# Designing Better Cancer Vaccines

UCL-CCC-Hackathon-2025
Team TC-AWARE

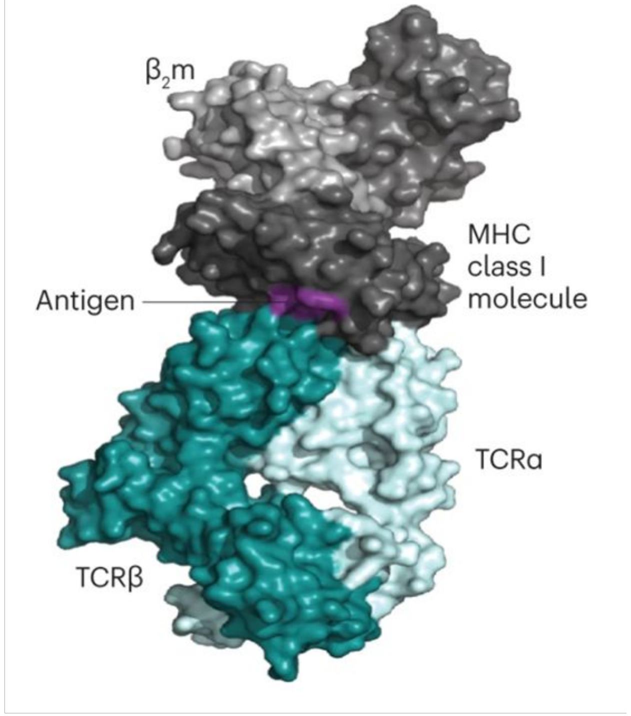
Matthew Cowley, Zhen Wei Yap, Mohammad Alawwami, Zarif Shafiei, Linh Hoang, Julia Sala-Bayo, Graham Bonomo-Jackson, Gleb Gmyzov, Nick Keatley

## Problem statement

Cancer vaccines offer an amazing opportunity to develop personalised cures for patients. Our immune system is well equipped to kill cancer but it needs help finding the right target. The most important step is identifying the right target.

- Hard to choose targets as the rough search space of cancer peptides is in excess of $10^{17}$

We have formulated an end-to-end pipeline that identifies and ranks potential targets (neoantigens)



Hudson et al. 2023 *Nat. Rev. Imm.*

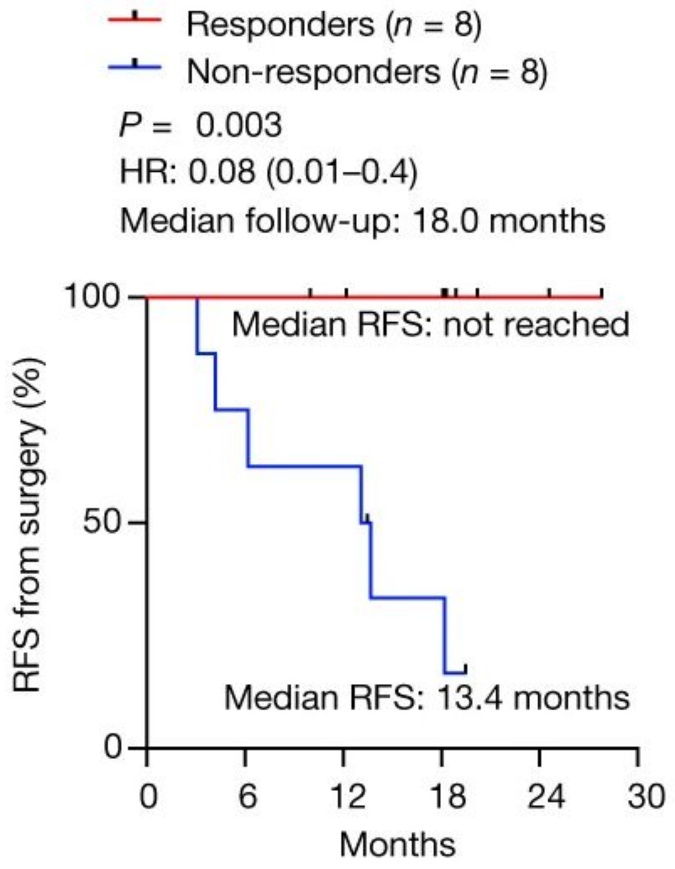
## How our proposal can help

Cancer vaccines can be highly effective, but there is a high rate of non-responders.

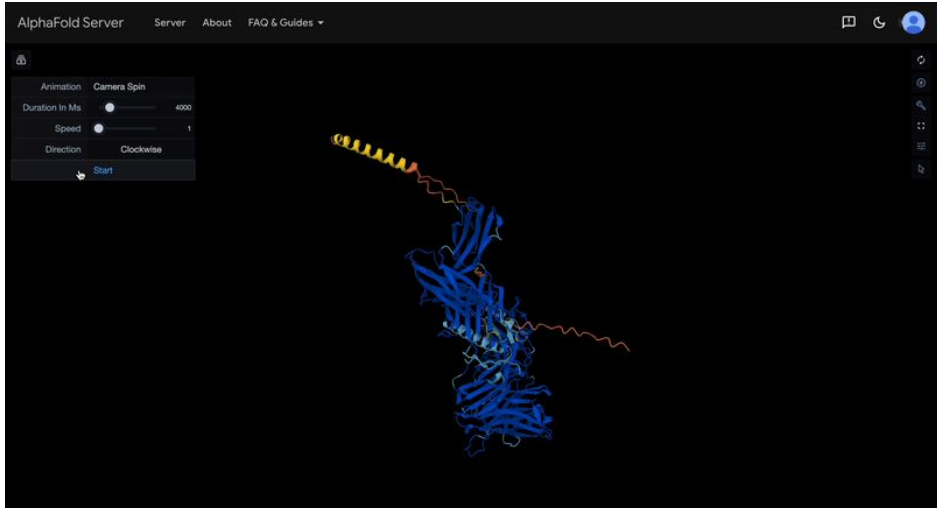
We propose to improve vaccine design by using patient immune receptors to select peptides which will drive a strong vaccine response.

Solution could reduce manufacturing time and costs ($100,000 per patient currently)

- Enables re-distribution of funds to better improve patient care



# Project Overview



Animation Camera Spin
Duration In Ms 4000
Speed 1
Direction Clockwise

| | |
|---|---|
| **Goal** | Find new weak spots on cancer cells that the patient's immune system hasn't noticed yet |
| **Output** | A prioritised list of cancer targets to put in a vaccine, but specifically excluding targets the immune system has already tried and failed to attack |
| **Key Innovation** | Personalised cancer vaccines that target the cancer's fundamental weak spots, specifically choosing targets the patient's immune system hasn't already failed to attack. |

## Data Requirements

**Training Data (model development)**

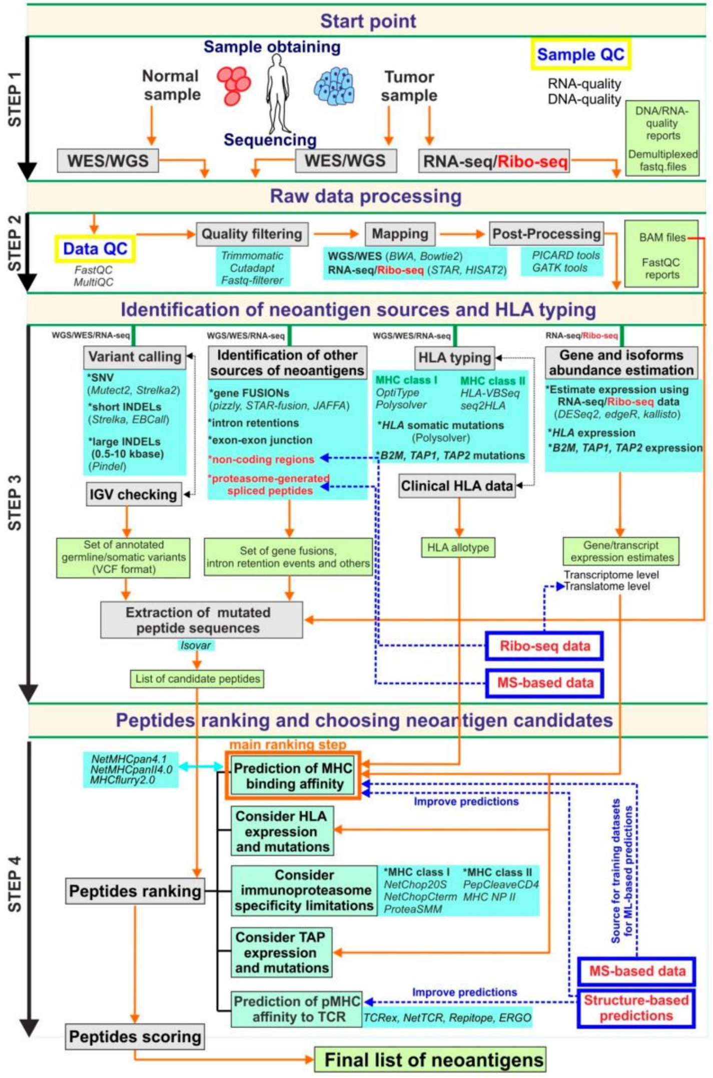
| Data Type | Source | Used For |
|---|---|---|
| Cancer genomes with validated neoantigens | Cancer genome studies | Neoantigen identification |
| HLA-peptide binding datasets (multi-individual) | IEDB, published datasets | MHC Find / binding prediction |
| TCR-peptide binding data | Viral/autoimmune contexts, structural DBs | TCR binding prediction |

## Data Requirements

**Patient-Specific Data (per-patient input)**

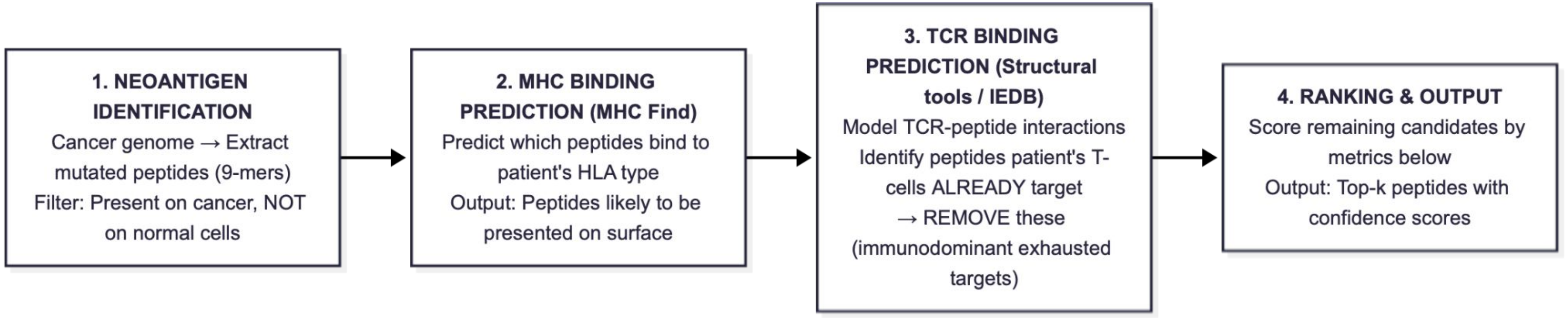
| **Data Type** | **Source** |
|---|---|
| Cancer genome | Tumour sample (NGS) |
| TCR repertoire | Blood sample sequencing |
| HLA/MHC type | Patient genotyping |

# In silico neoantigen prediction

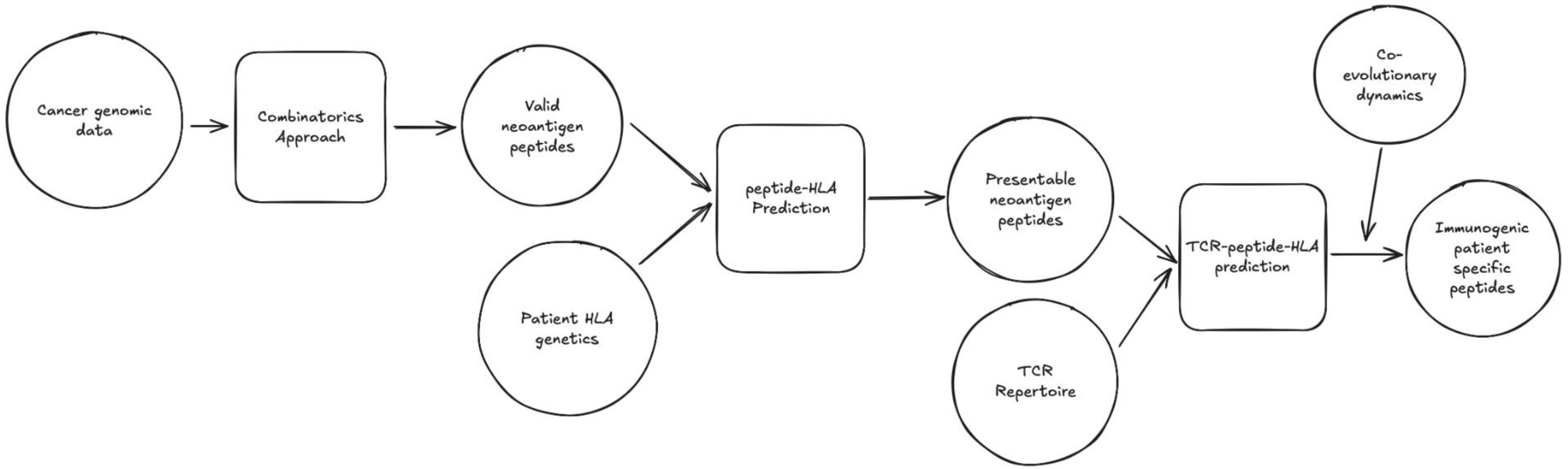


Start point
STEP 1
Sample obtaining
Sample QC
Normal sample
Tumor sample
RNA-quality
DNA-quality
Sequencing
DNA/RNA-quality reports
Demultiplexed fastq.files
WES/WGS
WES/WGS
RNA-seq/Ribo-seq
Raw data processing
STEP 2
Data QC
FastQC
MultiQC
Quality filtering
Trimmomatic
Cutadapt
Fastq-filterer
Mapping
WGS/WES (BWA, Bowtie2)
RNA-seq/Ribo-seq (STAR, HISAT2)
Post-Processing
PICARD tools
GATK tools
BAM files
FastQC reports
Identification of neoantigen sources and HLA typing
STEP 3
WGS/WES/RNA-seq
Variant calling
*SNV (Mutect2, Strelka2)
*short INDELs (Strelka, EBCall)
*large INDELs (0.5-10 kbase) (Pindel)
IGV checking
Set of annotated germline/somatic variants (VCF format)
WGS/WES/RNA-seq
Identification of other sources of neoantigens
*gene FUSIONs (pizzly, STAR-fusion, JAFFA)
*intron retentions
*exon-exon junction
*non-coding regions
*proteasome-generated spliced peptides
Set of gene fusions, intron retention events and others
WGS/WES/RNA-seq
HLA typing
MHC class I OptiType Polysolver
MHC class II HLA-VBSeq seq2HLA
*HLA somatic mutations (Polysolver)
*B2M, TAP1, TAP2 mutations
Clinical HLA data
HLA allotype
RNA-seq/Ribo-seq
Gene and isoforms abundance estimation
*Estimate expression using RNA-seq/Ribo-seq data (DESeq2, edgeR, kallisto)
*HLA expression
*B2M, TAP1, TAP2 expression
Gene/transcript expression estimates
Transcriptome level
Translatome level
Extraction of mutated peptide sequences
Isovar
List of candidate peptides
Ribo-seq data
MS-based data
Peptides ranking and choosing neoantigen candidates
STEP 4
main ranking step
NetMHCpan4.1
NetMHCpanII4.0
MHCflurry2.0
Prediction of MHC binding affinity
Improve predictions
Consider HLA expression and mutations
Peptides ranking
Consider immunoproteasome specificity limitations
*MHC class I NetChop20S NetChopCterm ProteaSMM
*MHC class II PepCleaveCD4 MHC NP II
Consider TAP expression and mutations
Source for training datasets for ML-based predictions
MS-based data
Improve predictions
Prediction of pMHC affinity to TCR
TCRex, NetTCR, Repitope, ERGO
Structure-based predictions
Peptides scoring
Final list of neoantigens

## High Level Processing/Development Pipeline


**1. NEOANTIGEN IDENTIFICATION**
Cancer genome → Extract mutated peptides (9-mers)
Filter: Present on cancer, NOT on normal cells
**2. MHC BINDING PREDICTION (MHC Find)**
Predict which peptides bind to patient's HLA type
Output: Peptides likely to be presented on surface
**3. TCR BINDING PREDICTION (Structural tools / IEDB)**
Model TCR-peptide interactions
Identify peptides patient's T-cells ALREADY target
→ REMOVE these (immunodominant exhausted targets)
**4. RANKING & OUTPUT**
Score remaining candidates by metrics below
Output: Top-k peptides with confidence scores

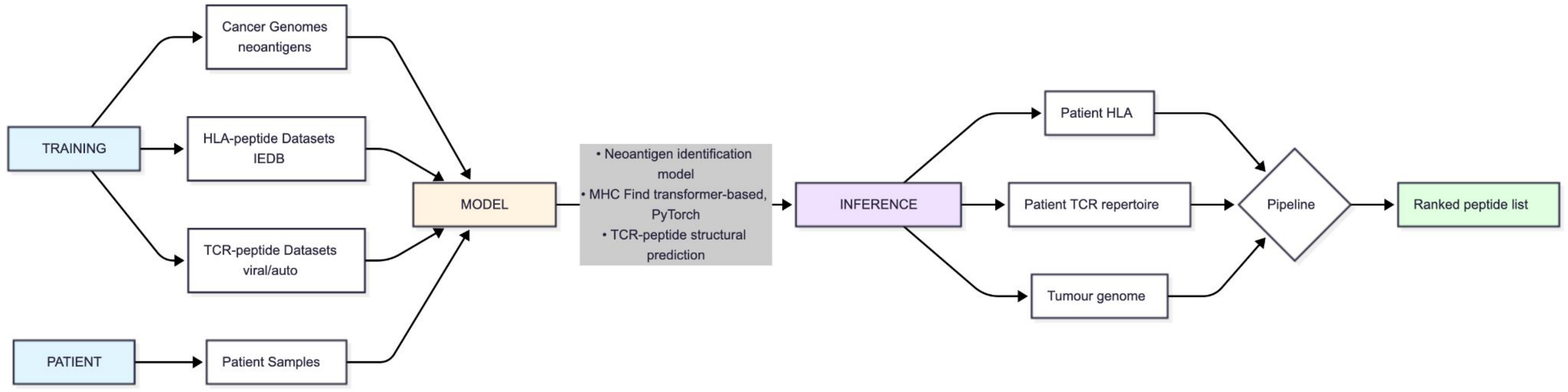
# Clinical Development Pipeline



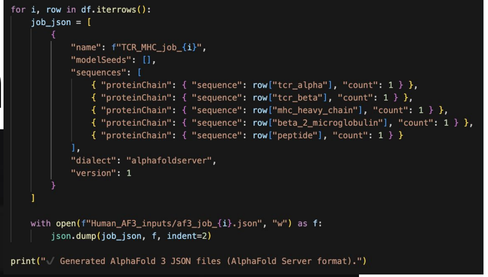
## Ranking Metrics

| Metric | Description | Why It Matters |
|---|---|---|
| **MHC-peptide binding affinity** | How strongly peptide binds to patient's HLA | Must be presented to be targeted |
| **TCR-peptide binding affinity** | Predicted interaction strength with naïve T-cells | Stronger binding → better response |
| **Surface presentation likelihood** | Probability of actual presentation | Processing/transport affects this |
| **Peptide concentration** | Abundance on cell surface | Higher = stronger immune signal |
| **Conservation (co-evolution)** | How essential/conserved across cancer subtypes | Harder for cancer to escape via mutation |
| **Clonality (phylogenetic position)** | Present in founder clone vs subclone | Truncal mutations = all cells targeted |
| **Specificity (cross-reactivity)** | Risk of targeting normal tissue | Safety — avoid autoimmunity |

# Consolidated Architecture Diagram


TRAINING
Cancer Genomes neoantigens
HLA-peptide Datasets IEDB
TCR-peptide Datasets viral/auto
PATIENT
Patient Samples
MODEL
• Neoantigen identification model
• MHC Find transformer-based, PyTorch
• TCR-peptide structural prediction
INFERENCE
Patient HLA
Patient TCR repertoire
Tumour genome
Pipeline
Ranked peptide list

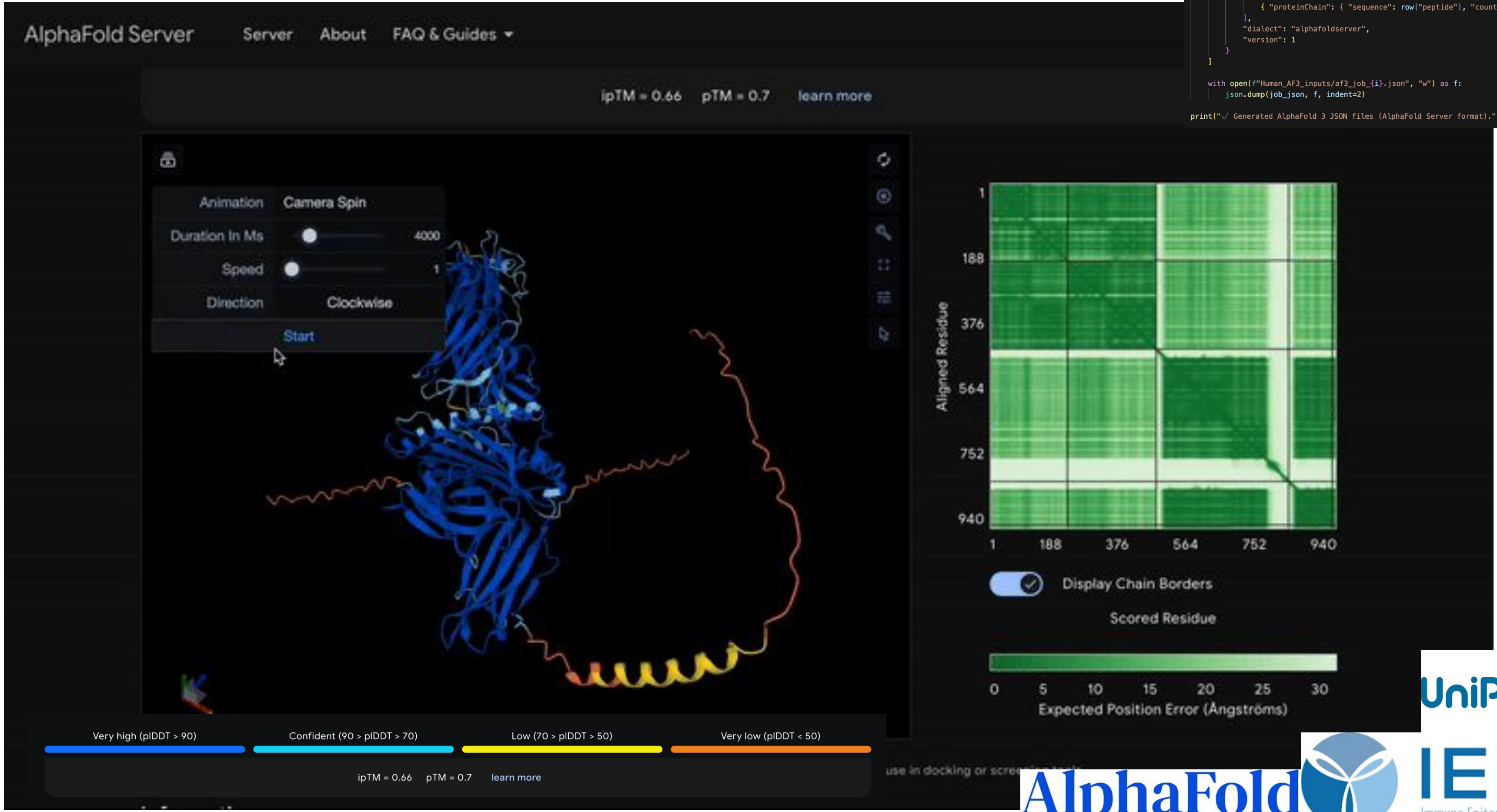
# PoC: Deep Learning Structure Prediction



```
for i, row in df.iterrows():
    job_json = [
        {
            "name": f"TCR_MHC_job_{i}",
            "modelSeeds": [],
            "sequences": [
                { "proteinChain": { "sequence": row["tcr_alpha"], "count": 1 } },
                { "proteinChain": { "sequence": row["tcr_beta"], "count": 1 } },
                { "proteinChain": { "sequence": row["mhc_heavy_chain"], "count": 1 } },
                { "proteinChain": { "sequence": row["beta_2_microglobulin"], "count": 1 } },
                { "proteinChain": { "sequence": row["peptide"], "count": 1 } }
            ],
            "dialect": "alphafoldserver",
            "version": 1
        }
    ]

    with open(f"Human_AF3_inputs/af3_job_{i}.json", "w") as f:
        json.dump(job_json, f, indent=2)

print("✓ Generated AlphaFold 3 JSON files (AlphaFold Server format).")
```



ipTM = 0.66 pTM = 0.7

| | |
|---|---|
| Animation | Camera Spin |
| Duration In Ms | 4000 |
| Speed | 1 |
| Direction | Clockwise |

Very high (pLDDT > 90) | Confident (90 > pLDDT > 70) | Low (70 > pLDDT > 50) | Very low (pLDDT < 50)

ipTM = 0.66 pTM = 0.7

## Thank you