slightly better predictions .

- The pipeline computes **RMSD** (Root Mean Square Deviation) to measure the structural deviation between the AF2-predicted models and the initial backbone structure generated by RFDiffusion and Predicted Aligned Error(PAE) which gives us the measure of how confident the AF2 is in it's predicted structures.
- We used the above two metrics to filter and shortlist our candidates:
  - We removed the designs with RMSD values > 5.
  - We removed the designs that had PAE values > 10.

  This filtering process ensured that only the most reliable and accurate designs were considered for further analysis.

LIABILITY REMOVAL

Objective

Minimize structural and functional risks in nanobody mask designs by addressing common liabilities that affect stability, expression, and performance.

Key Liabilities Addressed:

1.Consecutive Amino Acids: Removed sequences with 3+ identical residues (e.g., AAA, CCC) to prevent aggregation.

2.Excessive Unknown Residues (X): Limited sequences with high X content to maintain predictability and stability.

3.Post-Translational Modification (PTM) Motifs: Eliminated common motifs like NXS, NXT etc. linked to glycosylation and other undesired PTMs.

4.Isolated Cysteines: Removed isolated cysteines to avoid unintentional disulfide bonding that could destabilize structure.

The designs after the liability derisking steps were considered to be fit for manual inspection and MD Simulation.

MOLECULAR DYNAMICS SIMULATION

GROMACS

- As a final step in validation and filtering, we utilized MD simulations with GROMACS to evaluate the stability and dynamic behaviour of our nanobody designs.
- Unlike AlphaFold 2, which is influenced by its training data, GROMACS employs a physics-based force field, offering an unbiased assessment of structural stability. By simulating the designs under realistic conditions, we observed their dynamic behaviour over time, focusing specifically on potential changes in the conformations of the CDR loops, which are critical for binding functionality.
- This approach provided quantitative insights into any structural instabilities, guiding further refinement and shortlisting of the designs.

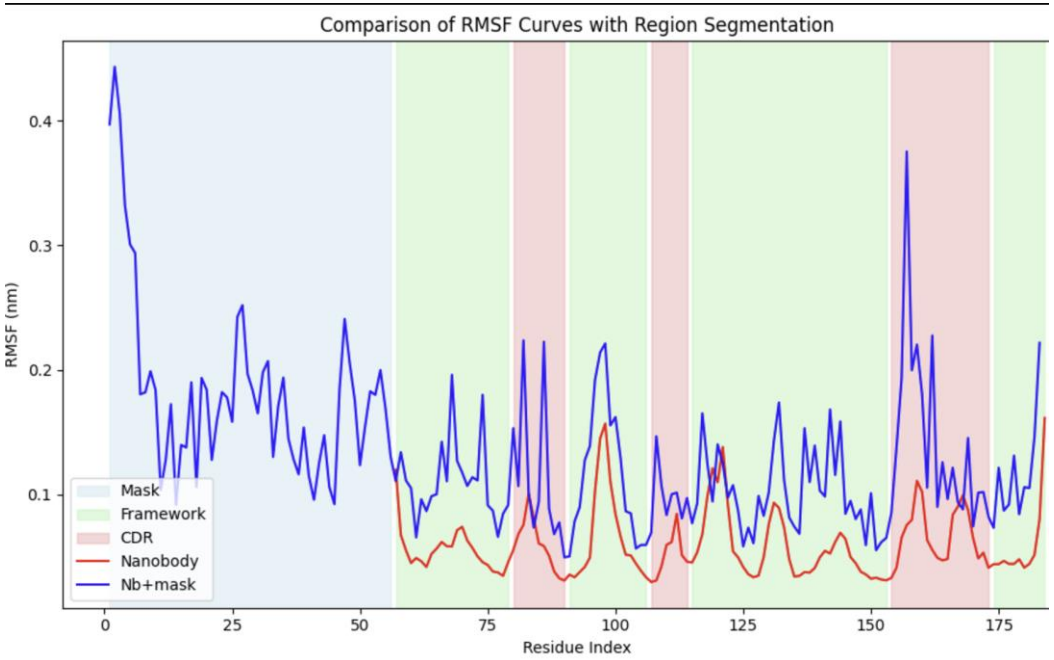
MD SIMULATION ANALYSIS

The simulations for all the designs were ran for 50 ns (50,000 steps), with Charmm36 Forcefield, the final trajectory was converted into a single pdb (1000 models) and was then analyzed and compared against the MD Simulation of the **nanobody** itself in the same conditions.

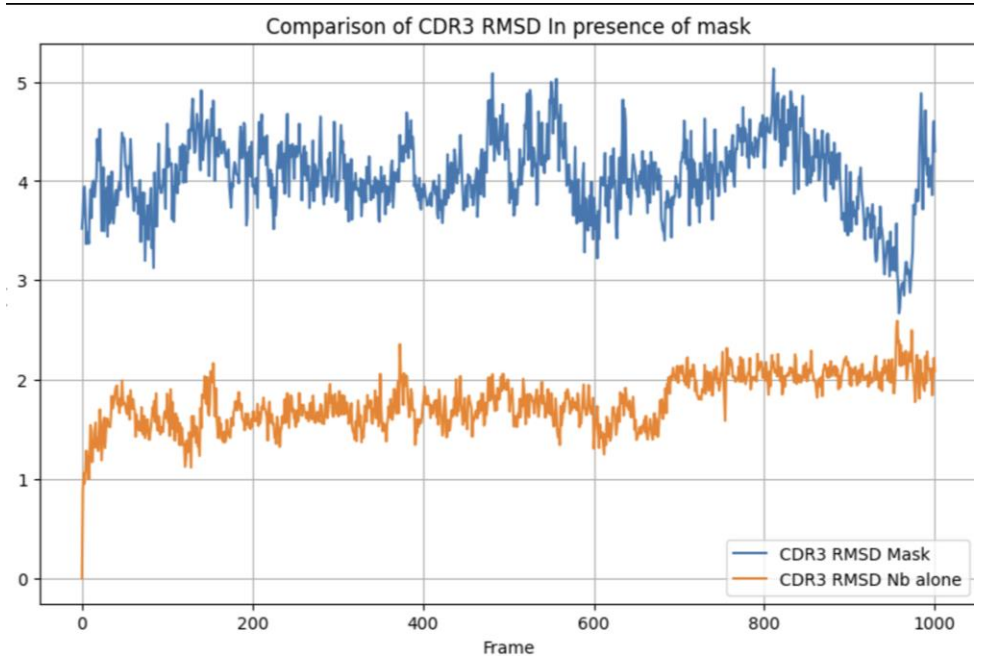
The following were considered for shortlisting:

- The designed mask throughout the trajectory had distance less than 5 Å from the framework 4
- The simulated protein had achieved a stabilized radius of gyration
- The simulated protein had high Root Mean Square Fluctuations in the CDR Regions
- Throughout the trajectory, the RMSD of the simulated protein's CDR3, when compared to the crystal structure's CDR3, was higher than the RMSD of the nanobody's CDR3 relative to the crystal structure. This indicates greater structural deviation in the simulated protein's CDR3 compared to the nanobody's CDR3.

MD SIMULATION ANALYSIS



- We can observe the above RMSF curve that the residues corresponding to the CDR regions undergo conformational changes under the presence of the mask.



- From the above plot we can infer that due to the presence of the mask, the CDR 3 RMSD is significantly higher when compared against the CDR 3 RMSD of nanobody for the entire trajectory.
- We also observed the trajectory around the frame 950-1000 mark and inferred that around the frame 980 the mask had no contacts with the framework 4. Hence, we notice the drop in the CDR 3 RMSD.

SUMMARY

To summarize, The pipeline developed by us was used for de-novo binder design to our choice of hotspots in Framework 4, covering CDR3. We generated more than 20,000 designs from RFDiffusion, then using Sol-MPNN generated 4 sequence for each design (80,000 files) we then use AF2 Reduced MSA for structure prediction of the generated sequences and then filter the results based on their RMSD values and PAE.

It was highly important for us to remove designs which could potentially cause aggregations, or could yield low expression, so we removed all the potential liability inducing designs.

We shortlisted, 30 designs out of all the samples to proceed with MD Simulations, then selected 10 best designs for experimental testing which had significantly changed the CDR conformations and the masks were less than 5 Å to the framework region.

THANK YOU

Abhishek Singh

+65 85026836

abhisteak@gmail.com