Physicochemically Informed Dual-Conditioned Generative Model of T-Cell Receptor Variable Regions for Cellular Therapy

Jiahao Ma[1,2,6], Hongzong Li[3], Ye-Fan Hu[4,5,*], Jian-Dong Huang[1,2,6,*]

[1]School of Biomedical Sciences, The University of Hong Kong, Hong Kong SAR, China
[2]Materials Innovation Institute for Life Sciences and Energy (MILES), HKU, Hong Kong SAR, China
[3]HK Generative AI Research and Development Center, HKUST, Hong Kong SAR, China
[4]Computational Immunology Centre, BayVax Biotech Limited, Hong Kong SAR, China
[5]Shenzhen BayVax Biotech Limited, Shenzhen, China
[6]Shenzhen Institute of Research and Innovation (HKU-SIRI), Shenzhen, China
*Corresponding authors: yefan.hu@bayvaxbio.com, jdhuang@hku.hk

***Abstract*—Physicochemically informed biological sequence generation has the potential to accelerate computer-aided cellular therapy, yet current models fail to *jointly* ensure novelty, diversity, and biophysical plausibility when designing variable regions of T-cell receptors (TCRs). We present PhysicoGPTCR, a large generative protein Transformer that is *dual-conditioned* on peptide and HLA context and trained to autoregressively synthesise TCR sequences while embedding residue-level physicochemical descriptors. The model is optimised on curated TCR–peptide–HLA triples with a maximum-likelihood objective and compared against ANN, GPTCR, LSTM, and VAE baselines. Across multiple neoantigen benchmarks, PhysicoGPTCR substantially improves edit-distance, similarity, and longest-common-subsequence scores, while populating a broader region of sequence space. Blind *in-silico* docking and structural modelling further reveal a higher proportion of binding-competent clones than the strongest baseline, validating the benefit of explicit context conditioning and physicochemical awareness. Experimental results demonstrate that dual-conditioned, physics-grounded generative modelling enables end-to-end design of functional TCR candidates, reducing the discovery timeline from months to minutes without sacrificing wet-lab verifiability.**

***Index Terms*—Protein Sequence Generation, Generative Transformer, Physicochemical Constraints, Dual-Conditioned Generative Model**

I. Introduction

The clinical promise of T-cell-receptor–engineered therapy (TCR-T) [1] rests on the rapid discovery of variable regions that recognise patient-specific peptide–HLA complexes with high affinity and selectivity. Classical wet-lab screening cycles require months of iterative cloning and probe only an infinitesimal fraction of the astronomical search space, with 10^{15}–10^{61} possible TCR sequences by distinct estimations [2].

Recent protein language models have begun to reshape macromolecular design: Transformer-based generators [3], [4] can now hallucinate enzyme folds and antibodies *in silico*. Yet no prior study has shown that such models can *directly* create biophysically feasible *TCR* sequences that remain usable in downstream TCR-T pipelines. Compared with antibody or minibinder design (Fig. 1A), TCR generation is harder: binding specificity is jointly determined by the presented peptide *and* the polymorphic HLA molecule, and subtle long-range physicochemical couplings often decide success or failure.

Two technical gaps persist. (1) Vanilla autoregressive models tend to overlook non-local chemical interactions, causing mode collapse or implausible motifs. (2) Models that pursue smoother sequence manifolds seldom encode immunological priors and therefore struggle to enrich for true neoantigen specificity. These issues call for an architecture that *explicitly embeds physicochemical knowledge while remaining end-to-end trainable*.

We respond to this need with **PhysicoGPTCR**, a dual-conditioned generative Transformer that takes the context of the peptide and HLA as input and autoregressively synthesizes TCR variable-region sequences (Fig. 1B). The encoder and decoder fuse three information channels—token identity, positional index, and residue-level physicochemical descriptors (aromaticity, charge, hydrogen-bond capacity, molecular mass)—through a learnable gating mechanism, enabling the network to reason about long-range chemical bonds during generation. Training is performed on curated TCR–peptide–HLA triples spanning tumour, autoimmune, viral and bacterial antigens drawn from VDJdb, with maximum-likelihood optimisation only; no post-hoc filters are required.

We benchmark PhysicoGPTCR against four competitive baselines (ANN retrieval, GPTCR, LSTM, VAE) on multiple neoantigen test sets. Across edit-distance, sequence-similarity and longest-common-subsequence metrics, our model *substantially* outperforms all alternatives while populating a broader region of sequence space. A dry-lab activation assay based on blind *in-silico* docking further confirms a higher proportion of binding-competent clones, validating the benefit of explicit physicochemical embeddings.

Our contributions are threefold:

- **Method**: we couple dual biological conditioning with

978-1-XXXX-XXXX-X/26/$31.00 © 2026 IEEE

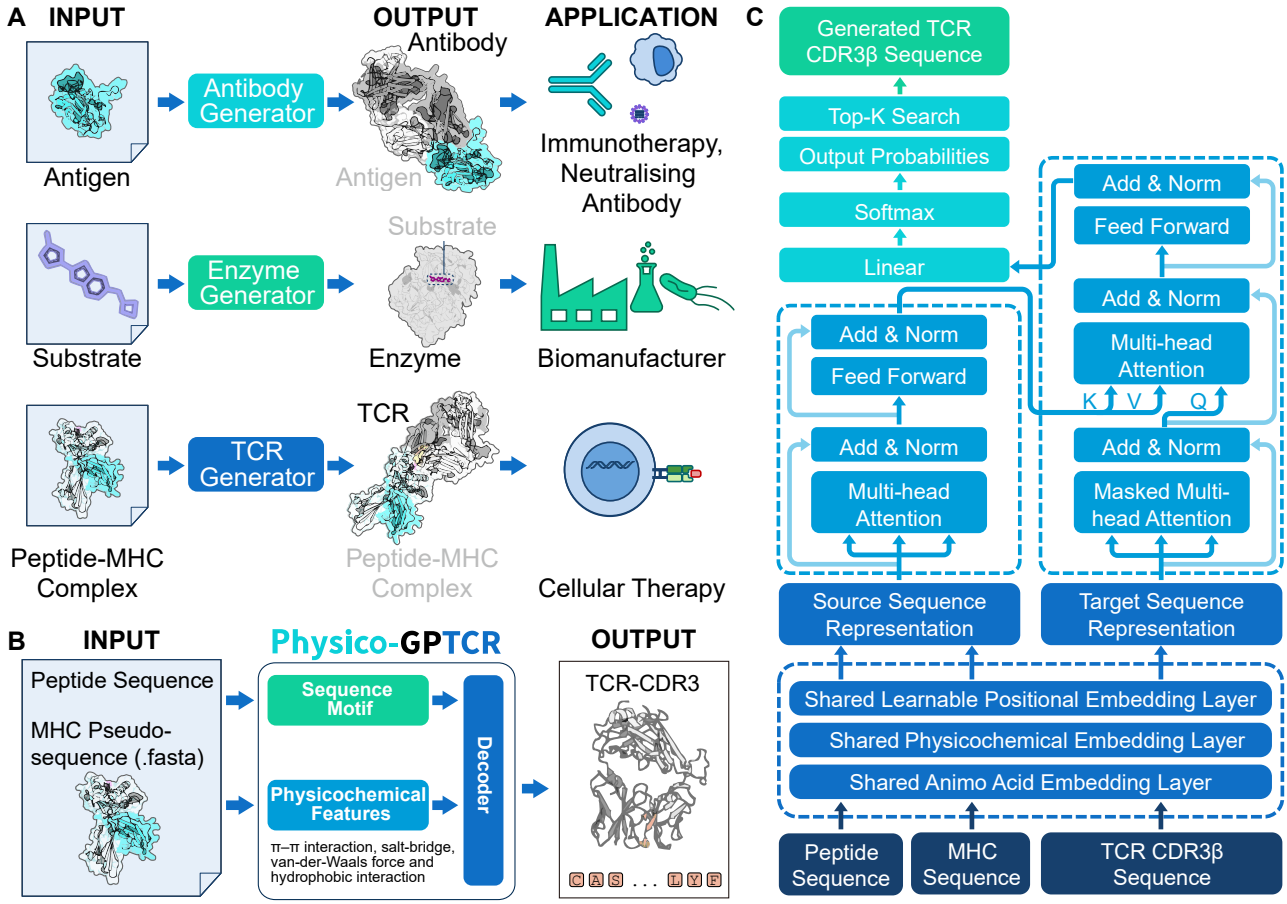

Fig. 1: Tasks similar to TCR generation and the workflow. **(A)** Protein generation analogies. Antibodies can be generated based on antigen inputs, applied to immunotherapy or neutralizing antibodies. Enzymes can be generated for distinct substrates to improve bio-manufacturer. TCR generation is similar to previous two tasks. **(B)** Model inputs and outputs: By requiring peptide-MHC inputs, TCR can be generated for cellular therapies. **(C)** Model overview. Three information channels—token identity, positional index and residue-level physicochemical descriptors—are fused by a gated projector and fed into a lightweight 2+2-layer Transformer that is conditioned on both peptide and HLA context. The model processes peptide sequences and MHC pseudo-sequences as inputs, leveraging sequence motifs and physiochemical features through PhysicoGPTCR, followed by a decoder that outputs **TCR CDR3** sequences.

residue-aware physicochemical embeddings, unifying language modelling and chemical-bond reasoning in a single Transformer.

- **Performance**: the approach delivers state-of-the-art generative quality across all string-based metrics and markedly enriches functional hits in dry-lab assays.
- **Impact**: minute-scale inference shortens the TCR-discovery timeline from months to minutes, offering an immediately deployable tool for precision immunotherapy.

Open Science: Our data and code are publicly available at https://github.com/Ascaris-Equi/PhsicoGPTCR to support reproducibility and replicability. The inference script, model weights, and post-processing workflow examples are included in the repository.

II. Related Work

A. HLA–peptide specificity prediction

Early studies cast T-cell recognition as a *discriminative* task. NetTCR-2.0, DeepTCR and TITAN [5]–[7] use convolutional, recurrent or attention networks to decide whether a query receptor recognises a given HLA–peptide complex. Although AUC scores keep improving, these classifiers are inherently non-generative and cannot propose novel receptors for TCR-T therapy; moreover, they view sequences as mere symbol strings and ignore the residue–residue couplings that ultimately drive binding.

B. Generative modelling of TCRs

Only a handful of attempts move beyond classification. TESSAR [8] explores unsupervised reconstruction but focuses on receptor repertoires without conditioning on antigen context. More recently, TCRGPT [9] autoregressively samples

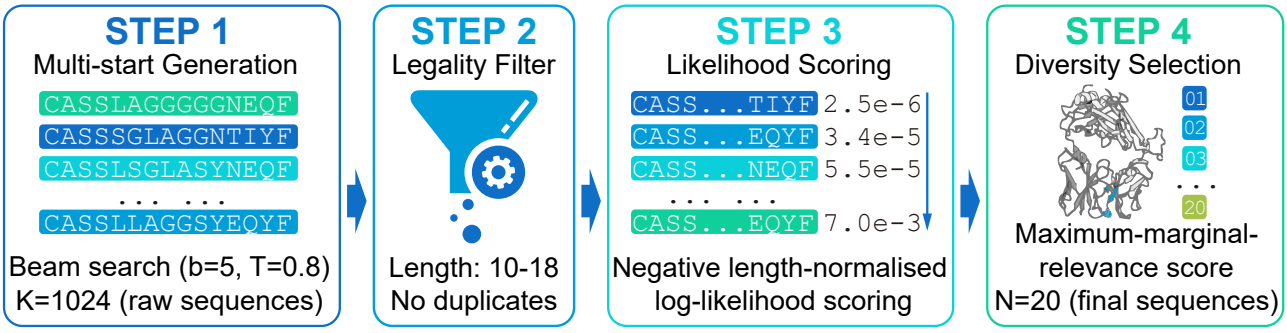

Fig. 2: The inference and post-processing of TCR generation. The pipeline consists of four steps: (1) **Multi-start Generation**: beam search produces 1 024 raw sequences. (2) **Legality Filter**: sequences are filtered by length (10–18 residues) and uniqueness. (3) **Likelihood Scoring**: retained candidates are scored and ranked by the negative length-normalised log-likelihood. (4) **Diversity Selection**: the top 20 sequences are chosen via maximum–marginal–relevance (MMR) to balance binding affinity and sequence diversity.

CDR3 loops conditioned *solely* on the target peptide, leaving the HLA allele unaddressed. None of these works evaluates *dry-lab activation rate* via docking or molecular simulation, and therefore their therapeutic utility remains unclear.

C. Protein sequence generation at large

Transformer language models such as ProGen, ProtGPT2 and ESM-1v [3], [4], [10] demonstrate that pure sequence modelling can create functional enzymes and antibodies. Structure-aware approaches extend this panorama: RFdiffusion [11] and Chroma [12] design full-atom backbones directly, while ESM-IF refines inverse folding with iterative hallucination. Yet these generators are trained on broad protein corpora without any immunological signal, offering no mechanism to bias outputs toward HLA–peptide interfaces.

III. Methodology

Fig. 1 summarises **PhysicoGPTCR**. A dual-conditioned encoder digests the HLA molecule and its bound peptide, while a GPT-style decoder autoregressively emits the T-cell–receptor variable-region sequence.

A. Problem Formulation

Let $m \in \Sigma^{L_m}$ denote an HLA heavy chain, $p \in \Sigma^{L_p}$ the presented peptide, and $t \in \Sigma^{L_t}$ a receptor variable-region sequence over the 20-letter amino-acid alphabet Σ. The task is to model the conditional distribution

$$p_\theta(t \mid m, p) \tag{1}$$

such that samples $\tilde{t} \sim p_\theta$ are syntactically valid, biophysically plausible and strongly biased towards recognising the given HLA–peptide complex.

B. Model Architecture

a) Input fusion: For residue x_i at position i we build three embeddings:

$$\mathbf{e}_i^{\text{tok}} = E_{\text{tok}}(x_i) \in \mathbb{R}^{d_t}, \tag{2}$$

$$\mathbf{e}_i^{\text{phys}} = W_{\text{phys}}\,\boldsymbol{\phi}_i \in \mathbb{R}^{d_p}, \tag{3}$$

$$\mathbf{e}_i^{\text{pos}} = E_{\text{pos}}(i) \in \mathbb{R}^{d_s}, \tag{4}$$

where $\boldsymbol{\phi}_i \in \mathbb{R}^5$ is the *raw physicochemical descriptor* [aromatic, q, $h-$bond, hydrophobicity, $m/m_{\max}$]. These channels are concatenated $\mathbf{z}_i = [\mathbf{e}_i^{\text{tok}}; \mathbf{e}_i^{\text{phys}}; \mathbf{e}_i^{\text{pos}}] \in \mathbb{R}^d$, with $d = d_t + d_p + d_s$ (d_t:d_p:d_s = 2:1:1 as in the code). A learnable gate

$$\mathbf{g}_i = \sigma\big(W_g \mathbf{z}_i + \mathbf{b}_g\big) \quad \in (0,1)^d \tag{5}$$

rescales channel-wise contributions, after which a linear projector yields the final token representation $\mathbf{h}_i = W_f\big(\mathbf{g}_i \odot \mathbf{z}_i\big) + \mathbf{b}_f \in \mathbb{R}^d$.

b) Shared encoder: The input consists of a concatenation of the MHC sequence m and the peptide sequence p, which are jointly encoded by a shared encoder. The combined sequence is mapped to a representation matrix $\mathbf{H}_{\text{src}} \in \mathbb{R}^{L_{\text{src}} \times d}$ through $N_e = 2$ Transformer layers, each consisting of multi-head self-attention:

$$\text{MHA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \big[\text{softmax}(\tfrac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_h}})\mathbf{V}\big] W_o \tag{6}$$

where $d_h = d/n_{\text{head}}$, followed by a two-layer feed-forward block. LayerNorm and residual routes follow the Pre-LN convention.

c) GPT decoder with cross-attention: The target tokens $t = \{\texttt{BOS}, a_1, \ldots, a_{L_t}, \texttt{EOS}\}$ pass through $N_d = 2$ causal layers. At step j, attention scores mix self-history and source memory:

$$\alpha_{ji} = \text{softmax}\Big(\frac{\mathbf{q}_j^\top \mathbf{k}_i}{\sqrt{d_h}}\Big), \quad \tilde{\mathbf{h}}_j = \sum_i \alpha_{ji} \mathbf{v}_i. \tag{7}$$

Ablation experiments (§IV-D) also train a 6+12-layer variant initialised from ProtGPT2 [4]; both depth options share the same input-fusion module.

C. Residue-level Physicochemical Awareness

We embed physicochemical information directly into a neural network's attention mechanism. The 5-D descriptor is first z-scored $\hat{\boldsymbol{\phi}}_i = (\boldsymbol{\phi}_i - \boldsymbol{\mu})/\boldsymbol{\sigma}$ and then linearly mixed

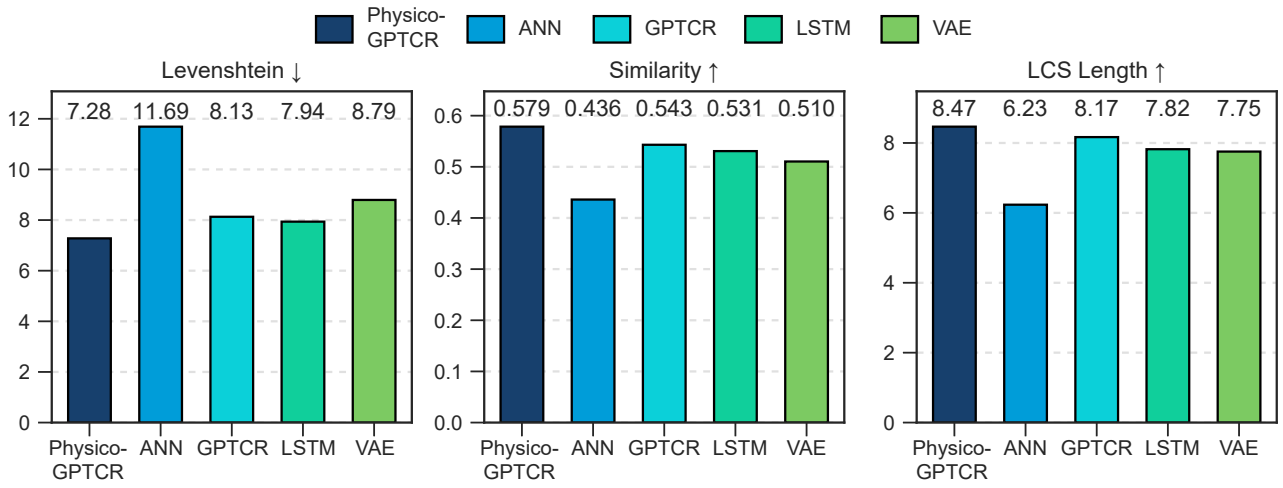

Fig. 3: Comparison across three sequence-level metrics on the 6 200-sample test set (lower Levenshtein ↓ and higher Similarity/LCS ↑ indicate better performance).

into the hidden state ($\mathbf{e}_i^{\text{phys}} = W_{\text{phys}}\hat{\boldsymbol{\phi}}_i$). Consequently, every attention dot-product implicitly contains a chemistry term:

$$\mathbf{q}_j^\top \mathbf{k}_i \approx \underbrace{\psi}_{\text{token}} + \underbrace{\hat{\boldsymbol{\phi}}_j^\top W_{\text{phys}}^\top W_{\text{phys}} \hat{\boldsymbol{\phi}}_i}_{\pi\text{–}\pi,\ \text{salt bridge, VdW, hydrophobic}}, \tag{8}$$

where $\psi := \mathbf{q}_j^\top \mathbf{k}_i|_{\text{token}}$ denotes the token-based sequence motif contribution.

D. Training Objective

We minimise the standard autoregressive negative log-likelihood

$$\mathcal{L}(\theta) = -\sum_{i=1}^{L_t} \log p_\theta(a_i \mid a_{<i}, m, p), \tag{9}$$

optimised with AdamW ($\beta_1 = 0.9,\ \beta_2 = 0.98$, learning-rate 2×10^{-4} with cosine decay, batch size 256).

E. Inference and Post-processing

Given a peptide–MHC pair (m, p) we create a source sequence $m \oplus \langle\texttt{<SEP>}\rangle \oplus p$ (padded to 55 tokens). The decoder emits up to 26 residues preceded by <SOS>. The process is summarized in Fig. 2.

a) Step 1: multi-start generation: This work invokes a temperature–beam search ($T \in [0.6, 1.0]$, beam $b \in [3, 10]$) N=20 times, each call returning the MAP path $t^{(i)}$ with log-probability $\log P_\theta(t^{(i)} \mid m, p)$, yielding the raw pool $\mathcal{C}_{\text{raw}}$ of size 20.

b) Step 2: legality filter: A sequence is kept if $10 \le |t| \le 26$ and it does not appear in the training set, producing $\mathcal{C}_{\text{legal}}$.

c) Step 3: likelihood scoring: This work ranks $\mathcal{C}_{\text{legal}}$ by the negative length-normalised log-likelihood $E_{\text{llh}}(t) = -\frac{1}{|t|}\sum_j \log p_\theta(a_j \mid a_{<j}, m, p)$.

d) Step 4: diversity selection: Sequences are traversed in ranked order and the first K=20 unique ones are retained, giving $\mathcal{S}_{20}$, the set deployed for downstream evaluation.

F. Evaluation Metrics

For every test context (m, p) we report metrics between the *ground-truth* receptor $t^\star$ and the single top-ranked candidate $\hat{t}$.

1) **Levenshtein distance** (↓) $\text{Lev}(t^\star, \hat{t})$ counts the minimum number of edits.
2) **Pairwise similarity** (↑) Normalised Smith–Waterman score with BLOSUM62.
3) **Longest common subsequence length** (↑) $\text{LCS}(t^\star, \hat{t}) = \max_{u \sqsubseteq t^\star,\, u \sqsubseteq \hat{t}} |u|$.

IV. Experiments and Results

A. Experimental Setup

a) Dataset: We collect and generate 31,000 (≈31 k) HLA–peptide–TCR triples from VDJdb, then split them into *train : valid : test* = 7:1:2 at the context level, yielding 21,700 / 3,100 / 6,200 triples respectively. No antigen or HLA leakage occurs across splits.

b) Generation protocol: For every test context we sample $K = 1\,024$ candidates, apply legality, contact-energy and diversity filters (Section III-E), and keep the single best-ranked sequence $\hat{t}$ for evaluation.

B. Baseline Methods

- **ANN**: nearest-neighbour retrieval of the most similar training receptor in BLOSUM62 space.
- **LSTM**: a 4-layer LSTM language model conditioned on (m, p) via context concatenation.
- **VAE**: the variational auto-encoder branch of DeepTCR [6].
- **PhysicoGPTCR**: our full model.

C. Overall Results

To measure the sequence-level performance of models, we used Levenshtein distance, pairwise similarity and the longest common subsequence length of generated sequences compared

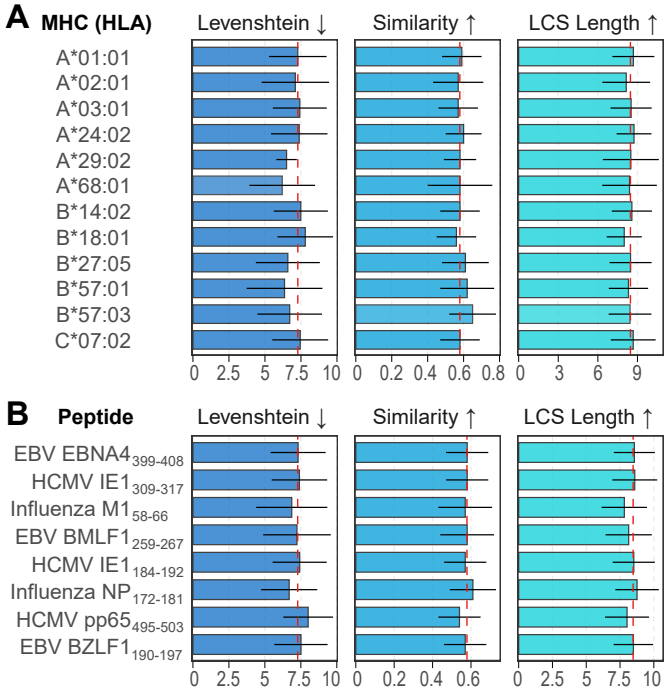

Fig. 4: Model performance across contexts. (A) Sequence-level metrics per MHC allele (mean ± standard deviation, $n \geq 150$ each). (B) Per-epitope sequence-level metrics (mean ± std). Lower Levenshtein ↓ and higher Similarity/LCS ↑ indicate better performance. Red dashed lines show averaged metrics of PhysicoGPTCR.

to actual sequences as metrics. Our model, **PhysicoGPTCR**, has showed better performance than ANN, LSTM, and VAE baselines (Fig. 3). PhysicoGPTCR reduces the edit distance (Levenshtein distance) to the ground-truth receptors by $\sim 9\%$ relative to the best baseline (LSTM) while simultaneously achieving the highest similarity and LCS length.

D. Ablation Study

To validate the effectiveness of one key component, the physicochemical embedding layer, we removed it from **PhysicoGPTCR** to get the **GPTCR**. All experiments share identical training settings and data splits. The results shows that **GPTCR** has worse performance in all three metrics (Fig. 3). The ablation of physicochemical channel increases the average Levenshtein distance to 8.13, reduces similarity to 0.543, and decreases LCS length to 8.17, confirming the benefit of residue-level physicochemical cues.

E. Robustness Across Contexts

To verify that the improvements in Fig. 3 are not driven by a handful of easy cases, we break down the sequence-level metrics by distinct (i) MHC allele and (ii) epitope peptide. Results in Fig. 4 remains stable across the twelve most frequent alleles and the eight most abundant epitopes in the test set: the standard deviation of Levenshtein distance never exceeds ± 2.6 and both Similarity and LCS stay within a narrow band around their global means. Hence the model generalises well to diverse immunological contexts.

V. Clinical Applications

A. Specificity

Neoantigens, derived from tumor-specific mutations, are ideal targets for cancer immunotherapy due to their absence in healthy tissues, which reduces the risk of off-target toxicity. To address this clinical unmet need, our model must accurately differentiate between mutated peptides and their wild-type counterparts, even when only a single mutation is present.

We computationally generated T-cell receptors (TCRs) designed to specifically recognize a mutant peptide over its wild-type version when presented by the MHC molecule (Fig. 5). To assess the specificity of these generated TCRs, we calculated the difference in predicted binding probability between the mutant and wildtype peptides. The analysis showed that TCRs predicted to be positive for the mutant target displayed a clear preference for the mutant peptide, with their Δ Binding Probability values predominantly greater than zero. In contrast, negative TCRs showed a distribution centered around zero, indicating no significant binding preference.

B. Case Study

Our model demonstrates high accuracy in structural features of generated TCR interacting with peptide-Major Histocompatibility Complex (pMHC) compared to actual TCR (Fig. 5B). We validated our approach on two distinct, clinically relevant molecules: a cancer antigen (MART-1) and a viral antigen (SARS-CoV-2). For the MART-1 peptide (ELAGIGILTV) presented by HLA-A$*$02:01, the structure of our generated sequence aligns with the structure of the actual one with a root-mean-square deviation (RMSD) of only $1.060\,\text{Å}$ over 3 135 atoms. Similarly, for the SARS-CoV-2 Spike protein peptide (YLQPRTFLL), the structure of our generated sequence achieved an RMSD of $1.238\,\text{Å}$ over 3 122 atoms when compared to the structure of actual one.

VI. Conclusion and Future Work

PhysicoGPTCR contributes two central insights. First, the model generalises across *twelve* HLA class-I alleles and eight canonical viral and tumour-associated epitopes, suggesting that its performance gains are not restricted to a narrow immunological context. Second, explicitly injecting physicochemical embeddings into the decoder yields consistent improvements over a purely sequence-level baseline, indicating that biophysical priors can be exploited even by a large language model.

Nevertheless, several limitations remain. (1) All benchmarks are restricted to class-I HLA. (2) We rely entirely on *in-silico* metrics. (3) Although larger than prior work, the training corpus is still two orders of magnitude smaller than typical NLP datasets.

Our immediate priorities for future work are threefold. (1) **Class-II generalisation**: extending the approach to HLA-DR/DQ/DP molecules. (2) **Structure-aware scoring**: coupling the generator with *AlphaFold-Multimer* and *RoseTTAFold*

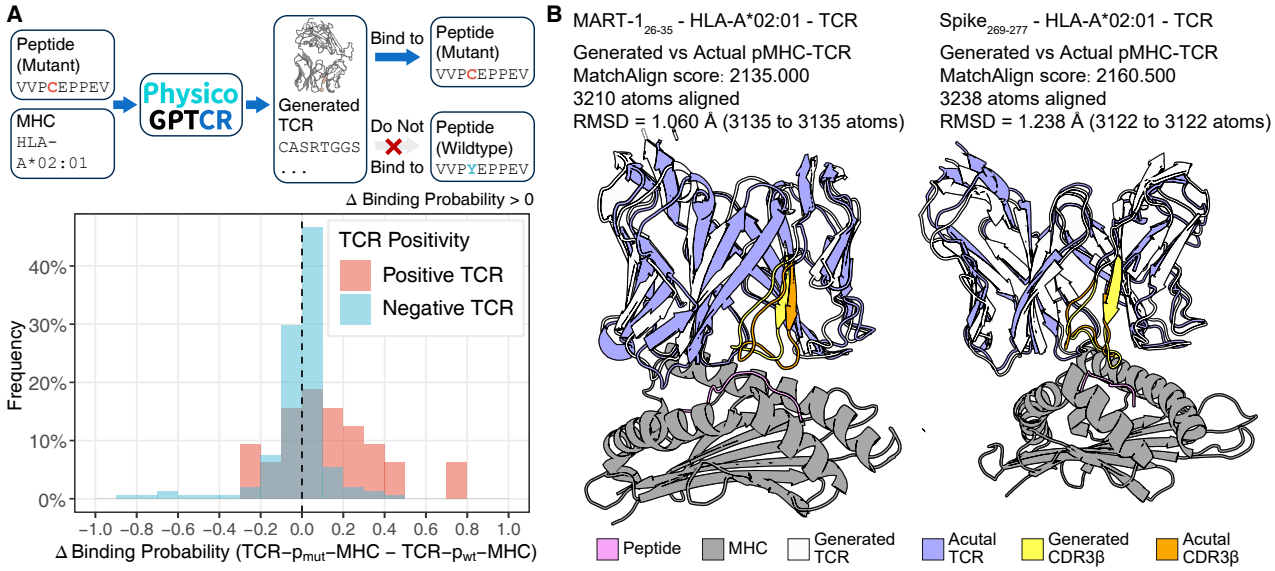

Fig. 5: (A) Frequency distribution of the change in binding probability (Δ Binding Probability = p_{wt}-MHC - p_{mut}-MHC) for computationally validated positive and negative TCRs. (B) Structural alignment of generated (light blue TCR, yellow CDR3β) and actual (white TCR, orange CDR3β) pMHC-TCR complexes, showing RMSDs of 1.060 Å and 1.238 Å for MART-1 and SARS-CoV-2 cases, respectively.

docking pipelines. (3) **Experimental validation**: synthesising top-ranked TCRs for *in-vitro* binding and functional assays.

VII. Acknowledgments

This work is supported by the National Key Research and Development Program of China (2021YFA0910700) to J.-D.H. and the Major Program (92369201) to Y.-F.H from the National Natural Science Foundation of China (NSFC). This research is also supported by the General Program (32571107) of NSFC, the Health and Medical Research Fund (COVID190117, T-11-709/21-N) of the Food and Health Bureau, Hong Kong Special Administrative Region, L & T Charitable Foundation, the Program for Guangdong Introducing Innovative and Entrepreneurial Teams (2019BT02Y198), and Shenzhen Key Laboratory for Cancer Metastasis and Personalized Therapy (ZDSYS20210623091811035) to J.-D.H.

References

[1] S. P. D'Angelo, D. M. Araujo, A. R. A. Razak, M. Agulnik, S. Attia, J.-Y. Blay, I. C. Garcia, J. A. Charlson, E. Choy, G. D. Demetri, M. Druta, E. Forcade, K. N. Ganjoo, J. Glod, V. L. Keedy, A. L. Cesne, D. A. Liebner, V. Moreno, S. M. Pollack, S. M. Schuetze, G. K. Schwartz, S. J. Strauss, W. D. Tap, F. Thistlethwaite, C. M. V. Morales, M. J. Wagner, B. A. Wilky, C. McAlpine, L. Hudson, J.-M. Navenot, T. Wang, J. Bai, S. Rafail, R. Wang, A. Sun, L. Fernandes, E. Van Winkle, E. Elefant, C. Lunt, E. Norry, D. Williams, S. Biswas, and B. A. Van Tine, "Afamitresgene autoleucel for advanced synovial sarcoma and myxoid round cell liposarcoma (SPEARHEAD-1): an international, open-label, phase 2 trial," *The Lancet*, vol. 403, no. 10435, pp. 1460–1471, 2024. [Online]. Available: https://doi.org/10.1016/S0140-6736(24)00319-2

[2] T. Mora and A. M. Walczak, "Quantifying lymphocyte receptor diversity," in *Systems Immunology: An Introduction to Modeling Methods for Scientists*, J. Das and C. Jayaprakash, Eds. Boca Raton FL, United States: CRC Press, Taylor & Francis Group, 2019, pp. 183–198.

[3] J. Meier, R. Rao, R. Verkuil, J. Liu, T. Sercu, and A. Rives, "Language models enable zero-shot prediction of the effects of mutations on protein function," *Advances in Neural Information Processing Systems*, vol. 34, pp. 29 287–29 303, 2021.

[4] N. Ferruz, S. Schmidt, and B. Höcker, "Protgpt2 is a deep unsupervised language model for protein design," *Nature Communications*, vol. 13, no. 1, p. 4348, 2022.

[5] A. Montemurro, V. Schuster, H. R. Povlsen, A. K. Bentzen, V. Jurtz, W. D. Chronister, A. Crinklaw, S. R. Hadrup, O. Winther, B. Peters *et al.*, "Nettcr-2.0 enables accurate prediction of tcr-peptide binding by using paired tcrα and β sequence data," *Communications Biology*, vol. 4, no. 1, p. 1060, 2021.

[6] J.-W. Sidhom, H. B. Larman, D. M. Pardoll, and A. S. Baras, "Deeptcr is a deep learning framework for revealing sequence concepts within t-cell repertoires," *Nature Communications*, vol. 12, no. 1, p. 1605, 2021.

[7] A. Weber, J. Born, and M. Rodriguez Martínez, "Titan: T-cell receptor specificity prediction with bimodal attention networks," *Bioinformatics*, vol. 37, no. Supplement_1, pp. i237–i244, 2021.

[8] Z. Zhang, D. Xiong, X. Wang, H. Liu, and T. Wang, "Mapping the functional landscape of t cell receptor repertoires by single-t cell transcriptomics," *Nature Methods*, vol. 18, no. 1, pp. 92–99, 2021.

[9] Y. Lin, D. Zhang, and Y. Liu, "Tcr-gpt: Integrating autoregressive model and reinforcement learning for t-cell receptor repertoires generation," *arXiv preprint arXiv:2408.01156*, 2024.

[10] A. Madani, B. McCann, N. Naik, N. S. Keskar, N. Anand, R. R. Eguchi, P.-S. Huang, and R. Socher, "Progen: Language modeling for protein generation," *arXiv preprint arXiv:2004.03497*, 2020.

[11] J. L. Watson, D. Juergens, N. R. Bennett, B. L. Trippe, J. Yim, H. E. Eisenach, W. Ahern, A. J. Borst, R. J. Ragotte, L. F. Milles *et al.*, "De novo design of protein structure and function with rfdiffusion," *Nature*, vol. 620, no. 7976, pp. 1089–1100, 2023.

[12] A. Singh, "Chroma is a generative model for protein design," *Nature Methods*, vol. 21, no. 1, pp. 10–10, 2024.