Briefings in Bioinformatics, 2026, 27, bbag616
https://doi.org/10.1093/bib/bbag616
Published: 11 November 2026
Case Study

OXFORD

CRAFT: an antigen-specific TCR generative model for conditional *de novo* design with physicochemical constraints

Xiaochuan Yang[1], Zaiduo Li[1], Zhijuan Li[2,3], Wei Zhang[2,3], Xianyu Zhang[4], Tiansong Yang[5], Mingliang Liu[1,*], Xiaokun Li[1,2,3,6,7,*]

[1]Key Laboratory of Information Fusion Estimation and Detection, Heilongjiang University, Xuefu Road, Harbin 150080, China
[2]School of Computer and Big Data, Heilongjiang University, Xuefu Road, Harbin 150080, China
[3]Postdoctoral Program of Heilongjiang Hengxun Technology Co., Ltd., Xuefu Road, Harbin 150080, China
[4]Department of Breast Surgery, Harbin Medical University Cancer Hospital, Haping Road, Harbin 150081, China
[5]Department of Rehabilitation, the First Affiliated Hospital of Heilongjiang University of Traditional Chinese Medicine, and Traditional Chinese Medicine Informatics Key Laboratory of Heilongjiang Province, Heping Road, Harbin 150040, China
[6]Faculty of Computing, Harbin Institute of Technology, West DaZhi Street, Harbin 150001, China
[7]Shandong Hengxun Technology Co., Ltd., Miaoling Road, Qingdao 266100, China

*Corresponding authors. Mingliang Liu, Key Laboratory of Information Fusion Estimation and Detection, Heilongjiang University, Xuefu Road, Harbin 150080, China. E-mail: 2002039@hlju.edu.cn; Xiaokun Li, Faculty of Computing, Harbin Institute of Technology, West DaZhi Street, Harbin 150001, China. E-mail: 24b303024@stu.hit.edu.cn.

Abstract

T-cell receptors (TCRs) are essential to adaptive immunity and mediate the recognition and elimination of malignant or infected cells through antigen-specific interactions. Despite the clinical progress of TCR-based immunotherapies, computational *de novo* design of antigen-specific TCRs remains difficult. The main difficulties include the complexity of peptide–major histocompatibility complex (pMHC)-dependent recognition, strict physicochemical requirements at the binding interface, and the small amount of paired training data available for model development. Current generative approaches depend largely on 1D sequence statistics and do not capture pMHC presentation context, interfacial physicochemical properties, and native receptor priors within a unified framework. We present conditional reconstruction of antigen fragments for TCRs (CRAFT), a decoupled generative model for pMHC-conditioned *de novo* design of antigen-specific TCR CDR3β sequences under physicochemical constraints. With large-scale pretraining, CRAFT learns base representations that help reduce the limitations imposed by data scarcity. We further incorporate antigen-derived contextual information to guide antigen-specific TCR CDR3β sequence design. The architecture comprises an antigen encoder (PrimeBERT) that captures the pMHC context and an autoregressive decoder (PrimeTCR) that preserves native receptor priors. The encoder and decoder are coupled through cross-attention and augmented with Atchley factor embeddings, allowing CRAFT to incorporate physicochemical constraints relevant to antigen-conditioned receptor design during sequence decoding. Across multiple evaluations, CRAFT demonstrates competitive sequence naturalness, together with favorable performance in specific binding, and predicted structural plausibility. Overall, CRAFT provides a computational framework for developing personalized tumor immunotherapies. It also holds potential to support clinical translation by enabling the efficient, targeted discovery of candidate TCRs.

Keywords CDR3β generation, neoantigen, physicochemical constraints, decoupled conditional generation, pMHC–contextualized decoding

Introduction

The continued development of T-cell engineering is reshaping cancer immunotherapy and advancing precision medicine [1–5]. Chimeric antigen receptor (CAR)-T cell therapy has demonstrated substantial efficacy in hematologic malignancies [6, 7]. However, CAR-T cell recognition relies on stable surface antigen expression. This dependency limits its application in highly heterogeneous solid tumors, where antigen loss and immune escape frequently occur [8, 9]. In contrast, engineered T-cell receptors (TCRs) circumvent this surface-target restriction by recognizing intracellularly derived peptides presented by major histocompatibility complex (MHC) molecules, including neoantigens arising from tumor-associated mutations [10, 11]. This recognition mechanism expands the addressable antigen space and supports mutation-directed individualized immunotherapy [12–14].

Received: April 23, 2026. **Revised:** July 25, 2026. **Accepted:** October 14, 2026
© The Author(s) 2026. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

Native TCR diversity stems primarily from V(D)J gene rearrangement of highly variable α and β chains [15–17]. This process creates a massive and sparse receptor sequence space. Within this space, the complementarity-determining region 3 of the β chain (CDR3β) directly engages the peptide–MHC (pMHC) complex and dictates the precision of antigen recognition. Accordingly, the CDR3β loop has emerged as a primary target for antigen-conditioned *de novo* candidate generation in computational TCR design [18, 19]. Exploring this enormous sequence space via traditional experimental screening is prohibitively resource-intensive. To address these bottlenecks, a central challenge in computational modeling is to efficiently capture the sequence constraints linked to pMHC recognition and incorporate immunogenic information into target-specific CDR3β design while preserving native-like receptor features [20].

Deep learning has substantially transformed the traditional trial-and-error paradigm of protein engineering [21–23]. Protein language models, such as ESM (Evolutionary Scale Model) [24] and ProGen [25], are trained via unsupervised learning on large-scale protein sequence corpora to learn long-range evolutionary constraints, local compositional rules, and higher-order sequence patterns [23]. They provide effective generation priors for *de novo* sequence design. Recent advances in computational protein modeling have enabled protein design to go beyond sequence-level optimization [26]. Methods such as RoseTTAFold All-Atom [27], ProteinMPNN [28], and RFdiffusion3 [29] can integrate structural information into the design process and also support the construction of complex functional topologies [30]. Related progress has also informed automated antibody design [31–36].

However, these advances in protein and antibody design are not directly transferable to TCR design. Unlike surface-binding antibodies [37], TCRs recognize pMHC interfaces that involve geometric complementarity, conformational variability, and inherently weak binding affinities [38–42]. These features make computational TCR generation substantially more challenging [43]. The difficulty is further compounded by a highly sparse sequence space, stringent interfacial physicochemical constraints, and limited paired training data [44]. Recent studies in computational immunology have therefore developed specialized frameworks for immunogenic TCR generation [19, 44, 45]. For example, GRATCR [46] adopts a Seq2Seq encoder–decoder architecture for pMHC-guided generation, whereas ER-TRANSFORMER [47] introduces bidirectional cross-attention to improve interface-level modeling. In addition, TCR-VALID [45] employs a variational autoencoder to sample candidates from a continuous latent space. Parallel to these deep learning architectures, many current approaches rely on sequence-similarity retrieval or 1D language-model frequency patterns [48–50]. Although these statistical strategies can recover local compositional regularities from native repertoires [51], they do not reliably preserve key interfacial features when designing TCRs for specific or previously unseen antigens [52].

Despite these conceptual advances, few current generative architectures jointly and explicitly represent the pMHC background and physicochemical constraints that shape TCR recognition [53–55]. Computationally, training a conditional generator directly on limited and heterogeneous paired data often overfits the model to local training distributions. This optimization approach erodes the native generative behavior and sequence diversity of the decoder. Therefore, the field requires a unified framework capable of preserving natural generative priors while accommodating explicit biological constraints through a stable optimization pathway.

In this study, we aim to generate TCR CDR3β sequences that retain native-like characteristics under antigenic conditions while remaining compatible with the relevant pMHC presentation context [56]. To achieve this, we propose CRAFT. CRAFT employs a decoupled design strategy. It separately models antigen-side conditioning and receptor-side autoregressive generation, ultimately coupling them through an explicit conditioning mechanism. Specifically, an antigen-side encoding module represents epitope sequences and their presentation background [57]. Simultaneously, a receptor-side generative module learns the autoregressive distribution of CDR3β sequences [58]. CRAFT incorporates antigen-derived information into the decoding process via a cross-attention conditioning mechanism. This mechanism allows conditional representations to interact directly with receptor-side decoding states at each generation step. Instead of relying on a monolithic architecture, CRAFT preserves the distinct representational strengths of both modalities and dynamically integrates them during sequence decoding.

CRAFT advances conditional TCR generation beyond 1D statistical fitting by incorporating both the antigen-presentation context and residue-level physicochemical constraints. Rather than relying on computationally expensive 3D structural supervision, this framework uses pMHC-conditioned representations and explicit physicochemical descriptors to guide sequence decoding toward more plausible candidates. By introducing antigen-side contextual information early in decoding [59], CRAFT promotes physicochemical compatibility while preserving the native generative behavior of the receptor. Introducing antigen-derived contextual cues at the start of sampling defines an allowable design range that maintains physicochemical coherence without altering the receptor's intrinsic generative behavior. Furthermore, the cross-attention mechanism offers an interpretable lens for examining position-level interaction patterns under antigenic conditioning. The following sections evaluate this framework across sequence naturalness, conditional generation fidelity, and predicted structural plausibility.

Results

Overview of the CRAFT framework

An overview of the CRAFT workflow and model architecture is shown in Fig. 1 and Supplementary Fig. S1. CRAFT is designed for the *de novo* generation of TCR CDR3β sequences in a defined pMHC context, with the aim of producing candidates that retain naturalness, specificity, and predicted structural plausibility under antigen-conditioned and physicochemical design objectives (Fig. 1a and b). Unlike sequence-only generative models [60], CRAFT explicitly incorporates the pMHC presentation background and residue-level physicochemical constraints derived from Atchley factors. This design allows conditional cues and sequence properties to be handled within a unified framework and biases early sampling toward more plausible candidates. At the model level, CRAFT adopts an asymmetric transformer encoder–decoder architecture, in which PrimeBERT encodes epitope and presentation-context information, whereas PrimeTCR performs conditional CDR3β generation (Fig. 1c and d). Within this architecture, the Epitope Blocks illustrated in the schematic correspond to the pretrained BERT encoder layers of PrimeBERT. In parallel, the TCR Blocks represent the GPT2-based autoregressive decoder layers that constitute PrimeTCR (Supplementary Fig. S1). PrimeBERT and PrimeTCR are first pretrained to learn general representations

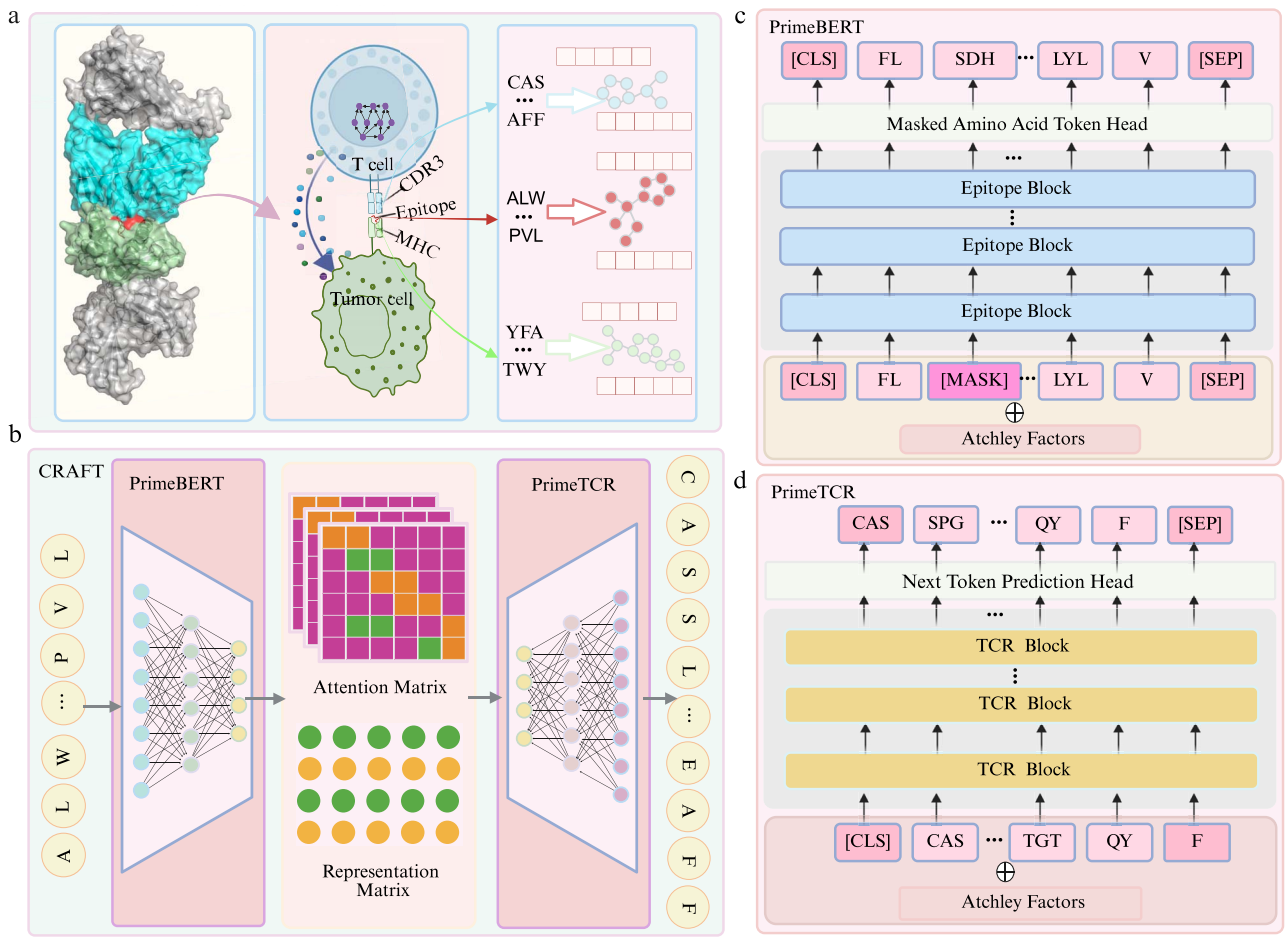


Figure 1. Overview of the CRAFT framework. (a) Schematic illustration of a TCR recognizing a tumor-associated epitope. (b) Architecture of CRAFT, which employs an encoder–decoder structure to generate antigen-specific TCR CDR3β sequences from pMHC inputs. (c) PrimeBERT, a masked language model pretrained on antigen peptides through masked token prediction, with Atchley factors incorporated as physicochemical features. (d) PrimeTCR, an autoregressive module pretrained on TCR β-chain sequences by next-token prediction, with sequences initiated by "CAS" and constrained to terminate with "F" or "W"; Atchley factors are incorporated as physicochemical features.

of epitopes and natural receptor sequences, respectively, and are then adapted to the antigen-conditioned CDR3β generation task through transfer learning-based fine-tuning on paired antigen–TCR data (Supplementary Table S1).

To systematically validate this design, the following sections evaluate CRAFT from four complementary perspectives: repertoire-level statistical conformity, immunogenic sequence generation, structural plausibility, and residue-level interpretability. These results suggest that CRAFT may provide an extensible computational foundation for advancing antigen-specific TCR engineering in precision immunotherapy.

CRAFT generates statistically natural and combinatorially consistent receptor libraries

CRAFT-generated CDR3β sequences broadly recapitulated the central length range of native TCR repertoires, with a bell-shaped distribution peaking at 13–16 amino acids (Fig. 2a). Compared with the natural repertoire, the generated sequences were modestly longer on average (14.756 versus 14.346 amino acids) and showed lower dispersion (SD, 1.434 versus 1.677 amino acids), indicating a more concentrated length distribution. Nevertheless, CRAFT largely avoided the long-tail truncation and aberrantly short sequences reported for earlier recurrent neural network architectures [61]; comparisons with other generative models are shown in Supplementary Fig. S2. This concentration may reduce implausible length extremes, but may also reflect limited exploration associated with autoregressive decoding and candidate selection. Generation probabilities (p_{gen}) were further estimated using Optimized Likelihood estimate of immunoGlobulin Amino-acid sequences (OLGA) [62]. The higher mean $\log_{10}(p_{\mathrm{gen}})$ of CRAFT-generated sequences relative to the natural repertoire (-10.011 versus -11.516) is consistent with reduced representation of low-generation-probability sequences.

We benchmarked the statistical conformity of CRAFT against four representative baselines (Supplementary Note 1): TCR-epiDiff [63], GRATCR [46], TcrDesign [64], and the large-scale pretrained model TCR-T5 [44]. The distribution of TCR-epiDiff in Fig. 2b covered a broader range, with higher variability across the generated outputs. GRATCR and TcrDesign deviated more clearly from this distributional pattern, reflecting weaker alignment with the probabilistic boundaries defined by natural sequences. While TCR-T5 achieved strong distributional coverage, CRAFT showed a highly comparable degree of fit under the evaluated settings. Furthermore, K-mer spectrum analysis [65] provided a complementary view of local compositional grammar. Although distribution fitting became progressively more challenging for all models at longer K-mer lengths, CRAFT and

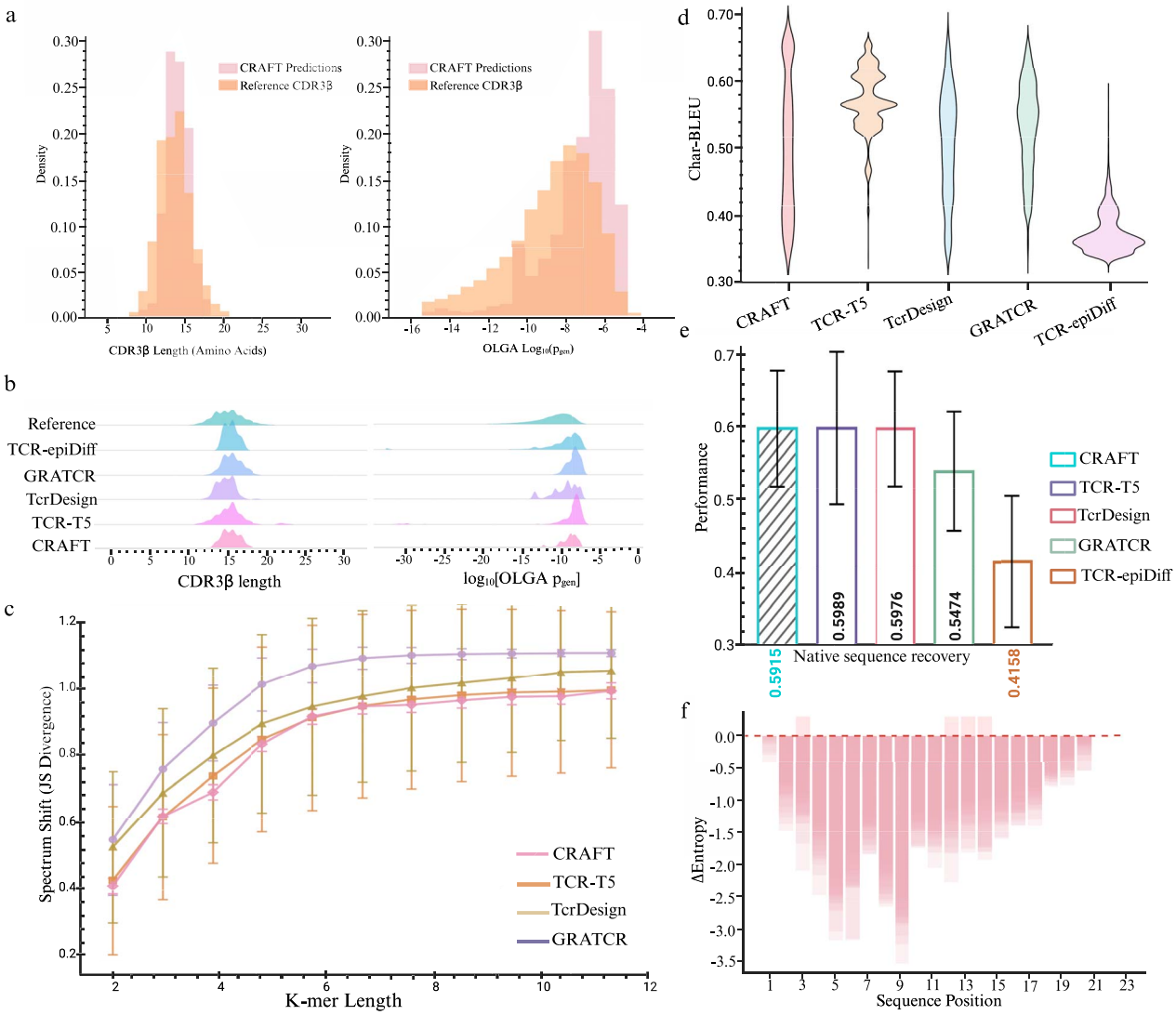


Figure 2. Generative fidelity and sequence-distribution analysis of CRAFT. (a) Overlapping density distributions of CDR3β length and OLGA generation probability ($\log_{10} p_{gen}$) for CRAFT-generated and reference sequences. (b) Ridge plots comparing CDR3β length and $\log_{10} p_{gen}$ distributions across CRAFT and baseline models. (c) K-mer spectrum shift measured by JS divergence, showing the relative preservation of local sequence compositional motifs by CRAFT. TCR-epiDiff was excluded from the K-mer spectrum comparison due to its length distribution deviating significantly from the reference, rendering the K-mer frequency estimates incomparable. (d) Char-BLEU scores for evaluating sequence similarity to natural counterparts. (e) Native sequence recovery performance across the compared models. (f) Positional Δentropy profile, representing the Shannon entropy difference between CRAFT-generated and reference sequences and illustrating the extent to which positional diversity is preserved relative to the reference repertoire.

TCR-T5 consistently achieved the lowest Jensen–Shannon (JS) divergence values in the short K-mer regime ($k = 2, 3, 4$). Specifically, CRAFT reached JS divergences of ~0.40, 0.58, and 0.64, respectively, thereby outperforming both GRATCR and TcrDesign (Fig. 2c). Together, the length distribution, p_{gen}, and K-mer analyses supported the statistical conformity and repertoire-level naturalness of CRAFT-generated sequences.

CRAFT balances sequence fidelity, conditional separability, and diversity in receptor library generation

Building on the close statistical agreement between CRAFT-generated sequences and the global profile of native TCR repertoires, we evaluated the model regarding sequence fidelity, conditional separation, and repertoire diversity. When evaluated using character-level BLEU (Char-BLEU) [66] and native sequence recovery (NSR) rate metrics, CRAFT, TcrDesign, and the large-scale pretrained model TCR-T5 formed the top-performing group among all evaluated methods (Fig. 2d–e). Specifically, in the Char-BLEU assessment, CRAFT exhibited a score distribution centered above 0.50, with the bulk of generated sequences achieving scores broadly comparable with those of TCR-T5 and TcrDesign. In contrast, TCR-epiDiff produced markedly lower scores, with its distribution concentrated in the lower range of the evaluation spectrum. Under the more stringent criterion of NSR rate, CRAFT reached 0.5915, closely matching both TCR-T5 (0.5989) and TcrDesign (0.5976), while substantially outperforming diffusion-based TCR-epiDiff (0.4158) and GRATCR (0.5474). Taken together, these findings indicated that CRAFT maintained robust sequence-level fidelity in preserving local sequence patterns and recovering

native receptors, consistently matching the performance of highly parameterized state-of-the-art baselines.

When sequence collections generated under different epitope settings were compared, CRAFT outputs remained closely tied to the conditioning input. Jaccard similarity analysis (Supplementary Table S2) yielded a mean pairwise score of 0.00065 across epitope-specific sets, indicating only limited sharing between antigen contexts. Repertoires produced under distinct epitope conditions therefore stayed clearly separated, with little convergence from one condition to another. At the same time, this conditional separation was not accompanied by a noticeable loss of candidate diversity. The fraction of nonredundant CRAFT sequences was 0.940, approaching the 0.960 observed in the natural reference repertoire, while within-epitope duplication remained low throughout. Together, these measurements indicate that CRAFT preserves condition-dependent separation without narrowing the diversity of the generated repertoire and continues to sample a broad expanse of sequence space.

A position-wise Shannon entropy profile provided a complementary view of the generated sequence distribution (Fig. 2f). Here, Δentropy was defined as $H_{\text{generated}} - H_{\text{natural}}$ at each aligned position; values close to zero indicate preservation of natural positional diversity, whereas negative and positive values indicate lower and higher diversity than the natural reference repertoire, respectively. Entropy differences near the two anchor segments of the CDR3β loop remained close to zero, whereas the central variable segment showed a downward shift ($\Delta\text{entropy} \approx -3.5$), producing a V-shaped profile. This localized entropy reduction is consistent with stronger epitope-conditioned sequence preferences in the central CDR3β region, but it may also reflect reduced local exploration and should therefore not be interpreted as intrinsically favorable. When considered together with the recovery rates, Jaccard indices, and uniqueness metrics, the results suggest that CRAFT introduced nonuniform conditional constraints while retaining substantial candidate-level diversity, rather than undergoing uniform repertoire-wide contraction. These sequence-level observations provided the basis for the subsequent structural analyses.

Benchmark evaluation of antigen-conditioned sequence specificity and plausibility across common, rare, and novel epitope settings

In benchmark evaluations, CRAFT was evaluated alongside TcrDesign, TCR-T5, GRATCR, TCR-epiDiff, and a random mutation baseline on a multilevel epitope test set partitioned into common, rare, and novel subsets (see **Materials and Methods**). To assess binding specificity toward target antigens, we used TCRBinder (a high-accuracy binding predictor) [19] to predict the binding probabilities of the generated CDR3β sequences. For each evaluation panel, CRAFT was compared independently with each of the five baselines using two-sided Welch's t-tests, followed by Holm–Bonferroni correction across the five pairwise comparisons. The value reported in each figure panel represents the largest P value among these comparisons, providing a conservative summary of the statistical significance across all CRAFT-versus-baseline comparisons. The baseline yielding the maximum raw P value, together with the corresponding adjusted P value and Cohen's d effect size, is reported in Supplementary Table S3. For common epitopes, CRAFT exhibited higher median binding probabilities than the reference models ($p = 2.3 \times 10^{-7}$, Fig. 3a). This advantage was also observed in the rare epitope subset ($p = 7.1 \times 10^{-6}$, Fig. 3b) and extended to the novel epitope subset (Supplementary Fig. S3a), indicating that the model maintained binding specificity across varying antigenic contexts. Sequence-level fidelity was subsequently evaluated using the score distributions of the BLOcks SUbstitution Matrix with a 62% clustering threshold (BLOSUM62) to measure the biochemical similarity of the generated sequences to naturally occurring reference sequences. Within the common epitope setting, CRAFT achieved a higher median similarity score than the reference methods ($P = 1.7 \times 10^{-7}$, Fig. 3c). As the evaluation shifted to the data-limited rare and novel conditions (Supplementary Fig. S3b and S3c), the score distributions of several reference methods, particularly TCR-epiDiff, showed a downward shift. In contrast, CRAFT maintained comparatively consistent score distributions across these evaluation regimes, suggesting that its sequence generation was less sensitive to representation frequency than that of several baseline methods. To further characterize variant diversity at the sequence level, we constructed a petal network visualization (Fig. 3d), which illustrated the clustering of CRAFT-generated variants targeting four representative epitopes. The corresponding visualizations for the baseline models in the common subset are available in Supplementary Fig. S4.

The 3D structural prediction metrics provided an orthogonal line of evidence supporting these sequence-level observations (Supplementary Table S4). In assessing local conformational quality, CRAFT achieved the highest predicted Local Distance Difference Test (pLDDT) score among all evaluated methods, reaching 0.9169. Furthermore, CRAFT yielded the lowest backbone root mean square deviation (RMSD) at 1.366 Å, indicating minimal geometric deviation from the natural structural references. In terms of the predicted Template Modeling (pTM) score, which reflects global topological accuracy, CRAFT reached 0.719, remaining close to TcrDesign, the top-scoring method at 0.721. Taken together, these readouts were consistent with structurally plausible and internally coherent CRAFT-generated candidates at the prediction level. A classification-based surrogate model was also used as an added assessment (Supplementary Table S4). Absolute scores stayed low across all methods, but CRAFT maintained a comparatively even pattern of performance (see **Materials and Methods**).

Ablation analysis of matched pMHC conditioning and core CRAFT components

To determine whether the performance of CRAFT arose from matched pMHC conditioning rather than from the unconditional receptor prior or the generation procedure alone, we compared the full model with controlled variants under the same data split, decoding settings, and post-processing protocol. These variants included a PrimeTCR-only model, a mismatched pMHC control, models without Atchley-factor inputs or physicochemical regularization, and simplified conditioning alternatives. Across common, rare, and novel epitope subsets, full CRAFT showed the most favorable TCRBinder-predicted binding-probability distributions (Fig. 4a–c). In particular, removing the antigen-side condition or replacing the cognate pMHC with a mismatched context reduced the predicted binding scores, indicating that the observed performance was not explained by the autoregressive receptor prior alone. The corresponding analyses of sequence-level metrics and structural rationality metrics showed concordant trends (Supplementary Figs S5–S6; Supplementary Tables S5 and S6). These results provide computational evidence that matched pMHC information contributes to target-specific CDR3β

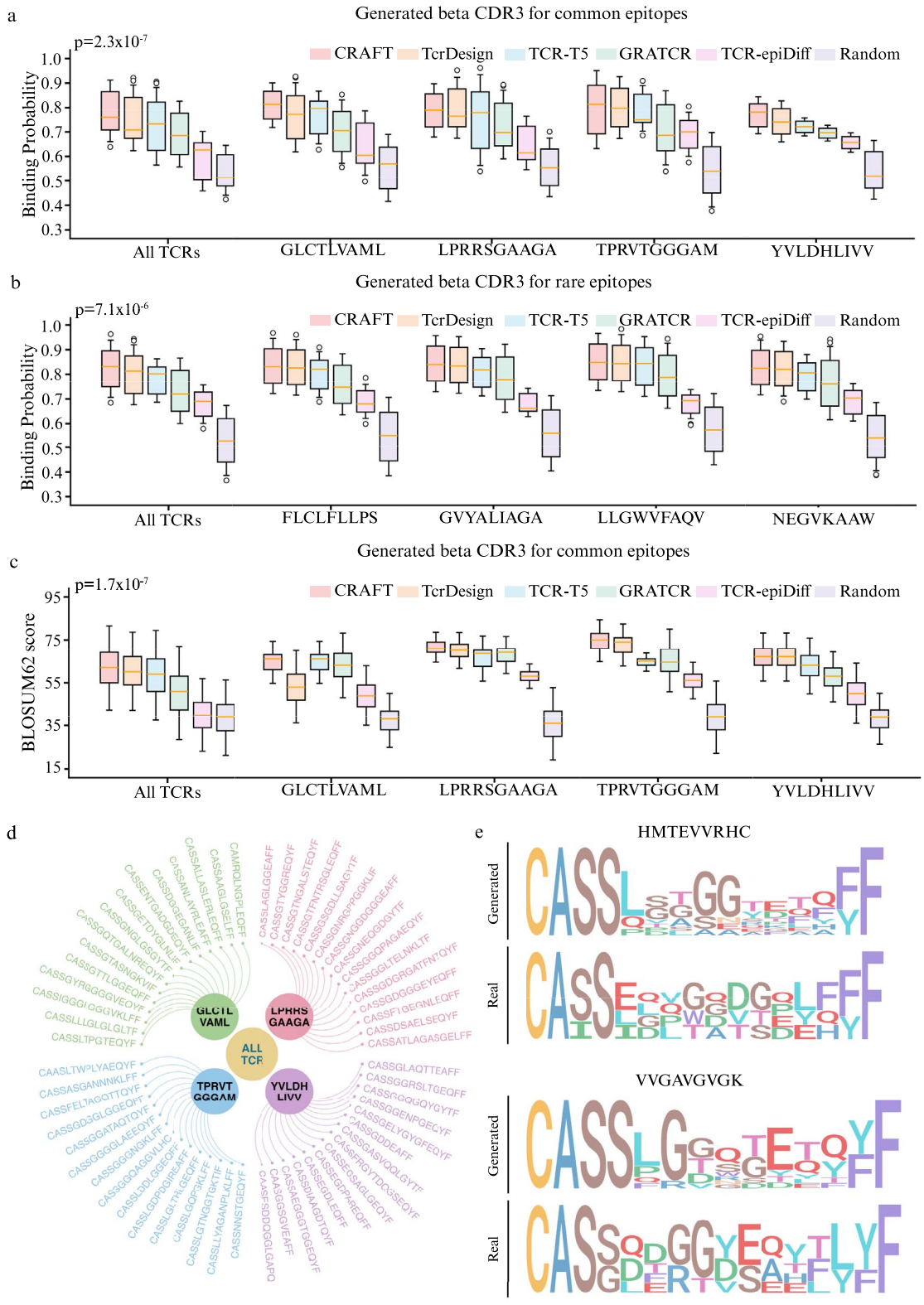


Figure 3. Epitope-specific sequence generation and functional plausibility of CRAFT. (a) Performance comparison of TCRBinder-predicted binding probabilities for artificial TCR CDR3β sequences generated by CRAFT, baseline models, and random mutations specifically for common epitopes. Binding probability is used to assess the functional binding capacity of the generated sequences. (b) Performance comparison of TCRBinder-predicted binding probabilities for the generated CDR3β sequences across the rare epitope subset. (c) Boxplots detailing the BLOSUM62 similarity scores between the generated and natural reference CDR3β sequences specifically for common epitopes. Statistical significance (P value) is indicated for pairwise comparisons. (d) Petal network visualization of CRAFT-generated CDR3β variants targeting four representative epitopes, illustrating sequence diversity and clustering under distinct antigenic conditions. (e) Sequence logos comparing the positional amino acid distributions of CRAFT-generated versus real CDR3β sequences for the specific epitopes HMTEVVRHC and VVGAVGVGK, highlighting preserved biochemical preferences.

generation, while experimental validation will be required to establish the functional properties of individual candidates.

We next performed component-specific ablations to resolve the contributions of the physicochemical representation, physicochemical regularization, decoupled pretraining, and conditional-decoding interface. To determine whether the fixed Atchley representation provided information beyond a generic residue embedding, we compared the full CRAFT model with three variants: one without Atchley inputs, one using learned 5D residue features, and one using fixed random 5D residue features. Full CRAFT retained more favorable binding-probability and BLOSUM62 distributions across the three epitope-frequency regimes (Supplementary Figs S7–S9; Supplementary Tables S7 and S8). We further evaluated the physicochemical regularization term by setting $\lambda = 0$ and sweeping λ from 0 to 2.0; the validation analysis identified $\lambda = 0.5$ as providing the most balanced performance across sequence recovery, predicted binding, similarity, and structural readouts (Supplementary Figs S10–S11). In parallel, CRAFT initialized with independently pretrained PrimeBERT and PrimeTCR outperformed variants without pretraining or with encoder-only or decoder-only pretraining, particularly in the rare and novel epitope subsets (Supplementary Figs S12–S14 and Supplementary Tables S9 and S10). Finally, replacing cross-attention with prefix or pooled-CLS conditioning, or removing residual fusion, reduced the sequence-level and predicted-binding performance relative to full CRAFT (Supplementary Figs S15–S17 and Supplementary Tables S11 and S12). Together, these analyses indicate that the complete model benefits from the coordinated use of matched pMHC conditioning, dual-branch pretraining, physicochemical regularization, and token-level cross-attention with residual fusion.

To assess how antigen conditioning shaped the regions of sequence space reached by CRAFT, we embedded >20 000 TCR CDR3β sequences generated with and without antigen conditioning into a Uniform Manifold Approximation and Projection (UMAP) space trained on authentic antigen–TCR pairings (Fig. 5a). The generated set occupied the main high-density neighborhoods associated with dominant natural clonotypes and extended into nearby low-occupancy areas, which kept the spatial pattern continuous around the major natural clusters. Representative sequences were then drawn from multiple cluster-defined regions and evaluated by ESM3-based [67] structural prediction. Predicted folding confidence remained high across these candidates, with TM-scores (Fig. 5a) ranging from 0.82 to 0.89 and pTM scores from 0.83 to 0.89. Evidence from similarity distributions, structure-related metrics, surrogate-model evaluation, and latent-space placement showed that CRAFT sustained immunogenic sequence generation across common, rare, and novel epitope settings. At the same time, the model retained sequence fidelity and predicted structural plausibility while enlarging the accessible candidate space.

Structural characterization and interface distribution analysis of CRAFT-generated TCR candidates for clinically relevant cancer epitopes

To examine the structural behavior of CRAFT-generated candidates in clinically relevant cancer settings, we selected two representative driver mutation-derived neoantigens (Supplementary Table S13) for further analysis: the TP53 R175H-derived peptide HMTEVVRHC [68] and KRAS G12-associated peptides, including VVGAVGVGK and related variants [69, 70]. The corresponding sequence logos are shown in Fig. 3e, with additional sequence views provided in Supplementary Fig. S18. For each target, CRAFT was used to generate a raw set of up to 1000 CDR3β candidates. Complex structures for the corresponding post-filtered valid sequences were then predicted using ESM3 [67] or AlphaFold3 [71], followed by downstream structural analyses. Across the resulting candidate sets, predicted structural confidence remained consistently high, yielding a cohort-average pLDDT above 0.90 and an average pTM above 0.80. Subsequently, representative high-confidence complexes were aligned against experimentally resolved native structures (Fig. 5b–c; Supplementary Figs S19 and S20). For TP53 R175H (PDB IDs: 6VQO and 7RM4) and KRAS G12 (PDB IDs: 9IKY and 7OW6), complex predictions containing the CRAFT-generated CDR3β candidates showed strong agreement with the native references, with TM-scores exceeding 0.93 and backbone RMSD values below 1.30 Å after global alignment using TM-align [72]. Structural superposition further revealed that, despite marked variation in amino acid composition among the generated CDR3β sequences, their overall orientation above the pMHC surface and the anchoring positions of the backbone scaffold remained closely aligned with the native conformations. These case studies indicate close predicted structural agreement among selected CDR3β-conditioned complexes in cancer-related neoantigen contexts.

We then analyzed multiple cancer-associated epitopes to compare interface-level structural distributions across different antigen backgrounds. For each target, solvent-accessible surface area (SASA) was derived from the predicted static structures of the valid post-filtered candidates (Fig. 6a). In this set of epitopes, SASA, as a readout of interface exposure and geometric compactness, followed a broadly similar distribution. For several targets, the generated ensembles showed lower median SASA values than those of the static reference structures of the native complexes (Fig. 6a, star markers), which corresponded to more compact interface geometries. Epitopes with relatively rigid interfaces, including KQWLVWLFL, also had a narrower interquartile SASA range, indicating a higher level of geometric consistency within the generated structures. We also assessed ΔSASA, a measure of buried SASA, and observed epitope-dependent differences in interface burial (Fig. 6b). Candidate sets for ALHGGWTTK and VVGAVGVGK had broad ΔSASA distributions, whereas HMTEVVRHC remained tightly constrained and stayed near the native reference. These SASA comparisons showed that CRAFT-generated interfaces were not confined to one fixed geometric state and instead preserved epitope-specific differences in compactness and conformational flexibility. This pattern was also supported by the structural alignments of representative neoantigens and by the interface-level differentiation observed across clinically relevant targets.

Attention in CRAFT aligns with putative TCR contact regions

We also examined whether the attention learned by CRAFT was concentrated in residues likely to mediate interaction. In the LLFGYPVYV-associated complex, the global attention heatmap (Fig. 6c) showed that the strongest signal was confined to a limited region connecting the antigenic peptide with the TCR β-chain. Across the β-chain framework region, attention remained weak and diffuse, whereas the CDR3β loop formed a distinct band of higher attention weights. This distribution indicates that model attention during sequence generation was preferentially localized to the variable CDR3β region rather than dispersed across the full β-chain scaffold.

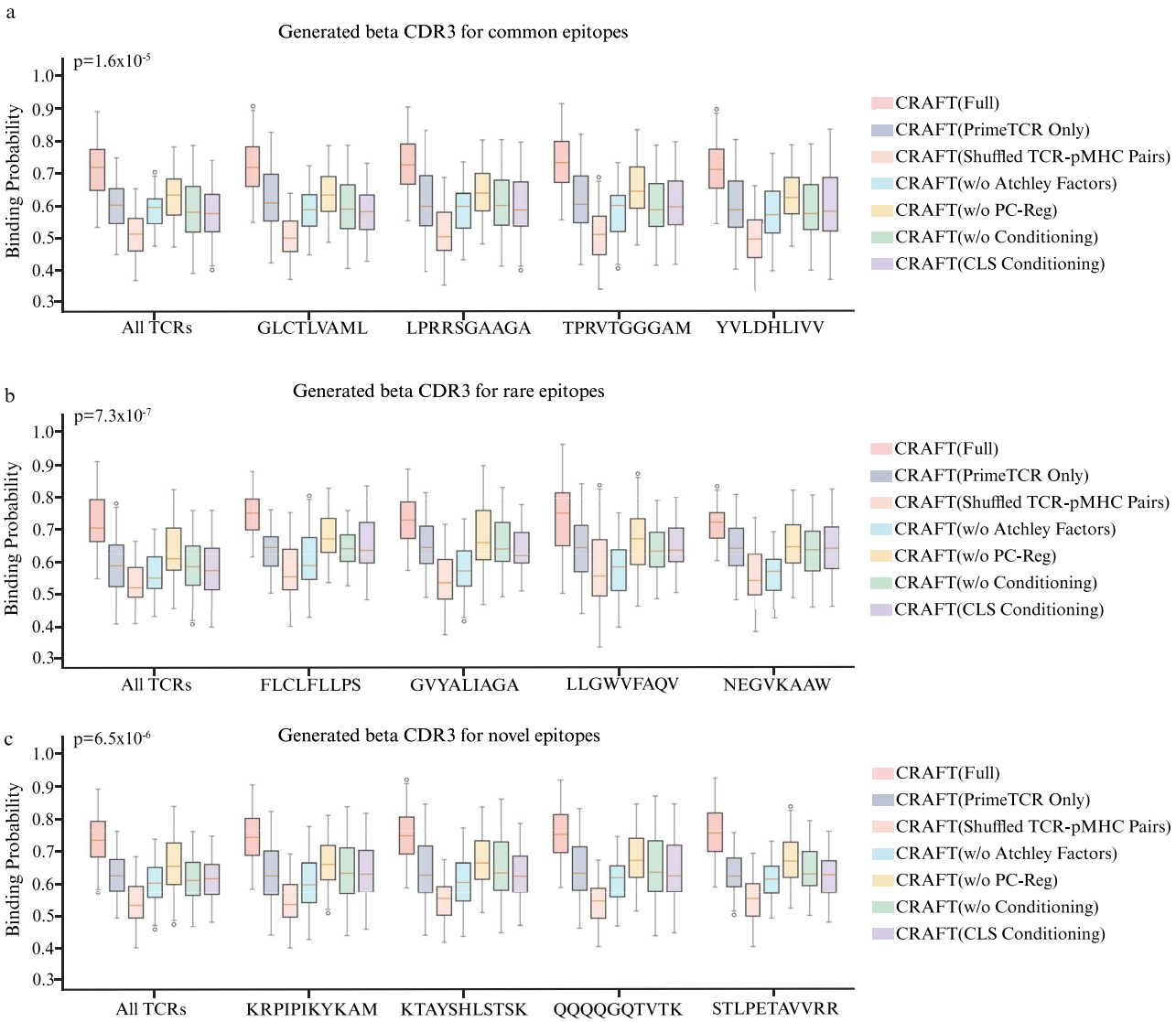


Figure 4. Component-wise ablation analysis of predicted antigen-specific binding by CRAFT. (a) Boxplots of TCRBinder-predicted binding probabilities for CDR3β sequences generated by full CRAFT and the indicated ablation variants in the common-epitope subset. (b) Corresponding comparison of predicted binding probabilities in the rare-epitope subset. (c) Corresponding comparison in the novel-epitope subset. In each panel, "All TCRs" represents the aggregate distribution across all evaluated epitopes within the corresponding frequency regime, whereas the remaining groups show distributions for four representative epitopes. Full CRAFT is compared with a PrimeTCR-only model, a shuffled TCR–pMHC-pairing control, a variant without Atchley-factor inputs, a variant without physicochemical regularization, an unconditional variant, and a variant using pooled CLS-based conditioning. Higher values correspond to higher antigen-specific binding probabilities predicted by TCRBinder. The P values shown in panels (a–c) are the largest P value obtained from Welch's t-tests comparing the full CRAFT model pairwise with the corresponding ablation variants. If the largest P value is statistically significant, all smaller P values are also statistically significant.

The residue interaction matrix centered on the CDR3β loop (Fig. 6d) contained several localized high-attention regions, in which defined antigen-side positions corresponded to specific receptor residues. In the predicted 3D structure (Fig. 6e), these regions partially overlapped with clustered putative interface contacts, linking the 1D attention pattern to structurally relevant interaction regions in this representative example.

To determine whether this correspondence extended beyond the LLFGYPVYV example, we further analyzed 30 TCR–pMHC complexes with available structures. Residues were classified as high-attention positions using normalized attention thresholds of 0.50, 0.55, and 0.60 and were compared with structurally annotated TCR–pMHC contact residues. At these thresholds, the mean numbers of high-attention residues per complex were 13, 10, and 8, respectively, of which an average of 6, 4, and 2 residues overlapped with the contact regions. The corresponding contact proportions among high-attention residues were 46.2%, 40.3%, and 25.0%, respectively. These results indicate that high-attention residues were partially enriched at the TCR–pMHC interface, particularly at the moderate thresholds of 0.50 and 0.55. However, the reduced correspondence at the more stringent threshold suggests that attention is not an exclusive proxy for direct physical contacts and may additionally reflect sequence context, conformational support, or longer-range dependencies. Overall, the attention analysis provides supportive, rather than causal, evidence

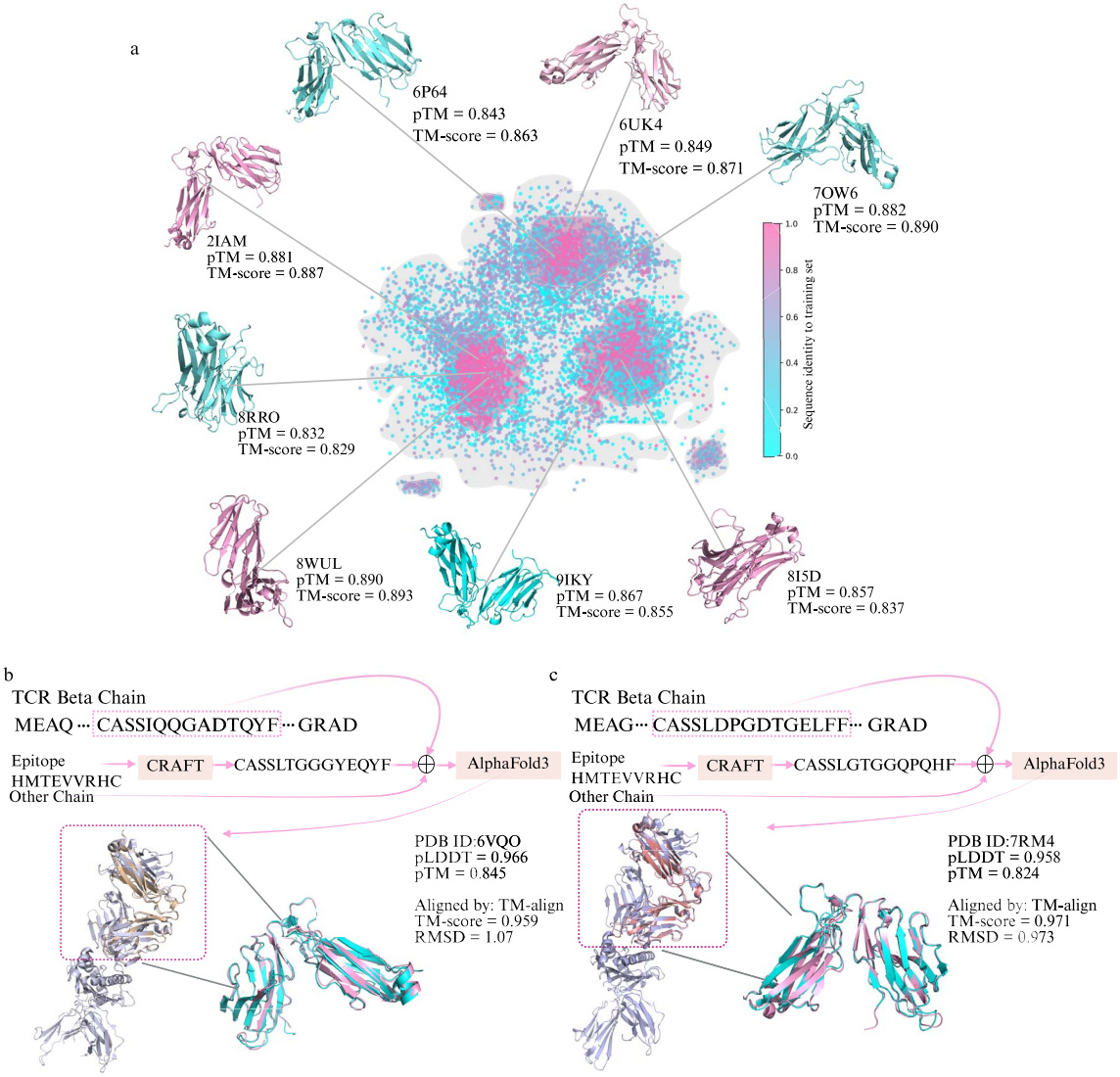


Figure 5. Structural diversity and predicted 3D modeling of CRAFT-generated TCRs. (a) UMAP projection of CRAFT-generated CDR3β sequences, with sequence identity to the training set indicated by the accompanying scale. Representative predicted structures across different identity ranges are highlighted, with the corresponding pTM (from ESM3) and TM-scores (computed by TM-align) indicated. (b) Structural analysis of a CRAFT-generated CDR3β loop targeting the HMTEVVRHC epitope (PDB ID: 6VQO). The generated sequence was integrated into a full-length TCR β-chain and modeled using AlphaFold3. Superimposition with the reference complex reveals high conformational consistency, as evidenced by a high TM-score and low RMSD (Å). (c) Structural modeling and evaluation for another representative complex of the HMTEVVRHC epitope (PDB ID: 7RM4). The CRAFT-generated CDR3β loop is shown superimposed on the reference structure.

that CRAFT captures structurally relevant patterns during antigen-conditioned sequence generation.

Discussion

In this study, we presented CRAFT, a computational framework for the antigen-conditioned *de novo* generation of specific TCR CDR3β candidates. By integrating self-supervised pretraining on large-scale unlabeled immune repertoires with targeted fine-tuning on paired antigen–TCR data, CRAFT showed favorable performance across multiple benchmark settings. Specifically, our evaluations showed consistent advantages in biochemical conformity, specificity, sequence-level fidelity under rare-epitope settings, and predicted structural plausibility in neoantigen-related settings. Importantly, CRAFT should not be interpreted as uniformly outperforming TcrDesign or TCR-T5 across all measures of sequence naturalness. Rather, it achieves comparable native-like sequence fidelity while maintaining competitive predicted local structural quality within a pMHC-conditioned and physicochemically guided generation framework. Its main advantage therefore lies in balancing conditional specificity, candidate diversity, and structural plausibility, rather than maximizing every individual evaluation metric. This work suggests that automated TCR generation can move beyond 1D sequence resemblance toward structurally informed, antigen-conditioned candidate generation.

The balanced performance of CRAFT across these evaluation dimensions is consistent with its overall design strategy, although the current evidence is derived primarily from sequence-level

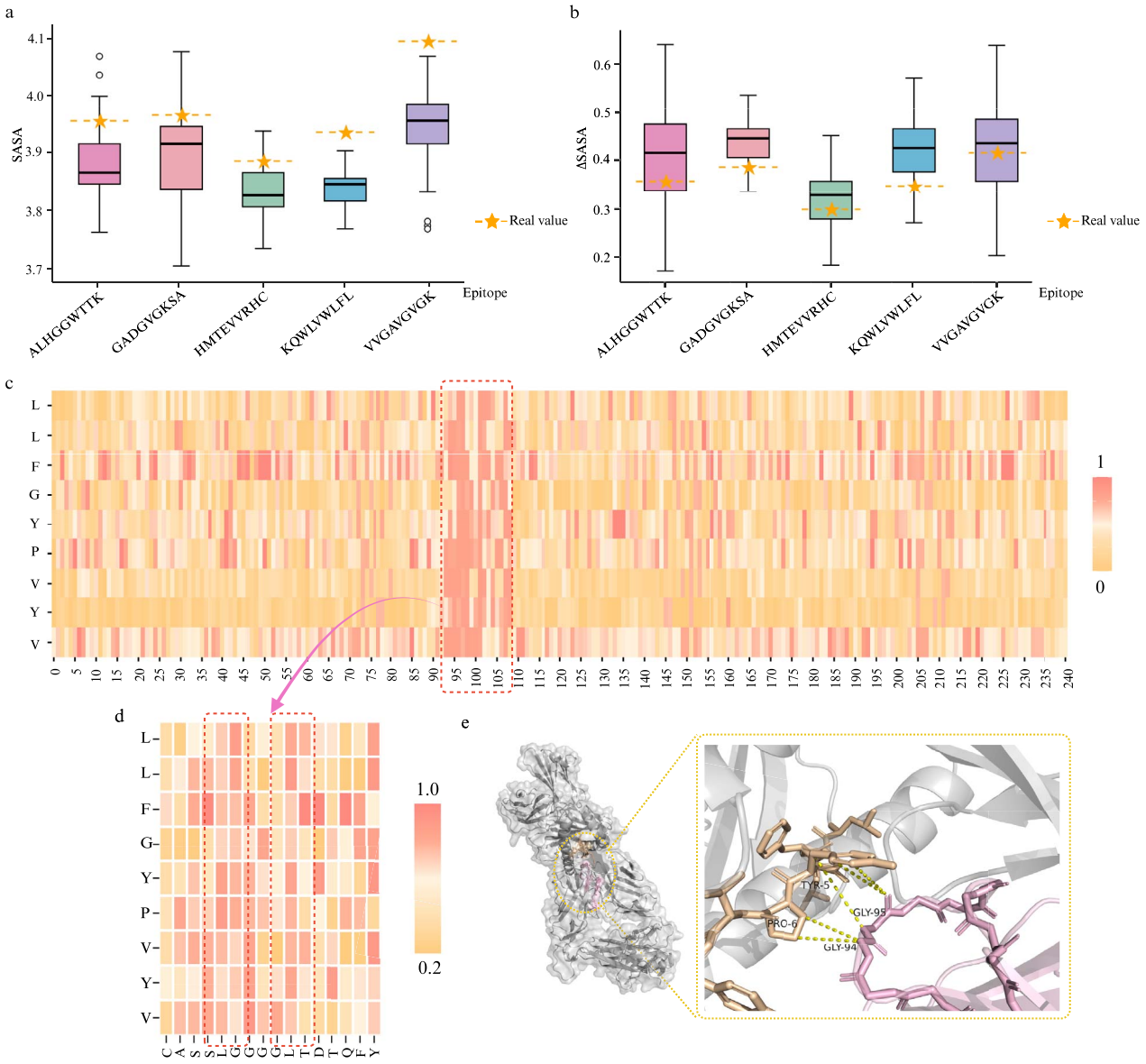


Figure 6. Biophysical property analysis and interaction-related interpretability of CRAFT. (a) Boxplots of SASA (Å)^2 calculated from the predicted structures of CRAFT-generated CDR3β candidates targeting five representative epitopes. Each boxplot represents the distribution of candidates generated for one epitope. The dashed lines and stars indicate the SASA values calculated from the corresponding experimentally resolved reference TCR–pMHC complexes. The reference values generally fall within or close to the distributions of the generated candidates. (b) Corresponding boxplots of the change in SASA upon binding (ΔSASA; Å^2) for CRAFT-generated candidates targeting the same five epitopes. The dashed lines and stars indicate the values calculated from the corresponding experimentally resolved reference complexes. The generated candidates show epitope-dependent distributions of interface burial relative to the corresponding reference values. For visualization, the raw SASA values in panel (a) and the raw ΔSASA values in panel (b) were divided by 10 000 and 1000, respectively. (c) Heatmap showing cross-attention weights between the epitope "LLFGYPVYV" (y-axis) and the full-length TCR β-chain containing the CRAFT-generated CDR3 beta sequence (x-axis). The dashed box highlights the CDR3β loop region, illustrating the concentration of attention signals within the putative interface-related region. (d) Enlarged heatmap displaying cross-attention values between the epitope "LLFGYPVYV" and residues within the CDR3β loop, with dashed boxes highlighting localized high-attention regions. (e) Visualization of the predicted TCR–pMHC interaction interface and putative residue contacts corresponding to the high-attention regions.

metrics and predicted structural analyses. Specifically, the functional separation between PrimeBERT for epitope-context representation and PrimeTCR for receptor generation allowed antigen-side cues to be integrated with the generative priors learned from natural TCR repertoires. This decoupled architecture accounts for the stable sequence behavior observed even in rare-epitope settings. Furthermore, the cross-attention-based conditioning scheme is consistent with the condition-dependent separation and the exploration of nearby sequence neighborhoods observed in the generated candidates. Moreover, the explicit physicochemical constraints introduced via Atchley factors may contribute to the high local conformational quality and the predicted specific interface-related geometric patterns observed in the generated complexes. Rather than claiming complete proof of the underlying biological mechanisms, we interpret these

architectural choices as a plausible explanation for the model's observed computational performance.

CRAFT showed favorable results in sequence-level fidelity, sequence specificity, generalization capacity, and predicted structural plausibility, but additional evaluation dimensions still need to be explored. Importantly, the structural, interface-related, and attention-based analyses in this study are computational rather than direct validation of pMHC–TCR binding or T-cell function. Experimental binding and functional assays and, where feasible, structural characterization will therefore be required to validate these predictions. At present, the framework is centered on TCR CDR3β sequence generation. A next step will be to extend it to full heterodimer design by combining hierarchical conditional decoding with structure-based co-generation so that α/β inter-chain coordination can be modeled within the same design process. We also plan to connect this computational framework with comprehensive *in vitro* and *ex vivo* validation in a closed iterative pipeline. Such a scheme will allow functional assessment of prioritized candidates and will provide a basis for subsequent preclinical evaluation.

CRAFT serves as a computational framework for antigen-specific *de novo* generation of TCR CDR3β candidates. In sequence-based and prediction-based analyses, the model performed well in sequence fidelity, specificity, conditional separation, and predicted structural plausibility, which supported candidate generation, prioritization, and structure-guided follow-up assessment. The value of this study lies in bringing sequence generation, conditional control, and structure-based evaluation into a single workflow that links raw candidate production with downstream filtering. With the continued growth of large-scale experimental validation data, future work can use this iterative modeling strategy to investigate wider translational uses and to probe deeper immunological mechanisms [73].

Materials and methods

Datasets and preprocessing

Pretraining datasets

Self-supervised pretraining was based on two independent unlabeled sequence collections. One collection was used for antigen epitope encoding, and the other was used for modeling TCR CDR3β sequences. Epitope records were gathered from the Immune Epitope Database (IEDB) [74], yielding >3 million unique sequences for PrimeBERT pretraining. Natural TCR CDR3β sequences were obtained from TCRdb 2.0 [75], which provided >4.5 million entries for autoregressive pretraining of PrimeTCR. These datasets were used to initialize the encoder and decoder modules. In the preprocessing step, only complete epitope records containing standard amino acid characters were kept, and repeated entries were removed to generate the final antigen corpus. No global deduplication was performed for the TCRdb 2.0 CDR3β dataset, which allowed the original repertoire frequency distribution to remain represented in the training data.

TCR–pMHC pairing dataset

A supervised TCR–pMHC paired dataset was compiled from several public databases, including McPAS-TCR [76], VDJdb [77], TRAIT [78], MIRA [79], IEDB [74], and STCRDab [80]. These resources provided epitope peptides together with matched TCR sequences and were used to form supervised inputs for positive-pair conditional generation. Records from STCRDab were kept in the integrated dataset because this resource contains experimentally supported antigen–TCR binding entries and could serve as a structural reference. All retrieved entries were first converted to a common sequence format and then organized into paired samples. After this harmonization step, each sample contained an epitope peptide, a paired TCR CDR3β sequence, and the corresponding observed antigen–TCR association.

The integrated dataset was then cleaned and standardized. Malformed, incomplete, and unparsable records were excluded. Noncanonical or ambiguous residue symbols that could still be parsed, including "-," "#," "∗," and "X," were converted to a single unknown token ([UNK]). All epitope and CDR3β sequences were truncated or padded to a maximum length of 30 amino acids so that input tensors had consistent dimensions. In CDR3β sequences, the conserved terminal residues were kept unchanged, namely the N-terminal cysteine (C) and the C-terminal phenylalanine (F) or tryptophan (W), in order to preserve the boundary features that characterize naturally recombined loops. After dataset merging, deduplication, and standardization, the final collection comprised >130 000 supervised training pairs and 24 000 independent validation pairs.

The training and validation sets were segregated at the epitope level, ensuring that all peptides present in the validation set were excluded from model training. To facilitate subsequent generalization benchmarking, the independent validation set was further partitioned into three mutually exclusive subsets according to the number of known paired CDR3β sequences available for each epitope: common, rare, and novel. Specifically, the common subset comprised unseen epitopes with relatively extensive paired-sequence support (> 20 pairs), the rare subset contained unseen epitopes with intermediate support (10–20 pairs), and the novel subset denoted unseen low-resource epitopes with limited paired-sequence support (< 10 pairs), including selected neoantigen-related contexts where applicable. Crucially, while these subsets represent different levels of biological sequence abundance, all of them remained completely unobserved during the training phase. This partition strategy enabled an epitope-level zero-shot evaluation of model generalization across unseen antigenic contexts with different levels of paired-sequence support, ranging from relatively well-supported to low-resource settings.

pMHC context representation

In this study, we incorporated the MHC presentation background alongside the epitope to define a pMHC context for epitope–receptor pairing. This formulation allowed the conditional input to represent antigen identity together with its associated presentation context during CDR3β generation. By integrating allele-specific sequences across diverse human leukocyte antigen subtypes, the model accommodated highly heterogeneous presentation contexts and enabled conditional epitope-receptor learning within a multi-allelic landscape. This configuration allowed the network to effectively capture the mappings between epitopes and CDR3β sequences across varied antigenic backgrounds.

At the sequence-representation level, the target CDR3β sequence was modeled as a conditional autoregressive process. Given a conditional input c, the model generated the receptor sequence from left to right, where c consisted of the epitope sequence and the corresponding pMHC context. Antigen identity and presentation context were combined into one conditional input and used for PrimeBERT encoding as well as the following conditional decoding step. The target sequence was expressed as standard amino acid residues and

generated autoregressively from the sequence prefix available at each decoding step.

Protein sequence representation was implemented using a deterministic residue-fragment tokenizer rather than a learned subword method such as BPE or WordPiece. Sequences were segmented from left to right using a longest-match strategy that preferentially selected tripeptide, dipeptide, and single-residue tokens. Each vocabulary contained 8426 tokens covering canonical residue fragments and special tokens. Peptide-side inputs and CDR3β sequences shared the same fragment inventory and segmentation rule but used separate token-to-index mappings and embedding spaces. For pMHC representation, the MHC pseudo-sequence was encoded residue by residue and concatenated before the fragment-tokenized epitope sequence. In parallel, amino acid residues were mapped to continuous physicochemical vectors using 5D Atchley factors. The resulting discrete pMHC representation and physicochemical features jointly defined the antigen-side conditional input for CDR3β generation. Details of vocabulary construction, special tokens, input serialization, and uncommon-residue handling are provided in Supplementary Note 2.

Model architecture of CRAFT

Technical implementation of the CRAFT architecture

CRAFT adopts an encoder–decoder architecture for antigen-conditioned generation of TCR CDR3β sequences. Its design includes two modules with separate roles. One module encodes epitope sequences and their presentation context on the antigen side. The other generates TCR CDR3β sequences in an autoregressive manner on the receptor side. Within this information flow, antigen-side conditions are initially transformed into contextual embeddings by the encoder. These embeddings are subsequently injected into the decoder via a conditioning mechanism, merging with receptor-side autoregressive states during generation. Structurally, this pipeline advances directly from antigen representation through conditioned integration into receptor generation. At the methodological level, CRAFT embeds antigen-derived cues directly into the sequence generation phase, thereby reducing reliance on extensive post hoc filtering after unconditional language modeling. The following subsections detail the implementations of PrimeBERT, PrimeTCR, and the cross-attention-based conditioned decoding.

PrimeBERT model

Operating as the antigen-side encoder within CRAFT, PrimeBERT processes a composite sequence comprising an MHC pseudo-sequence and an epitope peptide to generate a contextual peptide–MHC (pMHC) representation for downstream conditional decoding. Formally, let the concatenated input be defined as $x = [[\mathrm{CLS}], x_1, x_2, \ldots, x_L, [\mathrm{SEP}]]$, where x_i represents the sequential amino acid residues of the MHC pseudo-sequence followed directly by the epitope, and L denotes their combined length. For each position, the model constructs an input representation e_i comprising a token embedding, a positional embedding, and a projected Atchley factor term:

$$e_i = \mathrm{tok}(x_i) + \mathrm{pos}(i) + Wa_i \tag{1}$$

where $a_i \in \mathbb{R}^5$ defines the 5D Atchley factor vector corresponding to residue x_i, while W represents a learnable projection matrix mapping this vector into the hidden dimensional space of the model. Through this additive formulation, PrimeBERT explicitly embeds residue-level physicochemical attributes alongside sequence identity and positional information, yielding a unified pMHC input representation.

During encoding, PrimeBERT processes the resulting composite sequence through a bidirectional transformer architecture. Following propagation through the stacked self-attention layers, the encoder yields the hidden state H_{enc}, which functions as a contextual representation of the entire pMHC microenvironment. This hidden state encodes local residue information, positional dependencies, and physicochemical features introduced by the Atchley factors. These outputs subsequently provide the explicit antigen-side condition for downstream cross-attention mechanisms within the decoder. Functionally, PrimeBERT governs the specific pMHC representation learning, whereas the downstream decoding module executes the autoregressive generation of TCR CDR3β sequences.

PrimeTCR model

Functioning as the receptor-side decoder within CRAFT, PrimeTCR defines the autoregressive generation distribution for TCR CDR3 β sequences. Mathematically, let the target receptor sequence be defined as $y = [y_1, y_2, \ldots, y_T]$, where y_t represents the amino acid residue at position t, with T indicating the sequence length. PrimeTCR employs a decoder-only transformer architecture optimized via a causal language modeling (CLM) objective, thereby factorizing the full-sequence probability into a product of conditional next-token probabilities:

$$p(y) = \prod_{t=1}^{T} p(y_t \mid y_{<t}) \tag{2}$$

Under this formulation, the model conditions the prediction of residue y_t upon the previously generated prefix $y_{<t}$, establishing a standard left-to-right autoregressive generation process. In this paradigm, the corresponding training objective minimizes the autoregressive negative log-likelihood:

$$\mathcal{L}_{\mathrm{CLM}} = -\sum_{t=1}^{T} \log p(y_t \mid y_{<t}) \tag{3}$$

PrimeTCR models the receptor-side autoregressive probability distribution before antigen-side context is introduced into the generative pipeline.

Cross-attention-based conditioned decoder integration

Within CRAFT, the conditioned decoder emerges by fusing the receptor-side autoregressive PrimeTCR backbone with cross-attention sublayers receiving the explicit antigen-side context. Operationally, pretrained parameters from PrimeTCR were utilized to explicitly initialize the masked self-attention backbone of the decoder, whereas cross-attention sublayers were introduced as randomly initialized modules within the transformer stack and optimized during conditional training. In this design, receptor-side autoregressive generation and antigen-side conditioning are incorporated into one decoding process. PrimeTCR-derived generative priors are preserved on the receptor side, and encoder-derived contextual information is introduced on the antigen side during each step of decoding.

At each decoding step, PrimeTCR provides the causal hidden state H_{dec}, which functions as the Query. PrimeBERT provides the antigen-side contextual representation H_{enc}, which is used as both the Key and

the Value. The cross-attention sublayer combines the current decoding state with the antigen-side condition and generates the aggregated conditional representation H_{inject}. This operation is computed as follows:

$$\text{Attention}(Q, K, V, \text{mask}) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} + \text{mask}\right)V \tag{4}$$

where $Q = H_{\text{dec}}$, $K = H_{\text{enc}}$, and $V = H_{\text{enc}}$. This mechanism ensures decoder-side causal constraints remain preserved during generation, while simultaneously incorporating antigen-side contextual information into every decoding step via cross-attention.

A residual fusion step was applied before the output projection layer to integrate the conditional representation into the final vocabulary prediction. The aggregated conditional state H_{inject} is added directly to the causal hidden state H_{dec}. This fused representation then passes through a linear layer, yielding the exact output logits:

$$Y_{\text{logit}} = \text{Linear}(H_{\text{inject}} + H_{\text{dec}}) \tag{5}$$

Here, H_{inject} signifies the conditionally aggregated representation produced via cross-attention, whereas H_{dec} represents the current autoregressive decoding state. At the current position, the fused representation is used to derive the conditional vocabulary distribution that governs the next generated token. This additive design keeps the autoregressive PrimeTCR backbone intact and introduces explicit antigen-side context into each token-level prediction.

Training strategy

Pretraining strategy

CRAFT underwent a rigorous two-stage training procedure comprising initial self-supervised pretraining followed by conditional fine-tuning. During the first stage, distinct unlabeled antigen-side and receptor-side corpora explicitly initialized the network's encoder and decoder, respectively. During the second stage, paired antigen–TCR data strictly drove the supervised conditional optimization (Supplementary Table S1). This sequential design ensured pretraining established fundamental unimodal initializations across both network branches, whereas conditional fine-tuning provided antigen-dependent supervision driving targeted end-to-end parameter updating.

During initial pretraining, PrimeBERT and PrimeTCR underwent strictly independent optimization. PrimeBERT employed a masked language modeling (MLM) objective. In this stage, amino acid residues within the input epitope sequence were randomly masked at a probability of 15%, and the model was trained to reconstruct the masked positions using the remaining sequence context. The cross-entropy loss is defined as follows:

$$\mathcal{L}_{\text{MLM}} = -\sum_{m \in M} \log p(x_m \mid x_{\sim M}; \theta_{\text{enc}}) \tag{6}$$

where M defines the set of masked positions while θ_{enc} represents the encoder parameters. This stage was conducted exclusively on the antigen-side pretraining corpus and established encoder-side representation learning.

PrimeTCR was trained with a CLM objective on the TCR CDR3β sequences. For a target sequence $y = [y_1, \ldots, y_T]$, the training loss is defined as

$$\mathcal{L}_{\text{CLM}} = -\sum_{t=1}^{T} \log p(y_t \mid y_{<t}; \theta_{\text{dec}}) \tag{7}$$

where θ_{dec} denotes the decoder parameters. This phase relied entirely upon unlabeled receptor-side sequences without introducing any antigen-side conditioning, effectively establishing the fundamental unconditional autoregressive basis of the decoder. Post-training, the resulting PrimeBERT and PrimeTCR parameters directly initialized their respective encoder and decoder components, priming the subsequent conditional fine-tuning stage.

Fine-tuning

Following pretraining initialization, CRAFT advanced to a supervised fine-tuning stage utilizing antigen–TCR paired samples. A selective parameter-updating strategy maintained the stability of learned representations given the strictly limited scale of paired data, simultaneously concentrating the majority of trainable capacity within specific modules directly executing the conditional sequence generation. Within the encoder, only a subset of the pretrained parameters was updated during this phase, whereas the PrimeTCR decoding backbone, the conditional cross-attention sublayers, and the final vocabulary projection layer remained fully subject to standard backpropagation and parameter optimization. Under this targeted allocation, optimization was confined to modules involved in conditional alignment, decoder-state updating, and adaptation of the output distribution. This allowed supervised adaptation for precise immunogenic generation while keeping the pretrained network initialization unchanged.

The training loss comprised the standard autoregressive sequence-generation term $\mathcal{L}_{\text{gen}}$ and an Atchley-factor-based physicochemical regularization term $\mathcal{L}_{\text{phy}}$:

$$\mathcal{L} = \mathcal{L}_{\text{gen}} + \lambda \mathcal{L}_{\text{phy}} \tag{8}$$

At each decoding step, the predicted token distribution was obtained by applying softmax to the decoder logits and was projected into the 5D Atchley space through a fixed vocabulary-level Atchley lookup matrix. These projected vectors were mean pooled across non-padding decoding positions to obtain a differentiable physicochemical representation of the predicted TCR sequence, denoted by $\bar{\mathbf{a}}^{(i)}_{\text{TCR}}$ for sample i. The corresponding reference representation, $\bar{\mathbf{a}}^{(i)}_{\text{epi}}$, was defined as the mean Atchley vector of the epitope residues in the same sample. The pMHC sequence was incorporated through the encoder and cross-attention pathway, whereas the physicochemical target used by $\mathcal{L}_{\text{phy}}$ was calculated from the epitope peptide alone.

For each mini-batch of size B, the matched pair $\left(\bar{\mathbf{a}}^{(i)}_{\text{TCR}}, \bar{\mathbf{a}}^{(i)}_{\text{epi}}\right)$ was treated as the positive pair. One randomly permuted in-batch epitope representation, indexed by $\pi(i)$, was used as the control pair for each anchor. We used cosine similarity and a triplet-hinge formulation:

$$\mathcal{L}_{\text{phy}} = \frac{1}{B}\sum_{i=1}^{B}\left[\cos\left(\bar{\mathbf{a}}^{(i)}_{\text{TCR}}, \bar{\mathbf{a}}^{(\pi(i))}_{\text{epi}}\right) - \cos\left(\bar{\mathbf{a}}^{(i)}_{\text{TCR}}, \bar{\mathbf{a}}^{(i)}_{\text{epi}}\right) + \gamma\right]_+ \tag{9}$$

where $[\cdot]_+ = \mathbf{max}(0, \cdot)$, $\gamma = 0.1$, and π was independently resampled for each mini-batch. No temperature parameter was used in

$\mathcal{L}_{\text{phy}}$; the decoding temperature of 0.5 was used only during inference. We selected $\lambda = 0.5$ based on the sensitivity sweep over $\lambda \in \{0, 0.1, 0.5, 1.0, 2.0\}$. Complete details of the Atchley lookup matrix, token handling, masked pooling, and in-batch permutation procedure are provided in Supplementary Note 3.

Based on the sensitivity sweep over $\lambda \in \{0, 0.1, 0.5, 1.0, 2.0\}$ conducted on the held-out validation set, we used $\lambda = 0.5$ in all main experiments because it provided the most balanced overall performance across the evaluated metrics. During backpropagation, three components were optimized in parallel: the trainable subset of the pretrained encoder parameters, the pretrained decoder parameters, and the randomly initialized conditional module.

CRAFT was implemented in PyTorch and trained using two NVIDIA RTX 6000 GPUs with 24 GB of memory per GPU. PrimeBERT and PrimeTCR pretraining required ∼2 and 5 days, respectively, whereas CRAFT fine-tuning for 20 epochs required ∼30 h. Generating and filtering 1000 CDR3β candidates for each epitope required ∼4 min on average. Model parameter counts are provided in Supplementary Table S14, whereas the computational configuration, runtime, and inference procedures are described in Supplementary Note 4.

Inference and filtering of generated sequences

At inference time, target CDR3β sequences were generated with autoregressive beam search using a beam width of 50. Under this setting, multiple candidate paths were retained at each decoding step. Log-probability ranking was then used to order candidate sequences for downstream selection. The number of raw outputs was specified in advance for each type of analysis. For structural case evaluation and cross-epitope interface statistics, 1000 candidate sequences were generated for each queried antigen or epitope. For latent-space analysis, >20 000 sequences were generated and projected onto a UMAP space constructed from antigen–TCR binding pairs in the training set. All raw outputs were generated before post-processing, and no dynamic replenishment of valid sequences was introduced during decoding.

Post-processing was carried out in a fixed sequence. Within each antigen, the raw sequence pool was first screened for invalid characters and duplicate outputs. Sequences containing noncanonical, corrupted, or unresolved characters were removed, and exact-match duplicates generated for the same antigen were collapsed into a single entry. The remaining sequences were then filtered by endpoint and length requirements, so that only outputs with the expected anchoring pattern of an N-terminal cysteine and a C-terminal phenylalanine or tryptophan were kept. Beam rank was preserved after filtering. For structure-based analyses with a raw generation budget of n sequences (n denotes the user-specified number of raw sequences to be generated and can be set to any positive integer, such as 1 or 1000), all downstream calculations were performed on the resulting valid set. If more than n valid sequences remained, the top n by beam rank were retained. If fewer than 100 passed filtering, all valid sequences were used. The number of sequences entering structural analysis was therefore determined by the fixed raw generation budget together with the same filtering scheme, and no extra resampling was used to force a constant final size. The retained counts for each antigen or epitope are provided separately in the Supplementary Materials to maintain correspondence with the downstream structural and interface analyses.

The finalized sequence sets were then carried forward to downstream analyses. For practical prioritization toward experimental validation, valid candidates were assessed using a staged multi-criterion workflow rather than a single unvalidated composite score. Sequence novelty relative to the training corpus and structural confidence were first used to identify credible candidates. Among candidates satisfying these criteria, predicted binding probability was used as the primary ranking criterion, whereas similarity to natural TCR sequences served as a plausibility check rather than a quantity to be maximized. For the highest-ranked candidates, structural descriptors such as SASA and ΔSASA, together with sequence-level metrics such as NSR, were used as complementary secondary criteria to assess interface-related structural properties and sequence quality. This prioritization workflow was intended to nominate candidates for subsequent experimental validation and was distinct from the decoding rank used to construct the analysis-specific sequence sets.

For structure-oriented analysis, the filtered valid outputs directly supported subsequent TCR–pMHC structural prediction and representative structure comparison. For latent-space analysis, the expanded sequence pool was first mapped onto the UMAP space constructed from antigen–TCR pair representations in the training set. Representative sequences were then selected from distinct clustering regions and used for downstream evaluation of structural confidence. The same front-end inference procedure and basic post-processing rules were applied across downstream tasks, whereas candidate prioritization, structural prediction, and latent-space analysis were performed according to the requirements of each downstream analysis.

Evaluation overview

The comprehensive model evaluation proceeded across the sequence, structure, and representation levels. Sequence-level analyses assessed naturalness, reference-sequence fidelity, training-set novelty, and conditional separation of the generated outputs. Similarity to natural or reference sequences was used to characterize sequence fidelity and biological plausibility, whereas novelty was evaluated independently by comparing each generated CDR3β sequence with its nearest neighbor in the training set. Specifically, novelty at a sequence-identity threshold τ was defined as the proportion of generated sequences whose maximum identity to any training-set CDR3β sequence was below τ, with $\tau \in \{0.40, 0.50, 0.60, 0.70\}$. The corresponding novelty rates were 0.622, 0.713, 0.805, and 0.884, respectively. These results indicate that a substantial proportion of the generated sequences remained distinct from the training data across multiple sequence-identity thresholds, providing evidence against simple reproduction of training examples and supporting the sequence-level generalization ability of CRAFT. Structure-level evaluations examined predicted structural plausibility, interface-related features, and the behavior of representative complexes. Representation-level analyses included both latent-space projections and attention-pattern inspections. Detailed definitions of all evaluation metrics are provided in Supplementary Note 5.

Key Points

- CRAFT is a decoupled generative framework that enables the *de novo* design of target-specific TCR CDR3β sequences conditioned on complex pMHC contexts.
- The CRAFT framework seamlessly couples the PrimeBERT encoder and PrimeTCR decoder via cross-attention, injecting antigen-side

contextual information and using residual fusion to precisely integrate the pMHC context while strictly preserving native receptor generative priors.
- CRAFT incorporates Atchley-factor-based physicochemical constraints into the generation process, enabling interfacial property-related information to inform sequence design.
- CRAFT's ability to generate high-fidelity receptors strongly supports the prioritization of neoantigen-guided TCR candidates, thereby advancing the application of precision immunotherapy.

Acknowledgements

The authors thank the reviewers for their valuable comments and suggestions.

Author contributions

Xiaochuan Yang (Conceptualization, Formal analysis, Methodology, Software, Writing—original draft, Writing—review & editing), Zaiduo Li (Formal analysis, Methodology, Software, Writing—original draft, Writing—review & editing), Zhijuan Li (Formal analysis, Methodology, Software, Writing—original draft, Writing—review & editing), Wei Zhang (Formal analysis, Methodology, Software, Writing—original draft, Writing—review & editing), Xianyu Zhang (Conceptualization, Formal analysis, Methodology, Software, Writing—review & editing), Tiansong Yang (Conceptualization, Formal analysis, Methodology, Writing—review & editing), Mingliang Liu (Formal analysis, Methodology, Software, Writing—original draft, Writing—review & editing), and Xiaokun Li (Conceptualization, Formal analysis, Methodology, Writing—review & editing)

Supplementary material

Supplementary material is available at *Briefings in Bioinformatics* online.

Competing interests

The authors declare no conflicts of interest.

Funding

This study was supported by the Natural Science Foundation of Heilongjiang Province, China (No. PL2025F040), and the Double First Class Discipline Collaborative Innovation Foundation of Heilongjiang Province of China (No. LJGXCG2023-54).

Data availability

All the datasets used in this study are publicly available. IEDB: https://www.iedb.org/home_v3.php; TCRdb 2.0: https://guolab.wchscu.cn/TCRdb2/; McPAS-TCR: https://friedmanlab.weizmann.ac.il/McPAS-TCR/; VDJdb: https://vdjdb.com/; TRAIT: https://pgx.zju.edu.cn/traitdb/; MIRA: https://www.cancerrxgene.org/downloads/bulk_download; STCRDab: https://opig.stats.ox.ac.uk/webapps/stcrdab-stcrpred/. The code is available at https://github.com/fly9090/CRAFT.

References

1. Musher BL, Vasileiou S, Smaglo BG *et al.* Autologous multiantigen-targeted T-cell therapy for pancreatic cancer: a phase 1/2 trial. *Nat Med* 2026;**32**:258–69. https://doi.org/10.1038/s41591-025-04043-5
2. Ansari MS, Chauhan V, Singh A *et al.* AI-guided CAR designs and targeted pathway modulation to enhance multi-antigen CAR T cell durability and overcome antigen escape. *Nat Commun* 2026;**17**:1549. https://doi.org/10.1038/s41467-025-68272-5
3. Cabezas-Caballero J, Huhn A, Kutuzov MA *et al.* Generation of T cells with reduced off-target cross-reactivities by engineering co-signalling receptors. *Nat Biomed Eng* 2026;**10**:753–64. https://doi.org/10.1038/s41551-025-01563-w
4. Humer D, Klepsch V, Rieder D *et al.* NR2F6 deletion revives CAR-T cell function and induces antigen-agnostic immune memory in solid tumors. *Nat Commun* 2026;**17**:3264. https://doi.org/10.1038/s41467-026-69796-0
5. Nguyen TT, Ho P, Staudt S *et al.* Fine tuning towards the next generation of engineered T cells. *Nat Biomed Eng* 2025;**9**:1610–31. https://doi.org/10.1038/s41551-025-01492-8
6. Liu C, Wang Z, Zhang W *et al.* NKG2D CAR-T cells for solid tumor immunotherapy: advances, challenges, and future directions. *Front Immunol* 2026;**17**:1763843. https://doi.org/10.3389/fimmu.2026.1763843
7. Li J, Liu C, Zhang P *et al.* Optimizing CAR T cell therapy for solid tumours: a clinical perspective. *Nat Rev Clin Oncol* 2025;**22**:953–68. https://doi.org/10.1038/s41571-025-01075-1
8. Khan SH, Choi Y, Veena M *et al.* Advances in CAR T cell therapy: antigen selection, modifications, and current trials for solid tumors. *Front Immunol* 2025;**15**:1489827. https://doi.org/10.3389/fimmu.2024.1489827
9. Zhu X, Chen J, Li W *et al.* Hypoxia-responsive CAR-T cells exhibit reduced exhaustion and enhanced efficacy in solid tumors. *Cancer Res* 2024;**84**:84–100. https://doi.org/10.1158/0008-5472.CAN-23-1038
10. Haus-Cohen M, Reiter Y. Harnessing antibody-mediated recognition of the intracellular proteome with T cell receptor-like specificity. *Front Immunol* 2024;**15**:1486721. https://doi.org/10.3389/fimmu.2024.1486721
11. Wang P, Wang Y, Zhao X *et al.* Chimeric antigen receptor with novel intracellular modules improves antitumor performance of T cells. *Signal Transduct Target Ther* 2025;**10**:20. https://doi.org/10.1038/s41392-024-02096-5
12. Klebanoff CA, Chandran SS, Baker BM *et al.* T cell receptor therapeutics: immunological targeting of the intracellular cancer proteome. *Nat Rev Drug Discov* 2023;**22**:996–1017. https://doi.org/10.1038/s41573-023-00809-z
13. Borgers JSW, Lenkala D, Kohler V *et al.* Personalized, autologous neoantigen-specific T cell therapy in metastatic melanoma: a phase 1 trial. *Nat Med* 2025;**31**:881–93. https://doi.org/10.1038/s41591-024-03418-4
14. Braun DA, Moranzoni G, Chea V *et al.* A neoantigen vaccine generates antitumour immunity in renal cell carcinoma. *Nature.* 2025;**639**:474–82. https://doi.org/10.1038/s41586-024-08507-5
15. Shi Y, Strasser A, Green DR *et al.* Legacy of the discovery of the T-cell receptor: 40 years of shaping basic immunology and translational work to develop novel therapies. *Cell Mol Immunol* 2024;**21**:790–7. https://doi.org/10.1038/s41423-024-01168-4
16. Quast NP, Abanades B, Guloglu B *et al.* T-cell receptor structures and predictive models reveal comparable alpha and beta chain structural diversity despite differing genetic complexity. *Commun Biol* 2025;**8**:362. https://doi.org/10.1038/s42003-025-07708-6

17. Heather JM, Spindler MJ, Alonso MH *et al.* Stitchr: stitching coding TCR nucleotide sequences from V/J/CDR3 information. *Nucleic Acids Res* 2022;**50**:e68–8. https://doi.org/10.1093/nar/gkac190
18. Lu T, Zhang Z, Zhu J *et al.* Deep learning-based prediction of the T cell receptor-antigen binding specificity. *Nat Mach Intell* 2021;**3**:864–75. https://doi.org/10.1038/s42256-021-00383-2
19. Li X, Yang Q, Xu L *et al.* TCRdesign: an antigen-specific generative language model for *de novo* design of T-cell receptors. *Brief Bioinform* 2025;**26**:bbaf691. https://doi.org/10.1093/bib/bbaf691
20. Kim WJ, Crosse EI, De Neef E *et al.* Mis-splicing-derived neoantigens and cognate TCRs in splicing factor mutant leukemias. *Cell.* 2025;**188**:3422–3440.e24. https://doi.org/10.1016/j.cell.2025.03.047
21. Jumper J, Evans R, Pritzel A *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature.* 2021;**596**:583–9. https://doi.org/10.1038/s41586-021-03819-2
22. Wang M, Wang Z, Sun H *et al.* Deep learning approaches for *de novo* drug design: an overview. *Curr Opin Struct Biol* 2022;**72**:135–44. https://doi.org/10.1016/j.sbi.2021.10.001
23. Rives A, Meier J, Sercu T *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci* 2021;**118**:e2016239118. https://doi.org/10.1073/pnas.2016239118
24. Lin Z, Akin H, Rao R *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science.* 2023;**379**:1123–30. https://doi.org/10.1126/science.ade2574
25. Madani A, Krause B, Greene ER *et al.* Large language models generate functional protein sequences across diverse families. *Nat Biotechnol* 2023;**41**:1099–106. https://doi.org/10.1038/s41587-022-01618-2
26. Ingraham JB, Baranov M, Costello Z *et al.* Illuminating protein space with a programmable generative model. *Nature.* 2023;**623**:1070–8. https://doi.org/10.1038/s41586-023-06728-8
27. Krishna R, Wang J, Ahern W *et al.* Generalized biomolecular modeling and design with RoseTTAFold all-atom. *Science.* 2024;**384**:eadl2528. https://doi.org/10.1126/science.adl2528
28. Dauparas J, Anishchenko I, Bennett N *et al.* Robust deep learning-based protein sequence design using ProteinMPNN. *Science.* 2022;**378**:49–56. https://doi.org/10.1126/science.add2187
29. Butcher J, Krishna R, Mitra R *et al. De novo* design of all-atom biomolecular interactions with RFdiffusion3. bioRxiv [Preprint]. 2025. 2025.09.18.676967. https://doi.org/10.1101/2025.09.18.676967
30. Goverde CA, Pacesa M, Goldbach N *et al.* Computational design of soluble and functional membrane protein analogues. *Nature.* 2024;**631**:449–58. https://doi.org/10.1038/s41586-024-07601-y
31. Norman RA, Ambrosetti F, Bonvin AMJJ *et al.* Computational approaches to therapeutic antibody design: established methods and emerging trends. *Brief Bioinform* 2020;**21**:1549–67. https://doi.org/10.1093/bib/bbz095
32. Mason DM, Friedensohn S, Weber CR *et al.* Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning. *Nat Biomed Eng* 2021;**5**:600–12. https://doi.org/10.1038/s41551-021-00699-9
33. Nijkamp E, Ruffolo JA, Weinstein EN *et al.* ProGen2: exploring the boundaries of protein language models. *Cell Syst* 2023;**14**:968–978.e3. https://doi.org/10.1016/j.cels.2023.10.002
34. Shuai RW, Ruffolo JA, Gray JJ. IgLM: infilling language modeling for antibody sequence design. *Cell Syst* 2023;**14**:979–989.e4. https://doi.org/10.1016/j.cels.2023.10.001
35. Nabla Bio, Biswas S. *De novo* design of epitope-specific antibodies against soluble and multipass membrane proteins with high specificity, developability, and function. bioRxiv [Preprint]. 2025. 2025.01.21.633066. https://doi.org/10.1101/2025.01.21.633066
36. Bennett NR, Watson JL, Ragotte RJ *et al.* Atomically accurate *de novo* design of antibodies with RFdiffusion. *Nature.* 2026;**649**:183–93. https://doi.org/10.1038/s41586-025-09721-5
37. Huhn A, Nissley D, Wilson DB *et al.* The molecular reach of antibodies crucially underpins their viral neutralisation capacity. *Nat Commun* 2025;**16**:338. https://doi.org/10.1038/s41467-024-54916-5
38. Zareie P, Szeto C, Farenc C *et al.* Canonical T cell receptor docking on peptide-MHC is essential for T cell signaling. *Science.* 2021;**372**:eabe9124. https://doi.org/10.1126/science.abe9124
39. Sušac L, Vuong MT, Thomas C *et al.* Structure of a fully assembled tumor-specific T cell receptor ligated by pMHC. *Cell.* 2022;**185**:3201–3213.e19. https://doi.org/10.1016/j.cell.2022.07.010
40. Notti RQ, Yi F, Heissel S *et al.* The resting and ligand-bound states of the membrane-embedded human T-cell receptor-CD3 complex. *Nat Commun* 2025;**16**:10996. https://doi.org/10.1038/s41467-025-66939-7
41. Ma J, Ayres CM, Brambley CA *et al.* Dynamic allostery in the peptide/MHC complex enables TCR neoantigen selectivity. *Nat Commun* 2025;**16**:849. https://doi.org/10.1038/s41467-025-56004-8
42. Wang Y, Wang Y, Yuan W *et al.* Tuning the sensitivity of mechanosensory receptors through histidine scanning. *Cell.* 2026;**189**:1680–1700.e48. https://doi.org/10.1016/j.cell.2025.12.050
43. Chen X, Mao Z, Kolawole EM *et al.* Overcoming T cell tolerance to tumor self-antigens through catch-bond engineering. *Science.* 2026;**391**:eadx3162. https://doi.org/10.1126/science.adx3162
44. Karthikeyan D, Bennett SN, Reynolds AG *et al.* Conditional generation of real antigen-specific T cell receptor sequences. *Nat Mach Intell* 2025;**7**:1494–509. https://doi.org/10.1038/s42256-025-01096-6
45. Leary AY, Scott D, Gupta NT *et al.* Designing meaningful continuous representations of T cell receptor sequences with deep generative models. *Nat Commun* 2024;**15**:4271. https://doi.org/10.1038/s41467-024-48198-0
46. Zhou Z, Chen J, Lin S *et al.* GRATCR: epitope-specific T cell receptor sequence generation with data-efficient pre-trained models. *IEEE J Biomed Health Inform* 2025;**29**:2271–83. https://doi.org/10.1109/JBHI.2024.3514089
47. Yang J, He B, Zhao Y *et al. De novo* generation of T-cell receptors with desired epitope-binding property by leveraging a pre-trained large language model. bioRxiv [Preprint]. 2023. 2023.10.18.562845. https://doi.org/10.1101/2023.10.18.562845
48. Croce G, Bobisse S, Moreno DL *et al.* Deep learning predictions of TCR-epitope interactions reveal epitope-specific chains in dual alpha T cells. *Nat Commun* 2024;**15**:3211. https://doi.org/10.1038/s41467-024-47461-8
49. Glanville J, Huang H, Nau A *et al.* Identifying specificity groups in the T cell receptor repertoire. *Nature.* 2017;**547**:94–8. https://doi.org/10.1038/nature22976
50. Jensen MF, Nielsen M. Enhancing TCR specificity predictions by combined pan-and peptide-specific training, loss-scaling, and sequence similarity integration. *eLife.* 2024;**12**:RP93934. https://doi.org/10.7554/eLife.93934
51. Vu MH, Robert PA, Akbar R *et al.* Linguistics-based formalization of the antibody language as a basis for antibody language models. *Nat Comput Sci* 2024;**4**:412–22. https://doi.org/10.1038/s43588-024-00642-3
52. Karnaukhov VK, Shcherbinin DS, Chugunov AO *et al.* Structure-based prediction of T cell receptor recognition of unseen epitopes using TCRen. *Nat Comput Sci* 2024;**4**:510–21. https://doi.org/10.1038/s43588-024-00653-0

53. Reynisson B, Alvarez B, Paul S *et al.* NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res* 2020;**48**:W449–54. https://doi.org/10.1093/nar/gkaa379
54. Marzella DF, Crocioni G, Radusinović T *et al.* Geometric deep learning improves generalizability of MHC-bound peptide predictions. *Commun Biol* 2024;**7**:1661. https://doi.org/10.1038/s42003-024-07292-1
55. Aronson A, Hochner T, Cohen T *et al.* Structure modeling and specificity of peptide-MHC class I interactions using geometric deep learning. bioRxiv [Preprint]. 2022. 2022.12.15.520566. https://doi.org/10.1101/2022.12.15.520566
56. Lai B, Englund M, Bharanikumar R *et al.* DecoderTCR: compositional pretraining and entropy-guided decoding for TCR-pMHC interactions. In: Zhang T, Dudik M, Jaggi M *et al.* (eds). *Proceedings of the 43rd International Conference on Machine Learning*. Proceedings of Machine Learning Research, vol. **306**. PMLR, 2026, 62206–62229.
57. Zhang J, Ma W, Yao H. Accurate TCR-pMHC interaction prediction using a BERT-based transfer learning method. *Brief Bioinform* 2024;**25**:bbad436. https://doi.org/10.1093/bib/bbad436
58. Jiang Y, Li SC. Deep autoregressive generative models capture the intrinsics embedded in T-cell receptor repertoires. *Brief Bioinform* 2023;**24**:bbad038. https://doi.org/10.1093/bib/bbad038
59. Meynard-Piganeau B, Feinauer C, Weigt M *et al.* TULIP: a transformer-based unsupervised language model for interacting peptides and T cell receptors that generalizes to unseen epitopes. *Proc Natl Acad Sci* 2024;**121**:e2316401121. https://doi.org/10.1073/pnas.2316401121
60. Ferruz N, Schmidt S, Höcker B. ProtGPT2 is a deep unsupervised language model for protein design. *Nat Commun* 2022;**13**:4348. https://doi.org/10.1038/s41467-022-32007-7
61. Li C, Sutherland D, Richter A *et al. De novo* synthetic antimicrobial peptide design with a recurrent neural network. *Protein Sci* 2024;**33**:e5088. https://doi.org/10.1002/pro.5088
62. Sethna Z, Elhanati Y, Callan CGJr *et al.* OLGA: fast computation of generation probabilities of B- and T-cell receptor amino acid sequences and motifs. *Bioinformatics.* 2019;**35**:2974–81. https://doi.org/10.1093/bioinformatics/btz035
63. Seo SY, Rhee JK. TCR-epiDiff: solving dual challenges of TCR generation and binding prediction. *Bioinformatics.* 2025;**41**:i125–32. https://doi.org/10.1093/bioinformatics/btaf202
64. Diao K, Chen J, Zhao X *et al.* TcrDesign: *de novo* design of epitope-specific full-length T cell receptors. *Sci China Life Sci* 2026;**69**:2865–2880. https://doi.org/10.1007/s11427-025-3232-5
65. Sarkar A, Kang Y, Somia N *et al.* Designing DNA with tunable regulatory activity using score-entropy discrete diffusion. bioRxiv [Preprint]. 2025. 2024.05.23.595630. Version 2. https://doi.org/10.1101/2024.05.23.595630
66. Papineni K, Roukos S, Ward T *et al.* BLEU: a method for automatic evaluation of machine translation. In: Isabelle P, Charniak E, Lin D (eds). *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics 2002; 311–318. https://doi.org/10.3115/1073083.1073135
67. Hayes T, Rao R, Akin H *et al.* Simulating 500 million years of evolution with a language model. *Science.* 2025;**387**:850–8. https://doi.org/10.1126/science.ads0018
68. Hsiue EHC, Wright KM, Douglass J *et al.* Targeting a neoantigen derived from a common TP53 mutation. *Science.* 2021;**371**:eabc8697. https://doi.org/10.1126/science.abc8697
69. Tran E, Robbins PF, Lu YC *et al.* T-cell transfer therapy targeting mutant KRAS in cancer. *N Engl J Med* 2016;**375**:2255–62. https://doi.org/10.1056/NEJMoa1609279
70. Zhu J, Chen Z, Xu X *et al.* Structure guided analysis of KRAS G12 mutants in HLA-A∗ 11:01 reveals a length encoded immunogenic advantage in G12D. *Commun Biol* 2026;**9**:26. https://doi.org/10.1038/s42003-025-09285-0
71. Abramson J, Adler J, Dunger J *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold3. *Nature.* 2024;**630**:493–500. https://doi.org/10.1038/s41586-024-07487-w
72. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res* 2005;**33**:2302–9. https://doi.org/10.1093/nar/gki524
73. Sahin U, Schmidt M, Derhovanessian E *et al.* Individualized mRNA vaccines evoke durable T cell immunity in adjuvant TNBC. *Nature.* 2026;**651**:1088–96. https://doi.org/10.1038/s41586-025-10004-2
74. Vita R, Mahajan S, Overton JA *et al.* The immune epitope database (IEDB): 2018 update. *Nucleic Acids Res* 2019;**47**:D339–43. https://doi.org/10.1093/nar/gky1006
75. Yue T, Chen SY, Shen WK *et al.* TCRdb 2.0: an updated T-cell receptor sequence database. *Nucleic Acids Res* 2026;**54**:D504–10. https://doi.org/10.1093/nar/gkaf876
76. Tickotsky N, Sagiv T, Prilusky J *et al.* McPAS-TCR: a manually curated catalogue of pathology-associated T cell receptor sequences. *Bioinformatics.* 2017;**33**:2924–9. https://doi.org/10.1093/bioinformatics/btx286
77. Shugay M, Bagaev DV, Zvyagin IV *et al.* VDJdb: a curated database of T-cell receptor sequences with known antigen specificity. *Nucleic Acids Res* 2018;**46**:D419–27. https://doi.org/10.1093/nar/gkx760
78. Wei M, Wu J, Bai S *et al.* TRAIT: a comprehensive database for T-cell receptor-antigen interactions. *Genomics Proteomics Bioinformatics* 2025;**23**:qzaf033. https://doi.org/10.1093/gpbjnl/qzaf033
79. Dines JN, Manley TJ, Svejnoha E *et al.* The Immune RACE study: a prospective multicohort study of immune response action to COVID-19 events with the ImmuneCODE open access database. *medRxiv* [Preprint]. 2020. 2020.08.17.20175158. https://doi.org/10.1101/2020.08.17.20175158
80. Leem J, deOliveira SHP, Krawczyk K *et al.* STCRDab: the structural T-cell receptor database. *Nucleic Acids Res* 2018;**46**:D406–12. https://doi.org/10.1093/nar/gkx971

© The Author(s) 2026. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.
Briefings in Bioinformatics, 2026, **27**, bbag616
https://doi.org/10.1093/bib/bbag616
Case Study