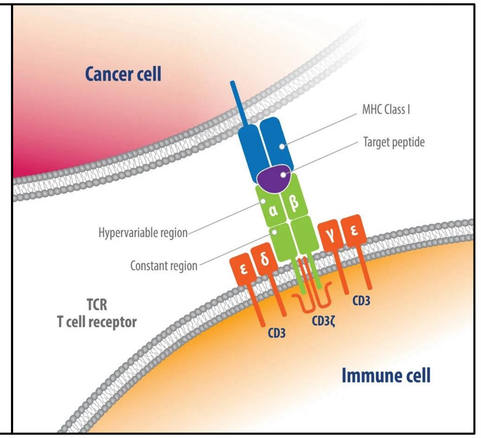
Fine-Tuning RFdiffusion for Structure-Based Design of Epitope-Specific T Cell Receptors

Anuj Prabhu, Computer Science, B.S.
Mentor: Dr. Heewook Lee, Assistant Professor
School of Computing and Augmented Intelligence (SCAI); Biodesign Center for Biocomputing, Security and Society (BSS)

Research Question:
This research investigates whether fine-tuning the RFdiffusion model on experimentally validated TCR-pMHC complex data can enable accurate, structure-conditioned generation of TCR CDR3β backbones with high binding complementarity to target epitopes.

Background:
- T cell receptors (TCRs) recognize small antigen peptides presented by MHC molecules on target cells.
- Binding specificity is driven by the α and β chains, especially the CDR3β region, which contacts the peptide directly.
- Proper recognition activates the immune system to destroy abnormal or infected cells.
- De novo TCR design could transform immunotherapy.
- Existing models lack 3D structural awareness, unlike RFdiffusion (Baker Lab, 2023), which generates protein backbones using rotationally equivariant diffusion.



Data Collection and Filtering:
- **Source:** Experimentally solved TCR-pMHC complexes from the TCR3d database (~300 observations).
- **Fields:** ***{Peptide, TRBV, TRBJ, CDR3β, PDB ID}***
- **Annotation:** V/J gene calls and numbering via ANARCI.
- **Processing:** TCRDock + AlphaFold refinement
- **Filtering:**
  - Cα RMSD ≤ 2 Å vs experimental
  - Human only (restrict to HLA-A*02:01 allele)
  - Binders (positive pairs) only
- **Output:** Extract Cα coords + rigid frames for pMHC (condition) and TCR/CDR3β (target) → inputs for RFdiffusion fine-tuning.

TCR-pMHC complexes

↓

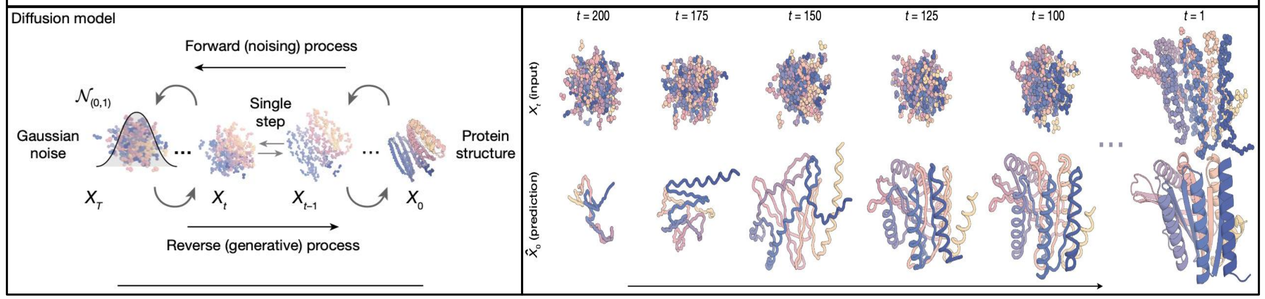
AlphaFold + TCRDock pipeline

↓

Cα coords + 3x3 rigid frames

LLFGYPVYV **(peptide)**	TRBV6-5 **(TRBV)**	TRBJ2-1 **(TRBJ)**	CASRPGLAGGRPEQYF **(CDR3β)**	1ao7 **(PDB ID)**

Model Framework and Training Plan:
- Fine-tuning RFdiffusion, a denoising diffusion probabilistic model that generates 3D protein backbones by iteratively denoising residue coordinates and orientations.
- Each residue represented as a rigid frame (Cα coordinate + 3×3 rotation matrix).
- Conditioned on epitope (pMHC) structure and interface hotspot residues to guide TCR-epitope complementarity.
- Trained to minimize MSE loss between predicted and native residue frames, learning to reverse the noise process.
- Uses self-conditioning to stabilize denoising and improve sample consistency.
- Generated backbones passed to ProteinMPNN for sequence design compatible with predicted structures.
- Evaluated using Cα RMSD, interface TM-score, and AlphaFold/TCRDock revalidation for structural plausibility.



Future Work:
- Implement and fine-tune the proposed epitope-conditioned RFdiffusion model using curated TCR pMHC structural data.

References:
- Watson, J.L., Juergens, D., Bennett, N.R. *et al.* De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100 (2023). https://doi.org/10.1038/s41586-023-06415-8
- J. Dauparas *et al.,* Robust deep learning-based protein sequence design using ProteinMPNN. *Science* **378**, 49-56 (2022). DOI: 10.1126/science.add2187
- https://tcr3d.ibbr.umd.edu/