EpitopeGen: Learning to Generate T Cell Epitopes: A Semi-Supervised Approach with Biological Constraints

Minuk Ma [1] **Wilson Tu** [1] **Carlos Vasquez-Rios** [1] **Jiarui Ding** [1]

Abstract

Single-cell TCR sequencing enables high-resolution analysis of T Cell Receptor (TCR) diversity and clonality, offering valuable insights into immune responses and disease mechanisms. However, identifying cognate epitopes for individual TCRs requires complex and costly functional assays. We address this challenge with EpitopeGen, a large-scale transformer model based on the GPT-2 architecture that generates potential cognate epitope sequences directly from TCR sequences. To overcome the scarcity of TCR-epitope binding pairs ($\approx 100,000$), EpitopeGen uses a semi-supervised learning method, termed BINDSEARCH, which searches over 70 billion potential pairs and incorporates high binding affinity predictions as pseudo-labels. To incorporate $CD8^+$ T cell biology into the model as an inductive bias, EpitopeGen employs a novel data balancing method, termed Antigen Category Filter, that carefully controls antigen category ratios in its training dataset. EpitopeGen significantly outperforms baseline approaches, generating epitopes with high binding affinity, diversity, naturalness, and biophysical stability. Code is available at `https://github.com/Ding-Group/EpitopeGen`.

1. Introduction

The adaptive immune system is a specialized defense mechanism in vertebrates that provides long-lasting protection against pathogens by recognizing and memorizing specific antigens. T cells play a vital role in identifying and eliminating infected cells through their unique T cell receptors (TCRs). $CD8^+$ T cells, in particular, inspect endogenous peptides displayed on class I Major Histocompatibility Complex (MHC) molecules, expressed ubiquitously across human cells (Chaplin, 2010). Upon recognition of abnormal peptides, such as those of viral or tumoral origin, cytotoxic $CD8^+$ T cells can initiate apoptosis in the target cells, given appropriate co-stimulatory signals (Andersen et al., 2006). The specific antigen fragment recognized by the immune system is termed an epitope. The Complementarity Determining Region 3 (CDR3) of the TCR is primarily responsible for epitope binding, with the interaction affinity determined by the physicochemical properties of both protein sequences. The extreme polymorphism of CDR3, resulting from VDJ recombination (Tonegawa, 1983; Parham & Ohta, 1996), enables a diverse range of immune responses but presents challenges for quantitative modeling due to the vast sequence diversity at the TCR-pMHC (peptide-loaded MHC) interface.

Prior works in computational TCR analysis can be broadly grouped into three categories: TCR diversity metrics, TCR clustering methods, and TCR-pMHC binding affinity prediction models. To quantify the focused nature of immune responses, previous works (Vujović et al., 2023a; Shirasawa et al., 2025; Reuben et al., 2020; Porciello et al., 2022; Amoriello et al., 2021; Twyman-Saint Victor et al., 2015) used diversity indices such as Shannon entropy (Shannon, 1948), Simpson's diversity index (Simpson, 1949), and Rényi diversity (Rényi, 1961; Greiff et al., 2015). However, they provide limited insight into antigen specificity. Analysis of TCR repertoire data has been facilitated by various TCR clustering methods (Huang et al., 2020; Mayer-Blackwell et al., 2021; Sidhom et al., 2021; Zhang et al., 2021a). These methods aim to group TCRs with potentially similar antigen specificity. However, these clustering approaches, while valuable for repertoire analysis and motif discovery, TCRs within the same cluster may not share antigen specificity. Recent machine learning advances have enabled TCR-pMHC binding affinity prediction (Lu et al., 2021; Gao et al., 2023; Chronister et al., 2021; Weber et al., 2021; Jokinen et al., 2021; Montemurro et al., 2021; Zhang et al., 2021b; Tong et al., 2020; Springer et al., 2020; Zhang et al., 2023; Peng et al., 2023; Cai et al., 2022; Moris et al., 2020). However, these binding affinity prediction models have limited utility in analyzing TCR repertoires, as repertoire data typically

*Equal contribution [1]Department of Computer Science, University of British Columbia, Vancouver, Canada. Correspondence to: Jiarui Ding <jiarui.ding@ubc.ca>.

Proceedings of the Workshop on Generative AI for Biology at the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

lack corresponding epitope information.

Recent works have attempted to generate TCR sequences from given epitopes (Yang et al., 2023; Zhou et al., 2025). While these works may benefit TCR design, a critical gap exists in analyzing TCR repertoires. We can observe immune responses (TCRs) through repertoire sequencing but cannot easily identify what triggered them (epitopes). For instance, tumor-infiltrating lymphocytes contain TCRs that potentially recognize tumor antigens, but the specific epitopes remain unknown. Similarly, immune monitoring after vaccination reveals activated TCR repertoires without direct information about cognate epitopes.

To address these limitations, we explore the generative modeling of epitope sequences, inspired by the success of large language models in open-ended text generation (Radford et al., 2019; Lewis et al., 2020; Raffel et al., 2020). By developing an efficient algorithm to identify cognate epitopes, we aim to bridge the gap between TCR repertoire data and functional analysis. Accurately identifying patient T cell epitopes will help better categorize T cells and deepen our understanding of T cell biology. Thus, this approach can provide valuable insights for advancing personalized medicine and enhancing patient outcomes. For example, identifying tumor-specific TCRs can improve the precision of cancer therapies, allowing better targeting of cancer cells (Hudson et al., 2023). Furthermore, accurately identifying TCRs that respond to viral epitopes can help in the design of more effective, customized vaccines (Grifoni et al., 2020).

We introduce EpitopeGen, a large-scale generative transformer model that predicts cognate epitope sequences from TCR sequences. We train a decoder-only transformer to learn the conditional probability distribution of epitope sequences given TCR inputs. The self-attention mechanism captures the relationships between TCR tokens and generated epitopes. We propose BINDSEARCH, a semi-supervised learning method that evaluated over 70 billion TCR-epitope pairs and selected high-confidence interactions based on predicted binding affinity. A key innovation in our approach is the Antigen Category Filter (ACF), a novel data balancing method that calibrates the distribution of antigen categories in the training set based on established immunological principles of $CD8^+$ T cell recognition. These distributional constraints were essential for establishing an appropriate prior in the generative model, ensuring biological plausibility when applied to repertoire-level analysis.

To the best of our knowledge, EpitopeGen represents the first sequence-to-sequence generative model for predicting epitopes from TCRs. We evaluate the generated epitopes across multiple dimensions, including binding affinity, chemical properties, and naturalness. The results show that the generated epitopes exhibit high binding affinity to the input TCRs and possess chemical properties similar to those of natural epitopes. In repertoire-level evaluations, EpitopeGen generates diverse epitopes that conform to natural antigen category distributions. As an orthogonal validation, the generated epitopes led to energetically stable complexes when evaluated using Rosetta (Leaver-Fay et al., 2011) simulation.

2. Related Work

2.1. TCR Diversity and Clustering

Janarthanam et al (Janarthanam et al., 2023) employ D50, the minimum number of unique clonotypes constituting 50% of the total, to track clonality changes in pediatric eosinophilic esophagitis patients during dietary interventions. TCRDivER (Vujović et al., 2023b) introduces a similarity-sensitive diversity measure that jointly considers clone size and sequence similarity. GLIPH2 (Huang et al., 2020) uses a statistical method to identify TCR motifs that are overrepresented in query sets compared to background TCR repertoires. DeepTCR (Sidhom et al., 2021) leverages deep learning-based autoencoders to learn latent TCR representations that facilitate clustering of similar sequences. While these methods effectively capture T cell clonal expansion patterns, they provide limited insight into antigen recognition.

2.2. TCR-pMHC Binding Affinity Prediction

Recent advances in deep learning have significantly improved TCR-pMHC binding affinity prediction. Early approaches introduced novel feature representations, with ImRex (Moris et al., 2020) using physicochemical properties of amino acid residues and TCRMatch (Chronister et al., 2021) developing k-mer-based similarity matching for TCR identification. To address data scarcity, several methods employed transfer learning (pMTnet (Lu et al., 2021)) and probabilistic modeling (TCRGP (Jokinen et al., 2021)), while NetTCR-2.0 (Montemurro et al., 2021) focused on high-frequency TCR-epitope pairs using shallow CNNs. Neural architectures have been developed to capture sequence dependencies, including LSTM networks (ERGO (Springer et al., 2020)), bimodal attention networks (TITAN (Weber et al., 2021)), and dual-stream self-attention (ATM-TCR (Cai et al., 2022)). Recent approaches incorporate pre-training and structural information. TABR-BERT (Zhang et al., 2023) applies self-supervised learning with transformer encoders, TEIM (Peng et al., 2023) leverages structural data for residue-level interaction prediction, and PanPep (Gao et al., 2023) employs meta-learning for improved generalization to novel epitopes. Despite these advances, current models remain limited by their dependence on paired TCR-epitope data and focus on classification/regression rather than generative tasks, restricting their utility for analyzing unpaired repertoire data.

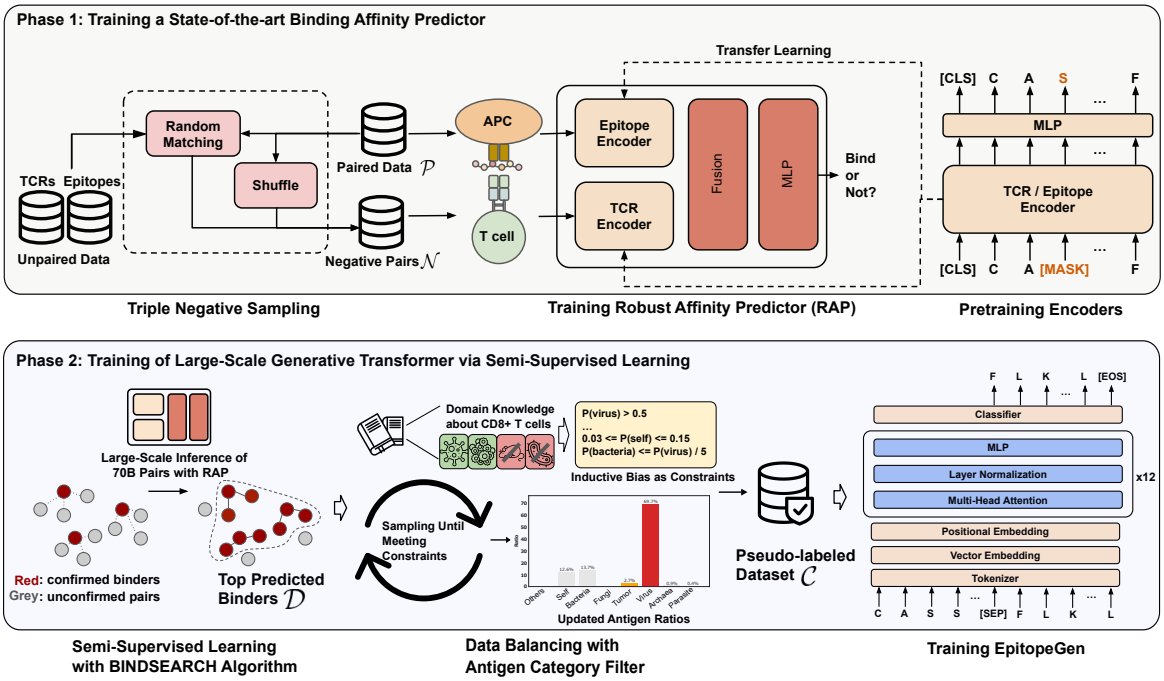

Figure 1. Overall architecture of EpitopeGen. The model employs a decoder-only transformer architecture to generate epitope sequences from TCR inputs, incorporating the Antigen Category Filter (ACF) to ensure biologically plausible distributions of predicted epitopes.

3. Method

Figure 1 shows our two-phase approach. In the first phase, we train a state-of-the-art binding affinity predictor (RAP). The performance benchmark of RAP can be found in Appendix A. Using RAP, in the second phase, we construct a large-scale pseudo-labeled dataset and train EpitopeGen. Training datasets for binding affinity predictors were compiled from four public sources: VDJdb, IEDB, PIRD, and McPAS-TCR. More details on datasets can be found in Appendix B.

3.1. Robust Affinity Predictor (RAP) training

Robust Affinity Predictor, our binding affinity predictor for pseudo-labeling, was developed by modifying TABR-BERT with three architectural changes: implementing a Softmax layer in the head architecture, removing MHC-related architectures, and utilizing PyTorch's CrossEntropyLoss instead of Contrastive loss. For the second modification, we retrained the BERT model (Devlin et al., 2019) solely on epitope sequences, a process that took two days using two NVIDIA L40S GPUs. The final model combined the predictions of five independently trained models through ensemble averaging.

A key challenge in training TCR-epitope binding predictors is the lack of confirmed non-binding pairs (negative pairs) in public datasets. To overcome this limitation, we developed Triple Negative Sampling (TNS), which generates diverse negative training examples through three complementary strategies. First, we pair known epitopes with TCRs from a large external pool. Second, we generate negative samples by pairing known TCRs with epitopes from a large external pool. Third, we randomly pair TCRs and epitopes within the dataset, based on the assumption that random TCR-epitope pairs are unlikely to bind. This diversified negative sampling approach helps reduce potential biases arising from relying on any single sampling strategy.

3.2. Semi-supervised learning method

The BINDSEARCH algorithm (**Algorithm 1**) generates a pseudo-labeled dataset of TCR-epitope pairs. Given sets of unpaired TCR sequences $\{t_i\}_{i=1}^{I}$ and epitope sequences $\{p_j\}_{j=1}^{J}$, the algorithm uses a binding affinity predictor function R (implemented as RAP) to estimate the binding affinity between TCRs and epitopes ($I = 6,831,478$, $J = 20,000,000$ were used). For each unpaired TCR t_i, the algorithm randomly samples $\beta = 10,000$ candidate epitopes from $\{p_j\}_{j=1}^{J}$. The binding affinity a is computed for each TCR-epitope pair (t_i, p) using R. Subsequently, the top $n_{\text{max\_tcr}} = 32$ pairs with the highest binding affinities are retained for each TCR. To mitigate redundancy in the resulting dataset, a filtering step is applied. Epitopes that occur more than $n_{\text{max\_epi}} = 100$ times in all pairs of TCR-epitopes are excluded. The value of $n_{\text{max\_epi}}$ was determined based

on the ratio of TCR to epitope observed in public datasets (specifically 116,057 epitopes to 1,141 TCR, resulting in a ratio of approximately 102). This process resulted in a pseudo-labeled dataset comprising $|\mathcal{D}| = 16,909,219$ TCR-epitope pairs. The algorithm took four days to run on ten NVIDIA L40S GPUs.

Algorithm 1 BINDSEARCH

Require:
$\{t_i\}_{i=1}^{I}$: Set of unpaired TCR sequences
$\{p_j\}_{j=1}^{J}$: Set of unpaired epitope sequences
R: Binding affinity predictor function (RAP)
β: Number of epitopes to check for each TCR (10000)
$n_{\text{max\_tcr}}$: Maximum number of epitopes per TCR (32)
$n_{\text{max\_epi}}$: Maximum occurrences of an epitope (100)
Ensure: $\mathcal{D}$: A dictionary of TCR-epitope pairs with binding affinities
function BINDSEARCH($\{t_i\}_{i=1}^{I}, \{p_j\}_{j=1}^{J}, R, \beta$)
  $\mathcal{D} \leftarrow \{\}$
  for $t_i \in \{t_i\}_{i=1}^{I}$ **do**
    $Q_i \leftarrow \text{RandomSample}(\{p_j\}_{j=1}^{J}, \beta)$
    $A_i \leftarrow \{\}$
    for $p \in Q_i$ **do**
      $a \leftarrow R(t_i, p)$
      $A_i \leftarrow A_i \cup (p, a)$
    end for
    $\mathcal{D}[t_i] \leftarrow \text{TopK}(A_i, n_{\text{max\_tcr}})$
  end for
  $\mathcal{D} \leftarrow \text{FilterRedundancy}(D, n_{\text{max\_epi}})$
  return $\mathcal{D}$
end function
function FILTERREDUNDANCY($\mathcal{D}, n_{\text{max\_epi}}$)
  $C \leftarrow \text{CountEpitopeOccurrences}(\mathcal{D})$
  for t_i, pairs $\in \mathcal{D}$.items() **do**
    $\mathcal{D}[t_i] \leftarrow (p, a) \in \text{pairs} : C[p] \leq n_{\text{max\_epi}}$
  end for
  return $\mathcal{D}$
end function=0

3.3. Antigen Category Filter

The intermediate dataset $\mathcal{D}$ was biased towards Eukaryotic species, a consequence of the peptide collection methods used by NetMHCPanv4.0 (Jurtz et al., 2017), MHCflurryv2.0 (O'Donnell et al., 2020), and SysteMHC (Huang et al., 2023). This bias likely reflects research priorities and funding rather than the biological distribution of antigens potentially recognized by $CD8^+$ T cells. To correct this discrepancy, we implemented the Antigen Category Filter (ACF) algorithm (**Algorithm 2**). ACF takes as input a set of redundancy-removed pseudo-labeled TCR-epitope pairs $\{(t_l, p_l)\}_{l=1}^{L}$ and a set of antigen categories with their target ratios $\{(c_n, r_n)\}_{n=1}^{N}$. The Antigen Category Filter begins by identifying species for each epitope. Then, it counts the number of pivot category and determines the target numbers to choose from all categories. Based on the target numbers, the samples are randomly drawn, which effectively adjusts the antigen category distribution in the dataset.

To determine the target antigen ratios, we considered five immunological insights: (1) Viral Dominance (Masopust et al., 2007; Moutaftsi et al., 2006; Addo et al., 2003), (2) Limited Bacteria (Friot et al., 2023; Shepherd & McLaren, 2020), (3) Endogenous Presence (Pittet et al., 1999; Rizzuto et al., 2009; Nelson et al., 2019; Kenison et al., 2024), (4) Rare Fungi and Parasites (Mittal et al., 2019; Stuckey Peter V. & Santiago-Tirado Felipe H., 2023; Walker et al., 2013; Morrison, 2009; Stuart et al., 2008), and (5) No Reported Pathogenic Archaea (Cavicchioli et al., 2003; Gill & Brinkman, 2011). In-depth discussion of target ratio ranges was provided in Appendix C. We call the resultant balanced dataset by Corpus or $\mathcal{C}$.

Algorithm 2 Antigen Category Filter (ACF)

Require:
$\mathcal{D} = \{(t_l, p_l)\}_{l=1}^{L}$: Set of redundancy-removed pseudo-labeled TCR-epitope pairs
$\{(c_n, r_n)\}_{n=1}^{N}$: Set of antigen categories and their target ratios, where $\sum_{n=1}^{N} r_n = 1$, $N = 9$, and r_1 is the ratio for the pivot
Ensure: $\mathcal{C}$: Corpus, or TCR-epitope pairs with balanced antigen category
function ACF($\{(t_l, p_l)\}_{l=1}^{L}, \{(c_n, r_n)\}_{n=1}^{N}$)
  $\{(t_l, p_l, c_l)\}_{l=1}^{L} \leftarrow \text{SearchCategory}(\{(t_l, p_l)\}_{l=1}^{L})$
  $M \leftarrow \text{CountPivotCategory}(\{(t_l, p_l, c_l)\}_{l=1}^{L})$
  $\{c'_n\}_{n=1}^{N} \leftarrow \{M \cdot r_n / r_1\}_{n=1}^{N}$ // Target numbers
  $\mathcal{C} \leftarrow \{\}$
  for $n \in 1, \ldots, N$ **do**
    $S_n \leftarrow (t_l, p_l) : (t_l, p_l, c_l) \in (t_l, p_l, c_l)_{l=1}^{L}$ and $c_l = c_n$
    $\mathcal{C} \leftarrow \mathcal{C} \cup \text{RandomSample}(S_n, c'_n)$
  end for
  return $\mathcal{C}$
end function
function SEARCHCATEGORY($\{(t_l, p_l)\}_{l=1}^{L}$)
  // Use blastp and NCBI database queries to determine species and categories
  // Return $\{(t_l, p_l, c_l)\}_{l=1}^{L}$ where c_l is the category for each pair
end function
function COUNTPIVOTCATEGORY($(t_l, p_l, c_l)_{l=1}^{L}$)
  return $|(t_l, p_l, c_l) \in \{(t_l, p_l, c_l)\}_{l=1}^{L} : c_l = \text{“Virus”}|$
end function=0

3.4. EpitopeGen architecture and training

EpitopeGen is a decoder-only transformer model specifically designed to generate epitope sequences while adhering to specific distributional constraints, including epitope di-

versity and biologically plausible antigen distributions.The model architecture is based on GPT-2, with amino acid sequences encoded using a BPE tokenizer (Sennrich et al., 2016) trained specifically on our amino acid corpus.

The Byte-Pair Encoding (BPE) algorithm iteratively merges the most frequent pairs of tokens, capturing the recurring subsequences as single tokens. This tokenization method allows for the representation of single amino acids or groups of amino acids as individual tokens, potentially capturing meaningful biological motifs. For a TCR-epitope pair (t, p), the input sequence is tokenized as:

$$\mathbf{x} = [\text{BPE}(t); [\text{SEP}]; \text{BPE}(p); [\text{EOS}]] \tag{1}$$

where $\text{BPE}(\cdot)$ denotes the BPE tokenization function, ';' represents concatenation, [SEP] delineates the boundary between TCR and epitope sequences, and [EOS] marks the end of each sequence. The tokenized sequences were processed using positional embeddings, where position-specific vectors are added to the token embeddings to maintain sequence order information.

The model defines a probability distribution $p_{\boldsymbol{\theta}}(\mathbf{x})$ over a sequence of tokens $\mathbf{x}$ of length n, which can be factorized as:

$$p_{\boldsymbol{\theta}}(\mathbf{x}) = \prod_{\tau=1}^{n} p_{\boldsymbol{\theta}}(\mathbf{x}_\tau | \mathbf{x}_{<\tau}) \tag{2}$$

where $\boldsymbol{\theta}$ represents the model parameters, $\mathbf{x}_\tau$ is the τ-th token in the sequence, and $\mathbf{x}_{<\tau}$ denotes all tokens before τ. This autoregressive formulation allows the model to generate epitope sequences token by token, conditioned on the input TCR sequence. The objective function for training is the negative log-likelihood:

$$\mathcal{L}(\boldsymbol{\theta}) = -\sum_{(t,p)\in\mathcal{C}} \log p_{\boldsymbol{\theta}}(\mathbf{x}). \tag{3}$$

Given the narrow length distribution of the TCR and epitope sequences compared to natural language paragraphs, each batch contained a single TCR-epitope pair. The AdamW optimizer (Loshchilov & Hutter, 2017) was used with parameters: initial learning rate ($\alpha = 1 \times 10^{-5}$), $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \times 10^{-8}$, and weight decay ($\lambda = 0.01$). Training took four hours for EpitopeGen and four days for EpitopeGenNoACF using four NVIDIA L40S GPUs.

4. EpitopeGen generates high-affinity, diverse, and biologically sane epitopes

We evaluated the binding affinities between the input TCRs and the generated epitopes across multiple test scenarios. We partitioned the test set of Corpus $\mathcal{C}$ into four subsets based on TCR and epitope exposure during training: UnseenEpi (TCRs seen, epitopes unseen), UnseenTCR (epitopes seen, TCRs unseen), SeenBoth (both seen, but not as a pair), and UnseenBoth (neither seen). EpitopeGen generated epitope sequences for each TCR in the test sets, and the binding affinity was measured using RAP. For comparison, we also measured the binding affinities between each TCR and 100 randomly sampled epitopes and calculated the percentile ranks. Table 1 shows the average percentile rank of binding affinity values of the generated epitopes. The percentile ranks for UnseenEpi, UnseenTCR, SeenBoth, and UnseenBoth were 81.5, 81.3, 81.9, and 81.2, respectively. These results indicated that EpitopeGen can generate epitopes with high binding affinities for both previously encountered and novel TCRs.

Table 1. Average percentile rank of binding affinity ($\pm$ 95% CI) for EpitopeGen across different test set splits

UnseenEpi	**UnseenTCR**	**SeenBoth**	**UnseenBoth**
81.46 ± 0.90	81.35 ± 0.20	81.91 ± 0.44	81.25 ± 0.43

We evaluated EpitopeGen using test sets of VDJdb, IEDB, PIRD, and McPAS-TCR. Two baseline methods were implemented for comparison: RandGen, which generates random amino acid sequences based on the training set's epitope length distribution, and BLOSUMGen, which assigns epitopes from the training set based on TCR sequence similarity using BLOSUM62 substitution matrix alignment. EpitopeGen-generated epitopes demonstrated consistently high binding affinities across multiple test sets, with mean percentile ranks exceeding 80%. In evaluations using VDJdb, PIRD, and McPAS-TCR test sets, EpitopeGen outperformed both RandGen and BLOSUMGen while maintaining comparable performance on IEDB (Table 2). External validation using independent experimental datasets from Glanville et al. (Glanville et al., 2017) and Nolan et al. (Nolan et al., 2020) further confirmed EpitopeGen's robustness, achieving mean percentile ranks of 85.8 and 84.1, respectively.

To demonstrate EpitopeGen's predictive capabilities, we analyzed TCR sequences associated with NLVPMVATV, the most frequently occurring epitope in the VDJdb dataset. In Figure 2, we compared TCRs that generated this epitope against experimentally validated TCRs using Logomaker (Tareen & Kinney, 2020). The generated sequences exhibited characteristic amino acid patterns with notable variability in their central regions. Both the generated and experimentally validated sequences displayed conserved motifs: an N-terminal 'CASS' pattern, a central 'LGGGGYE' sequence, and a C-terminal 'QFF' motif. This consistent pattern alignment demonstrates EpitopeGen's ability to generate epitope sequences from TCRs that are similar to experimentally validated ones.

We further evaluated EpitopeGen's performance using repertoire-level test sets, specifically the 10x dataset (10x

Table 2. Average percentile rank of binding affinity (± 95% CI) between generated epitopes and TCRs across benchmark datasets. Higher values indicate stronger binding.

Method	VDJdb	IEDB	PIRD	McPAS-TCR	Glanville	MIRA
RandGen	61.20 ± 2.30	59.86 ± 0.56	62.54 ± 2.39	63.22 ± 1.58	61.55 ± 1.26	59.92 ± 1.30
BLOSUMGen	47.50 ± 2.56	**88.79** ± 0.53	47.76 ± 2.86	50.31 ± 1.97	48.07 ± 1.49	**90.19** ± 0.93
EpitopeGen	**79.86** ± 1.71	**88.18** ± 0.34	**83.42** ± 1.69	**81.51** ± 1.19	**84.80 ± 0.83**	82.36 ± 0.92

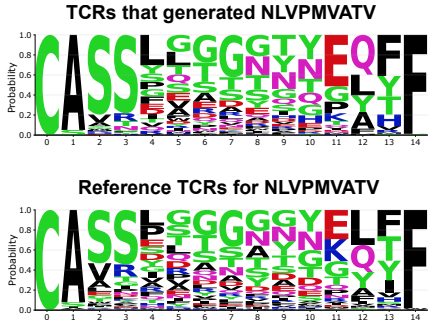

Figure 2. Sequence motif comparison for TCRs recognizing the NLVPMVATV epitope. (Top) Logomaker plot of TCRs generated by EpitopeGen. (Bottom) Logomaker plot of reference TCRs from experimental data. The similar motif patterns indicate that EpitopeGen successfully captures the key sequence features of TCRs recognizing this epitope.

Genomics, 2022) released by 10x Genomics, Inc. This evaluation closely mirrors real-world application scenarios in which EpitopeGen is used to infer epitopes for an individual's entire TCR repertoire. For comparative analysis, we fine-tuned ProGen2 (Nijkamp et al., 2023), a leading protein language model, on the publicly available training set. We also developed three variants of EpitopeGen. EpitopeGenNoACF was trained on the intermediate dataset $\mathcal{D}$ without applying Antigen Category Filter, thus lacking inductive bias on the proper distribution of antigen categories. EpitopeGenNoACFFinetune was derived by fine-tuning EpitopeGenNoACF using $\mathcal{C}$. EpitopeGenMHC incorporated both TCR and MHC information for epitope sequence generation.

The generated epitopes by EpitopeGen predominantly originated from viruses (Figure 3), with smaller proportions from tumoral, self, and bacterial sources. This distribution aligns with the immunological principles of Viral Dominance, Limited Bacterial Presence, and Endogenous Epitope Presence. In contrast, all other models (ProGen2Finetuned, EpitopeGenNoACF, and EpitopeGenNoACFFinetune) generated approximately 37.7% and 30% of epitopes from 'Other' (mostly Eukaryotic) and 'Bacteria' with notably fewer viral antigens. These skewed antigen category distributions of generated epitopes reflect the initial bias in the dataset $\mathcal{D}$ before the application of Antigen Category Filter.

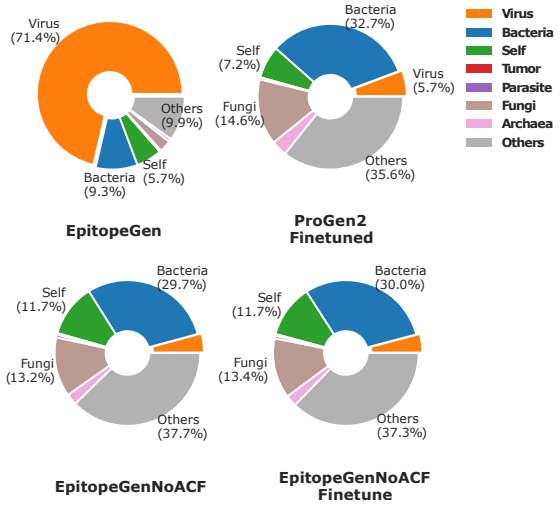

Figure 3. Source antigen distribution of predicted epitopes from EpitopeGen, ProGen2 Finetuned, EpitopeGenNoACF, and EpitopeGenNoACFFinetune. The Antigen Category Filter (ACF) in EpitopeGen helps maintain biologically realistic distributions.

To assess epitope diversity, we employed six diversity indices: Shannon diversity (Shannon, 1948; Greiff et al., 2015), Rényi diversity (α=2) (Rényi, 1961), Simpson's diversity index (Simpson, 1949), the Epi-to-TCR ratio (unique epitopes/number of TCRs), avg_repetition_top_1_percent, and top_10_concentration (proportion of epitopes in most frequent 10%). Figure 4 shows that EpitopeGen generated epitopes showed superior diversity, achieving an epitope-to-TCR ratio of 0.5. In contrast, EpitopeGenMHC showed significantly lower diversity indices, indicating the generation of redundant epitopes for different TCRs. This limitation stems from the lack of (TCR, epitope, MHC) triplets in the currently available datasets.

To prove the effectiveness of semi-supervised learning, we evaluated models trained with different proportions of unlabeled data. The Pseudo-labeling 0% model, trained solely on the public paired dataset, showed considerable redundancy, with the top 1% epitopes repeating approximately 2,000 times on average and a top_10_concentration exceeding 0.95 (Figure 4, right). In contrast, Pseudo-labeling 100%

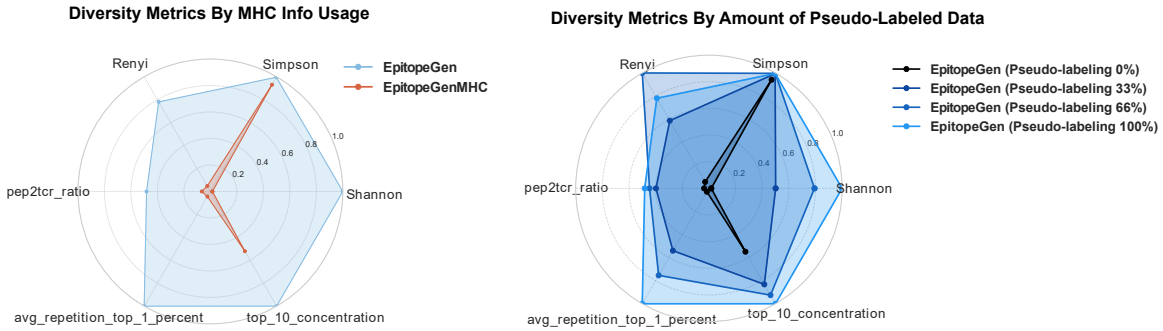

Figure 4. Diversity metrics comparison. Radar plots comparing six diversity indices of generated epitopes using EpitopeGen versus EpitopeGenMHC, with model variants trained with different proportions of pseudo-labeled data (0%, 33%, 66%, and 100%).

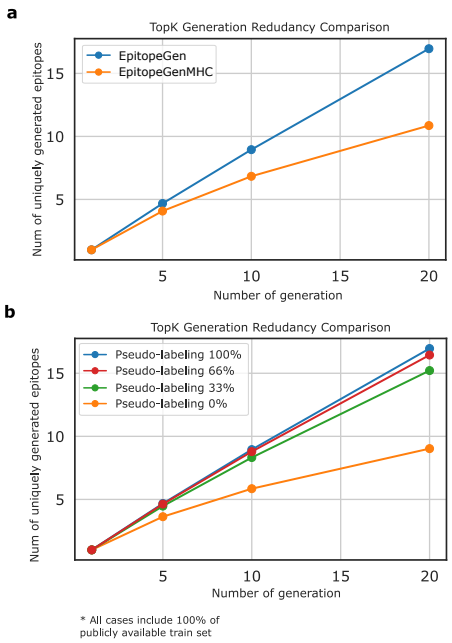

Figure 5. Redundancy in Top K generations. a, between EpitopeGen and EpitopeGenMHC, b, by the proportion of pseudo-labeled data.

achieved much greater diversity with a top_10_concentration below 0.50. Epitope diversity improved progressively with increasing proportions of pseudo-labeled data, highlighting the advantage of incorporating unlabeled data in ensuring the generation of diverse epitopes.

Figure 5 shows the number of uniquely generated epitopes by the number of generation attempts. EpitopeGen showed lower redundancy in its top-20 generations compared to EpitopeGenMHC (Figure 5a). Additionally, redundancy in the top-k generations decreased progressively as the proportion of pseudo-labeled training data increased (Figure 5b). These results show that EpitopeGen produces more diverse epitopes in its top-k predictions compared to the baselines.

5. EpitopeGen generates natural epitopes

We next examined the naturalness of the generated epitopes by comparing their various properties with those of naturally occurring epitopes collected from the test sets. The generated epitopes had an average length of 10.08, which aligns well with the typical length range (Trolle et al., 2016) (8 to 12 amino acids) of the epitopes loaded onto MHC class I molecules (Figure 6, left). Additionally, the amino acid usage patterns of the generated epitopes closely resembled those of natural epitopes (Pearson correlation = 0.911; Figure 7, blue).

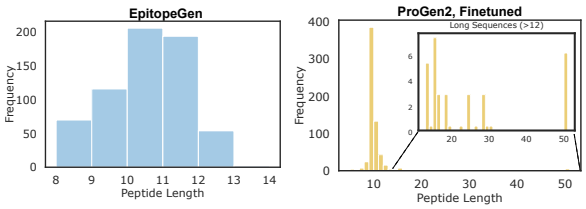

Figure 6. Length distribution of generated epitopes.

Recent protein language models, such as ProtGPT2 (Ferruz et al., 2022), ProGen (Madani et al., 2023), and ProGen2 (Nijkamp et al., 2023), demonstrated protein sequence generation capabilities, but lack precise control mechanisms for specialized tasks such as the generation of $CD8^+$ T cell epitopes. To evaluate the utility of pre-trained models, we fine-tuned ProGen2 on public training data. Our experiments revealed that the fine-tuned model often generated epitopes exceeding the biological length constraints, exhibiting a long-tailed length distribution (Figure 6). This behavior stems from ProGen2's pre-training on general protein sequences, which are typically longer than T-cell epitopes. Specifically, 4.41% of the generated sequences were longer than 12, while 1.47% were shorter than 8 amino acids. The model also showed different amino acid usage patterns compared to natural epitopes (Figure 7, yellow). When tested on the 10x dataset, the fine-tuned ProGen2 model exhibited

severe bias, with 83.80% of generated sequences starting with 'G'. These findings suggest that strongly conditional generation tasks, where a single amino acid difference can significantly impact binding properties, require enhanced supervision through carefully curated training data that satisfy biological constraints.

To assess the chemical feasibility of the generated epitopes, we analyzed their several key properties using the `ProtParam` package (Wilkins et al., 1999). Table 3 shows that the distributions of these chemical properties in EpitopeGen-generated epitopes closely mirrored those of natural epitopes, while randomly generated epitopes showed significantly different distributions.

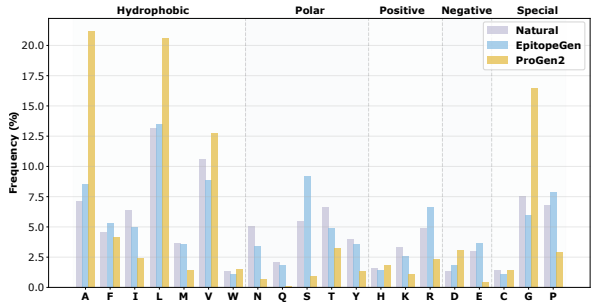

Figure 7. Amino acid usage comparison between natural epitopes, EpitopeGen-generated epitopes, and ProGen2 (Finetuned)-generated epitopes. EpitopeGen-generated epitopes share a more similar amino acid distribution to natural epitopes.

Source	Extinction Coefficient	Aromaticity	Secondary Structure
Natural	0.00	0.09	0.44
EpitopeGen	0.00	0.09	0.40
p-value	**(1.00)**	**(1.00)**	**(0.19)**
RandGen	1490.00	0.11	0.33
p-value	(1.9e-7)	(4.3e-3)	(5.5e-5)

Table 3. Chemical properties of natural and generated epitopes p-values represent statistical comparison with natural epitopes. Bold p-values indicate properties where generated epitopes cannot be statistically distinguished from natural epitopes ($p > 0.05$).

6. EpitopeGen generates biophysically stable epitopes

We use Molecular Dynamics (MD) simulations (McCammon et al., 1977) to measure biophysical properties at the interface of TCR-pMHC, which provides orthogonal information compared to deep learning methods. Specifically, we utilized `InterfaceAnalyzer` (Stranges & Kuhlman, 2012), from the `Rosetta` suite (Leaver-Fay et al., 2011), to measure free energy and hydrophobicity. Gibbs free energy (dG_separated) quantifies the energy difference before and after TCR-pMHC binding (Alford et al., 2017). Hydrophobic interaction was measured due to its importance in protein folding and docking (Dill, 1990). For comparison, we sampled 20 random epitopes from the VDJdb test set as controls. Table 4 shows that EpitopeGen-generated epitopes showed lower Gibbs free energy compared to randomly sampled epitopes, with a median percentile rank of 22.5%. This suggests that EpitopeGen-generated epitopes form more energetically stable complexes compared to randomly sampled ones. Furthermore, these epitopes exhibited pronounced hydrophobic interactions, with a mean percentile rank of 82.5%. This observation supports the idea that the generated epitopes form stronger hydrophobic interactions with the CDR3β region, potentially burying hydrophobic regions and contributing to binding stability.

Source	Percentile Rank	
	Binding ↓ (dG_sep)	**Hydrophobic ↑ (dSASA_hp)**
Random	50.00	50.00
VDJdb	30.00	70.00
EpitopeGen	**22.50**	**82.50**

Table 4. Percentile ranks compared against randomly sampled peptides with identical TCRs. Lower binding energy (dG_sep ↓) and higher hydrophobic interface area (dSASA_hp ↑) are favorable.

7. Conclusion

Precise identification of T-cell binding partners is crucial for understanding human adaptive immunity and developing targeted therapies, including immunotherapies and personalized vaccines. We introduced EpitopeGen, a large-scale generative transformer designed to predict potential epitope sequences from TCR data, thereby enhancing the utility of single-cell TCR sequencing analysis. To address the fundamental challenge of limited paired TCR-epitope training data, we proposed a semi-supervised learning method that incorporates a large number of unpaired data, complemented by a novel Antigen Category Filter algorithm that ensures biologically plausible antigen category distributions in the data used to train EpitopeGen.

Our key innovations include the Robust Affinity Predictor for reliable binding prediction, BINDSEARCH for leveraging unpaired data, and the Antigen Category Filter for maintaining biologically appropriate antigen category distributions. Together, these components enable EpitopeGen to generate epitopes that simultaneously satisfy multiple critical criteria: high binding affinity, sequence diversity, natural amino acid composition, and biophysical stability. Importantly, when applied to repertoire-scale TCR data, EpitopeGen produced epitopes with antigen category distributions aligning with established immunological principles.

Acknowledgements

This work was supported by a Discovery grant from the Natural Sciences and Engineering Research Council (NSERC) of Canada, and a department startup fund from the University of British Columbia (to J.D.). J.D. is a Canada Research Chair and is supported by the Canadian Institutes of Health Research through the Canada Research Chair Program. The computational resource is partially supported by the Canada Foundation for Innovation & John. R. Evans Leader Fund (to J.D.). This research was supported in part through the computational resources and services provided by Advanced Research Computing at the University of British Columbia.

Impact Statement

This paper presents work aiming to advance the field of machine learning and T cell biology. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

10x Genomics. A new way of exploring immunity: Linking highly multiplexed antigen recognition to immune repertoire and phenotype. Application note, 10x Genomics, 2022.

Addo, M. M., Yu, X. G., Rathod, A., Cohen, D., Eldridge, R. L., Strick, D., Johnston, M. N., Corcoran, C., Wurcel, A. G., Fitzpatrick, C. A., Feeney, M. E., Rodriguez, W. R., Basgoz, N., Draenert, R., Stone, D. R., Brander, C., Goulder, P. J. R., Rosenberg, E. S., Altfeld, M., and Walker, B. D. Comprehensive epitope analysis of human immunodeficiency virus type 1 (HIV-1)-specific t-cell responses directed against the entire expressed HIV-1 genome demonstrate broadly directed responses, but no correlation to viral load. *J Virol*, 77(3):2081–2092, February 2003.

Alford, R. F., Leaver-Fay, A., Jeliazkov, J. R., O'Meara, M. J., DiMaio, F. P., Park, H., Shapovalov, M. V., Renfrew, P. D., Mulligan, V. K., Kappel, K., Labonte, J. W., Pacella, M. S., Bonneau, R., Bradley, P., Dunbrack, Jr, R. L., Das, R., Baker, D., Kuhlman, B., Kortemme, T., and Gray, J. J. The rosetta All-Atom energy function for macromolecular modeling and design. *J Chem Theory Comput*, 13(6): 3031–3048, May 2017.

Amoriello, R., Chernigovskaya, M., Greiff, V., Carnasciali, A., Massacesi, L., Barilaro, A., Repice, A. M., Biagioli, T., Aldinucci, A., Muraro, P. A., Laplaud, D. A., Lossius, A., and Ballerini, C. TCR repertoire diversity in multiple sclerosis: High-dimensional bioinformatics analysis of sequences from brain, cerebrospinal fluid and peripheral blood. *eBioMedicine*, 68, June 2021.

Andersen, M. H., Schrama, D., Thor Straten, P., and Becker, J. C. Cytotoxic T cells. *J Invest Dermatol*, 126(1):32–41, January 2006.

Cai, M., Bang, S., Zhang, P., and Lee, H. ATM-TCR: TCR-Epitope binding affinity prediction using a Multi-Head Self-Attention model. *Front Immunol*, 13:893247, July 2022.

Cavicchioli, R., Curmi, P. M., Saunders, N., and Thomas, T. Pathogenic archaea: do they exist? *BioEssays*, 25(11): 1119–1128, 2003.

Chaplin, D. D. Overview of the immune response. *J Allergy Clin Immunol*, 125(2 Suppl 2):S3–23, February 2010.

Chen, S.-Y., Yue, T., Lei, Q., and Guo, A.-Y. TCRdb: a comprehensive database for t-cell receptor sequences with powerful search function. *Nucleic Acids Res*, 49 (D1):D468–D474, January 2021.

Chronister, W. D., Crinklaw, A., Mahajan, S., Vita, R., Koşaloğlu-Yalçın, Z., Yan, Z., Greenbaum, J. A., Jessen, L. E., Nielsen, M., Christley, S., Cowell, L. G., Sette, A., and Peters, B. TCRMatch: Predicting T-Cell receptor specificity based on sequence similarity to previously characterized receptors. *Front Immunol*, 12:640725, March 2021.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019.

Dill, K. A. Dominant forces in protein folding. *Biochemistry*, 29(31):7133–7155, August 1990.

Ferruz, N., Schmidt, S., and Höcker, B. ProtGPT2 is a deep unsupervised language model for protein design. *Nature Communications*, 13(1):4348, July 2022.

Friot, A., Djebali, S., Valsesia, S., Parroche, P., Dubois, M., Baude, J., Vandenesch, F., Marvel, J., and Leverrier, Y. Antigen specific activation of cytotoxic CD8(+) T cells by staphylococcus aureus infected dendritic cells. *Front Cell Infect Microbiol*, 13:1245299, October 2023.

Gao, Y., Gao, Y., Fan, Y., Zhu, C., Wei, Z., Zhou, C., Chuai, G., Chen, Q., Zhang, H., and Liu, Q. Pan-Peptide meta learning for t-cell receptor–antigen binding recognition. *Nature Machine Intelligence*, 5(3):236–249, March 2023.

Gill, E. E. and Brinkman, F. S. L. The proportional lack of archaeal pathogens: Do viruses/phages hold the key? *BioEssays*, 33(4):248–254, 2011.

Glanville, J., Huang, H., Nau, A., Hatton, O., Wagar, L. E., Rubelt, F., Ji, X., Han, A., Krams, S. M., Pettus, C., Haas,

N., Arlehamn, C. S. L., Sette, A., Boyd, S. D., Scriba, T. J., Martinez, O. M., and Davis, M. M. Identifying specificity groups in the T cell receptor repertoire. *Nature*, 547(7661):94–98, July 2017.

Greiff, V., Bhat, P., Cook, S. C., Menzel, U., Kang, W., and Reddy, S. T. A bioinformatic framework for immune repertoire diversity profiling enables detection of immunological status. *Genome Med*, 7(1):49, May 2015.

Grifoni, A., Weiskopf, D., Ramirez, S. I., Mateus, J., Dan, J. M., Moderbacher, C. R., Rawlings, S. A., Sutherland, A., Premkumar, L., Jadi, R. S., Marrama, D., de Silva, A. M., Frazier, A., Carlin, A. F., Greenbaum, J. A., Peters, B., Krammer, F., Smith, D. M., Crotty, S., and Sette, A. Targets of T cell responses to SARS-CoV-2 coronavirus in humans with COVID-19 disease and unexposed individuals. *Cell*, 181(7):1489–1501.e15, May 2020.

Huang, H., Wang, C., Rubelt, F., Scriba, T. J., and Davis, M. M. Analyzing the mycobacterium tuberculosis immune response by t-cell receptor clustering with GLIPH2 and genome-wide antigen screening. *Nature Biotechnology*, 38(10):1194–1202, October 2020.

Huang, X., Gan, Z., Cui, H., Lan, T., Liu, Y., Caron, E., and Shao, W. The SysteMHC Atlas v2.0, an updated resource for mass spectrometry-based immunopeptidomics. *Nucleic Acids Research*, 52(D1):D1062–D1071, 11 2023. ISSN 0305-1048.

Hudson, D., Fernandes, R. A., Basham, M., Ogg, G., and Koohy, H. Can we predict T cell specificity with digital biology and machine learning? *Nature Reviews Immunology*, 23(8):511–521, August 2023.

Janarthanam, R., Kuang, F. L., Zalewski, A., Amsden, K., Wang, M.-Y., Ostilla, L., Keeley, K., Hirano, I., Kagalwalla, A., Wershil, B. K., Gonsalves, N., and Wechsler, J. B. Bulk t-cell receptor sequencing confirms clonality in pediatric eosinophilic esophagitis and identifies a food-specific repertoire. *Allergy*, 78(9):2487–2496, May 2023.

Jokinen, E., Huuhtanen, J., Mustjoki, S., Heinonen, M., and Lähdesmäki, H. Predicting recognition between T cell receptors and epitopes with TCRGP. *PLoS Comput Biol*, 17(3):e1008814, March 2021.

Jurtz, V., Paul, S., Andreatta, M., Marcatili, P., Peters, B., and Nielsen, M. NetMHCpan-4.0: Improved Peptide-MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. *J Immunol*, 199 (9):3360–3368, October 2017.

Kenison, J. E., Stevens, N. A., and Quintana, F. J. Therapeutic induction of antigen-specific immune tolerance. *Nature Reviews Immunology*, 24(5):338–357, May 2024.

Leaver-Fay, A., Tyka, M., Lewis, S. M., Lange, O. F., Thompson, J., Jacak, R., Kaufman, K., Renfrew, P. D., Smith, C. A., Sheffler, W., Davis, I. W., Cooper, S., Treuille, A., Mandell, D. J., Richter, F., Ban, Y.-E. A., Fleishman, S. J., Corn, J. E., Kim, D. E., Lyskov, S., Berrondo, M., Mentzer, S., Popović, Z., Havranek, J. J., Karanicolas, J., Das, R., Meiler, J., Kortemme, T., Gray, J. J., Kuhlman, B., Baker, D., and Bradley, P. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol*, 487:545–574, 2011.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, Online, July 2020. Association for Computational Linguistics.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017.

Lu, T., Zhang, Z., Zhu, J., Wang, Y., Jiang, P., Xiao, X., Bernatchez, C., Heymach, J. V., Gibbons, D. L., Wang, J., Xu, L., Reuben, A., and Wang, T. Deep learning-based prediction of the T cell receptor–antigen binding specificity. *Nature Machine Intelligence*, 3(10):864–875, October 2021.

Madani, A., Krause, B., Greene, E. R., Subramanian, S., Mohr, B. P., Holton, J. M., Olmos, J. L., Xiong, C., Sun, Z. Z., Socher, R., Fraser, J. S., and Naik, N. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 41(8): 1099–1106, August 2023.

Masopust, D., Murali-Krishna, K., and Ahmed, R. Quantitating the magnitude of the lymphocytic choriomeningitis virus-specific cd8 t-cell response: It is even bigger than we thought. *Journal of Virology*, 81(4):2002–2011, 2007.

Mayer-Blackwell, K., Schattgen, S., Cohen-Lavi, L., Crawford, J. C., Souquette, A., Gaevert, J. A., Hertz, T., Thomas, P. G., Bradley, P., and Fiore-Gartland, A. Tcr meta-clonotypes for biomarker discovery with *tcrdist3* enabled identification of public, hla-restricted clusters of sars-cov-2 tcrs. *eLife*, 10:e68605, nov 2021. ISSN 2050-084X.

McCammon, J. A., Gelin, B. R., and Karplus, M. Dynamics of folded proteins. *Nature*, 267(5612):585–590, June 1977.

Mittal, J., Ponce, M. G., Gendlina, I., and Nosanchuk, J. D. Histoplasma capsulatum: Mechanisms for pathogenesis. *Curr Top Microbiol Immunol*, 422:157–191, 2019.

Montemurro, A., Schuster, V., Povlsen, H. R., Bentzen, A. K., Jurtz, V., Chronister, W. D., Crinklaw, A., Hadrup, S. R., Winther, O., Peters, B., Jessen, L. E., and Nielsen, M. NetTCR-2.0 enables accurate prediction of TCR-peptide binding by using paired TCRα and β sequence data. *Communications Biology*, 4(1):1060, September 2021.

Moris, P., De Pauw, J., Postovskaya, A., Gielis, S., De Neuter, N., Bittremieux, W., Ogunjimi, B., Laukens, K., and Meysman, P. Current challenges for unseen-epitope TCR interaction prediction and a new perspective derived from image classification. *Briefings in Bioinformatics*, 22(4):bbaa318, 12 2020. ISSN 1477-4054.

Morrison, D. A. Evolution of the apicomplexa: where are we now? *Trends in Parasitology*, 25(8):375–382, August 2009.

Moutaftsi, M., Peters, B., Pasquetto, V., Tscharke, D. C., Sidney, J., Bui, H.-H., Grey, H., and Sette, A. A consensus epitope prediction approach identifies the breadth of murine TCD8+-cell responses to vaccinia virus. *Nature Biotechnology*, 24(7):817–819, July 2006.

Nelson, C. E., Thompson, E. A., Quarnstrom, C. F., Fraser, K. A., Seelig, D. M., Bhela, S., Burbach, B. J., Masopust, D., and Vezys, V. Robust iterative stimulation with Self-Antigens overcomes CD8(+) T cell tolerance to self- and tumor antigens. *Cell Rep*, 28(12):3092–3104.e5, September 2019.

Nijkamp, E., Ruffolo, J. A., Weinstein, E. N., Naik, N., and Madani, A. ProGen2: Exploring the boundaries of protein language models. *Cell Systems*, 14(11):968–978.e3, November 2023.

Nolan, S., Vignali, M., Klinger, M., Dines, J. N., Kaplan, I. M., Svejnoha, E., Craft, T., Boland, K., Pesesky, M., Gittelman, R. M., Snyder, T. M., Gooley, C. J., Semprini, S., Cerchione, C., Mazza, M., Delmonte, O. M., Dobbs, K., Carreño-Tarragona, G., Barrio, S., Sambri, V., Martinelli, G., Goldman, J. D., Heath, J. R., Notarangelo, L. D., Carlson, J. M., Martinez-Lopez, J., and Robins, H. S. A large-scale database of t-cell receptor beta (TCRβ) sequences and binding associations from natural and synthetic exposure to SARS-CoV-2. August 2020.

O'Donnell, T. J., Rubinsteyn, A., and Laserson, U. Mhcflurry 2.0: Improved pan-allele prediction of mhc class i-presented peptides by incorporating antigen processing. *Cell Systems*, 11(1):42–48.e7, 2020. ISSN 2405-4712.

Parham, P. and Ohta, T. Population biology of antigen presentation by MHC class I molecules. *Science*, 272 (5258):67–74, April 1996.

Peng, X., Lei, Y., Feng, P., Jia, L., Ma, J., Zhao, D., and Zeng, J. Characterizing the interaction conformation between t-cell receptors and epitopes with deep learning. *Nature Machine Intelligence*, 5(4):395–407, April 2023.

Pittet, M. J., Valmori, D., Dunbar, P. R., Speiser, D. E., Liénard, D., Lejeune, F., Fleischhauer, K., Cerundolo, V., Cerottini, J. C., and Romero, P. High frequencies of naive Melan-A/MART-1-specific CD8(+) T cells in a large proportion of human histocompatibility leukocyte antigen (HLA)-A2 individuals. *J Exp Med*, 190(5):705–715, September 1999.

Porciello, N., Franzese, O., D'Ambrosio, L., Palermo, B., and Nisticò, P. T-cell repertoire diversity: friend or foe for protective antitumor response? *Journal of Experimental & Clinical Cancer Research*, 41(1):356, December 2022.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. 2019.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1), jan 2020. ISSN 1532-4435.

Rényi, A. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pp. 547–561, Berkeley, California, 1961. University of California Press.

Reuben, A., Zhang, J., Chiou, S.-H., Gittelman, R. M., Li, J., Lee, W.-C., Fujimoto, J., Behrens, C., Liu, X., Wang, F., Quek, K., Wang, C., Kheradmand, F., Chen, R., Chow, C.-W., Lin, H., Bernatchez, C., Jalali, A., Hu, X., Wu, C.-J., Eterovic, A. K., Parra, E. R., Yusko, E., Emerson, R., Benzeno, S., Vignali, M., Wu, X., Ye, Y., Little, L. D., Gumbs, C., Mao, X., Song, X., Tippen, S., Thornton, R. L., Cascone, T., Snyder, A., Wargo, J. A., Herbst, R., Swisher, S., Kadara, H., Moran, C., Kalhor, N., Zhang, J., Scheet, P., Vaporciyan, A. A., Sepesi, B., Gibbons, D. L., Robins, H., Hwu, P., Heymach, J. V., Sharma, P., Allison, J. P., Baladandayuthapani, V., Lee, J. J., Davis, M. M., Wistuba, I. I., Futreal, P. A., and Zhang, J. Comprehensive T cell repertoire characterization of non-small cell lung cancer. *Nature Communications*, 11(1):603, January 2020.

Rizzuto, G. A., Merghoub, T., Hirschhorn-Cymerman, D., Liu, C., Lesokhin, A. M., Sahawneh, D., Zhong, H.,

Panageas, K. S., Perales, M.-A., Altan-Bonnet, G., Wolchok, J. D., and Houghton, A. N. Self-antigen-specific CD8+ T cell precursor frequency determines the quality of the antitumor immune response. *J Exp Med*, 206(4): 849–866, March 2009.

Sennrich, R., Haddow, B., and Birch, A. Neural machine translation of rare words with subword units. In Erk, K. and Smith, N. A. (eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics.

Shannon, C. E. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

Shepherd, F. R. and McLaren, J. E. T cell immunity to bacterial pathogens: Mechanisms of immune control and bacterial evasion. *Int J Mol Sci*, 21(17), August 2020.

Shirasawa, M., Yoshida, T., Matsutani, T., Takeyasu, Y., Goto, N., Yagishita, S., Kitano, S., Kuroda, H., Hida, T., Kurata, T., and Ohe, Y. Diversity of TCR repertoire predicts recurrence after CRT followed by durvalumab in patients with NSCLC. *npj Precision Oncology*, 9(1):17, January 2025.

Shugay, M., Bagaev, D. V., Zvyagin, I. V., Vroomans, R. M., Crawford, J. C., Dolton, G., Komech, E. A., Sycheva, A. L., Koneva, A. E., Egorov, E. S., Eliseev, A. V., Van Dyk, E., Dash, P., Attaf, M., Rius, C., Ladell, K., McLaren, J. E., Matthews, K. K., Clemens, E. B., Douek, D. C., Luciani, F., van Baarle, D., Kedzierska, K., Kesmir, C., Thomas, P. G., Price, D. A., Sewell, A. K., and Chudakov, D. M. VDJdb: a curated database of t-cell receptor sequences with known antigen specificity. *Nucleic Acids Res*, 46(D1):D419–D427, January 2018.

Sidhom, J.-W., Larman, H. B., Pardoll, D. M., and Baras, A. S. DeepTCR is a deep learning framework for revealing sequence concepts within t-cell repertoires. *Nature Communications*, 12(1):1605, March 2021.

Simpson, E. H. Measurement of diversity. *Nature*, 163 (4148):688–688, April 1949.

Springer, I., Besser, H., Tickotsky-Moskovitz, N., Dvorkin, S., and Louzoun, Y. Prediction of specific tcr-peptide binding from large dictionaries of tcr-peptide pairs. *Frontiers in Immunology*, 11, 2020. ISSN 1664-3224.

Stranges, P. B. and Kuhlman, B. A comparison of successful and failed protein interface designs highlights the challenges of designing buried hydrogen bonds. *Protein Sci*, 22(1):74–82, November 2012.

Stuart, K., Brun, R., Croft, S., Fairlamb, A., Gürtler, R. E., McKerrow, J., Reed, S., and Tarleton, R. Kinetoplastids: related protozoan pathogens, different diseases. *The Journal of Clinical Investigation*, 118(4):1301–1310, 4 2008.

Stuckey Peter V. and Santiago-Tirado Felipe H. Fungal mechanisms of intracellular survival: what can we learn from bacterial pathogens? *Infection and Immunity*, 91(9): e00434–22, July 2023.

Tareen, A. and Kinney, J. B. Logomaker: beautiful sequence logos in python. *Bioinformatics*, 36(7):2272–2274, April 2020.

Tickotsky, N., Sagiv, T., Prilusky, J., Shifrut, E., and Friedman, N. McPAS-TCR: a manually curated catalogue of pathology-associated T cell receptor sequences. *Bioinformatics*, 33(18):2924–2929, September 2017.

Tonegawa, S. Somatic generation of antibody diversity. *Nature*, 302(5909):575–581, April 1983.

Tong, Y., Wang, J., Zheng, T., Zhang, X., Xiao, X., Zhu, X., Lai, X., and Liu, X. SETE: Sequence-based ensemble learning approach for TCR epitope binding prediction. *Comput Biol Chem*, 87:107281, June 2020.

Trolle, T., McMurtrey, C. P., Sidney, J., Bardet, W., Osborn, S. C., Kaever, T., Sette, A., Hildebrand, W. H., Nielsen, M., and Peters, B. The length distribution of class i–restricted t cell epitopes is determined by both peptide supply and mhc allele–specific binding preference. *The Journal of Immunology*, 196(4):1480–1487, 02 2016. ISSN 0022-1767.

Twyman-Saint Victor, C., Rech, A. J., Maity, A., Rengan, R., Pauken, K. E., Stelekati, E., Benci, J. L., Xu, B., Dada, H., Odorizzi, P. M., Herati, R. S., Mansfield, K. D., Patsch, D., Amaravadi, R. K., Schuchter, L. M., Ishwaran, H., Mick, R., Pryma, D. A., Xu, X., Feldman, M. D., Gangadhar, T. C., Hahn, S. M., Wherry, E. J., Vonderheide, R. H., and Minn, A. J. Radiation and dual checkpoint blockade activate non-redundant immune mechanisms in cancer. *Nature*, 520(7547):373–377, April 2015.

Vita, R., Mahajan, S., Overton, J. A., Dhanda, S. K., Martini, S., Cantrell, J. R., Wheeler, D. K., Sette, A., Sette, A., Peters, B., and Peters, B. The immune epitope database (iedb): 2018 update. *Nucleic Acids Research*, 47, 2019.

Vujović, M., Marcatili, P., Chain, B., Kaplinsky, J., and Andresen, T. L. Signatures of T cell immunity revealed using sequence similarity with TCRDivER algorithm. *Communications Biology*, 6(1):357, March 2023a.

Vujović, M., Marcatili, P., Chain, B., Kaplinsky, J., and Andresen, T. L. Signatures of T cell immunity revealed using sequence similarity with TCRDivER algorithm. *Communications Biology*, 6(1):357, March 2023b.

Walker, D. M., Oghumu, S., Gupta, G., McGwire, B. S., Drew, M. E., and Satoskar, A. R. Mechanisms of cellular invasion by intracellular parasites. *Cell Mol Life Sci*, 71 (7):1245–1263, November 2013.

Weber, A., Born, J., and Rodriguez Martínez, M. TITAN: T-cell receptor specificity prediction with bimodal attention networks. *Bioinformatics*, 37(Supplement˙1):i237–i244, 07 2021. ISSN 1367-4803.

Wilkins, M. R., Gasteiger, E., Bairoch, A., Sanchez, J. C., Williams, K. L., Appel, R. D., and Hochstrasser, D. F. Protein identification and analysis tools in the ExPASy server. *Methods Mol Biol*, 112:531–552, 1999.

Yang, J., He, B., Zhao, Y., Jiang, F., Wang, Z., Guo, Y., Xu, Z., Yuan, B., Song, J., Zhang, Q., and Yao, J. De novo generation of t-cell receptors with desired epitope-binding property by leveraging a pre-trained large language model. *bioRxiv*, 2023. doi: 10.1101/2023.10.18.562845. URL https://www.biorxiv.org/content/early/2023/10/20/2023.10.18.562845.

Zhang, H., Zhan, X., and Li, B. GIANA allows computationally-efficient TCR clustering and multi-disease repertoire classification by isometric transformation. *Nature Communications*, 12(1):4699, August 2021a.

Zhang, J., Ma, W., and Yao, H. Accurate TCR-pMHC interaction prediction using a BERT-based transfer learning method. *Briefings in Bioinformatics*, 25(1):bbad436, 12 2023. ISSN 1477-4054.

Zhang, W., Wang, L., Liu, K., Wei, X., Yang, K., Du, W., Wang, S., Guo, N., Ma, C., Luo, L., Wu, J., Lin, L., Yang, F., Gao, F., Wang, X., Li, T., Zhang, R., Saksena, N. K., Yang, H., Wang, J., Fang, L., Hou, Y., Xu, X., and Liu, X. PIRD: Pan immune repertoire database. *Bioinformatics*, 36(3):897–903, 2019.

Zhang, W., Hawkins, P. G., He, J., Gupta, N. T., Liu, J., Choonoo, G., Jeong, S. W., Chen, C. R., Dhanik, A., Dillon, M., Deering, R., Macdonald, L. E., Thurston, G., and Atwal, G. S. A framework for highly multiplexed dextramer mapping and prediction of t cell receptor sequences to antigen specificity. *Science Advances*, 7(20): eabf5835, 2021b.

Zhou, Z., Chen, J., Lin, S., Hong, L., Wei, D.-Q., and Xiong, Y. Gratcr: Epitope-specific t cell receptor sequence generation with data-efficient pre-trained models. *IEEE Journal of Biomedical and Health Informatics*, 29 (3):2271–2283, 2025. doi: 10.1109/JBHI.2024.3514089.

A. Performance of RAP

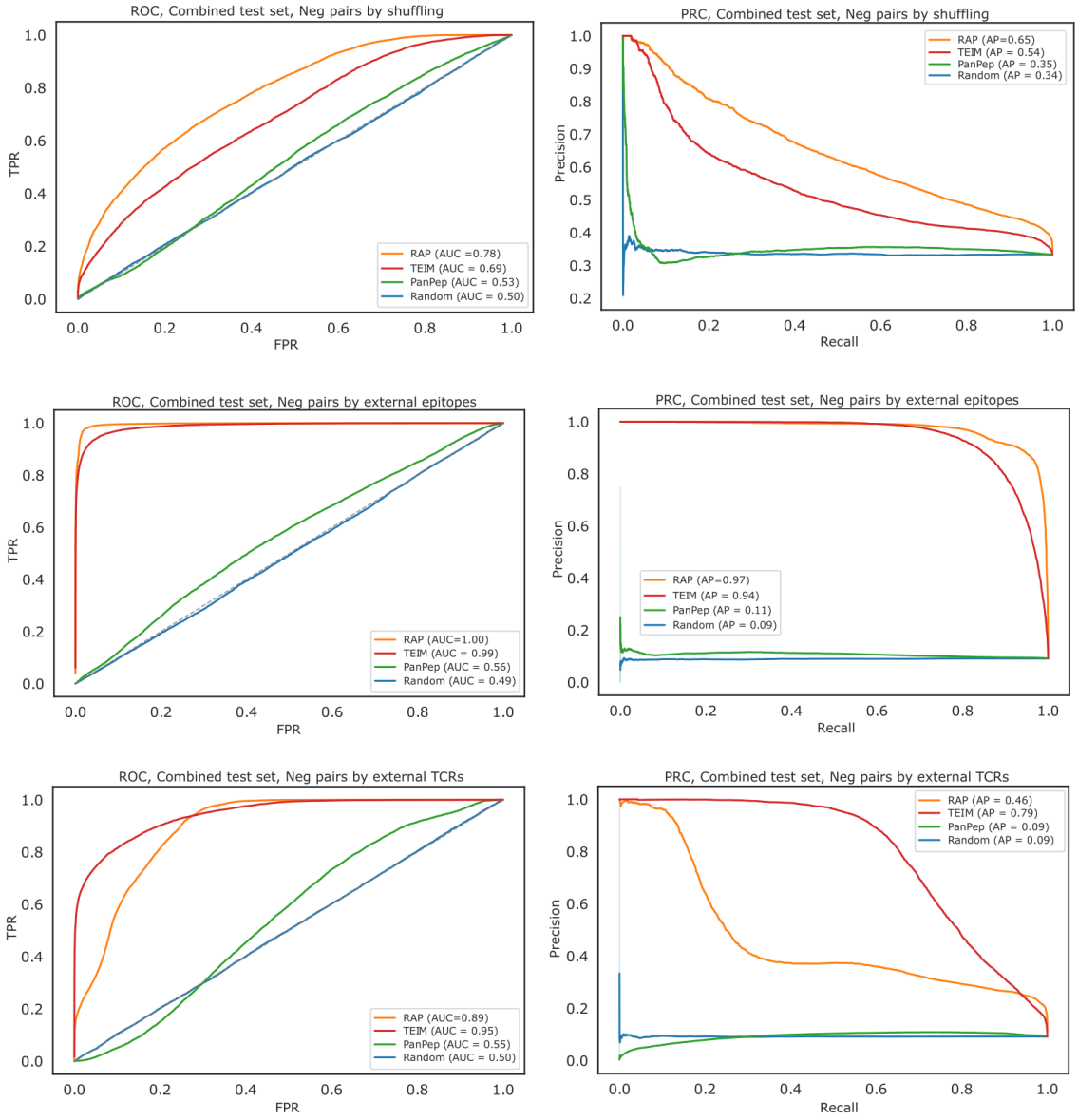

Figure 8. AUROC and AUPRC of RAP and other leading binding affinity predictors.

Figure 8 presents the performance of RAP compared to several state-of-the-art TCR–peptide binding affinity predictors, evaluated across three distinct test setups. These test sets differ in how negative (non-binding) TCR–peptide pairs were constructed. In the first setup, negative samples were generated by randomly shuffling TCR and peptide sequences, under the assumption that such arbitrary pairs are unlikely to bind. The second setup uses test-set TCRs paired with epitope sequences not seen during training, mitigating potential overfitting to peptide sequences. In the third setup, peptides from the test set are paired with previously unseen TCRs, aiming to reduce overfitting to TCR sequences. Each setup was evaluated using AUROC and AUPRC as performance metrics. RAP consistently outperforms all baselines in the first two setups and achieves the second-best performance in the third. These results highlight the robustness and competitiveness of RAP across a range of evaluation scenarios.

B. Datasets

B.1. Public paired datasets

Four publicly available datasets of VDJdb (Shugay et al., 2018), PIRD (Zhang et al., 2019), IEDB (Vita et al., 2019), and McPAS-TCR (Tickotsky et al., 2017) were used to develop the Robust Affinity Predictor (RAP) and evaluate the generated epitopes.

VDJdb is a curated database of T-cell receptor (TCR) sequences with known antigen specificities. This resource aggregates data from previously published studies, providing a comprehensive collection of TCR-antigen interactions. The dataset contains rich metadata that offers detailed information about each TCR-antigen pair, including TCR information (gene, cdr3, v.segm, j.segm, v.end, and j.start), antigen information (antigen.epitope, antigen.gene, and antigen.species), MHC context (mhc.a, mhc.b, mhc.class, and complex.id), quality control and others (species, reference.id, vdjdb.score, and label). Among these, vdjdb.score is a confidence score (0-3) assigned to each entry, reflecting the reliability of the association between the TCR and the antigen in that entry.

The Immune Epitope Database (IEDB) is a comprehensive repository of experimentally determined immune epitope data, encompassing information on both B-cell and T-cell epitopes from human and animal model studies. The database covers various immunological contexts, including infections, allergies, autoimmune diseases, and transplantations. We utilized the IEDB database export v3 (`https://www.iedb.org/database_export_v3.php`), which provides detailed information on epitope sequences (trimmed_seq, original_seq), immune receptor characteristics (receptor_group), antigen source details (source_organisms, source_antigens), labels, and epitopes.

The Pan Immune Repertoire Database (PIRD) was developed to provide a structured repository for sequencing data of the T-cell receptor (TCR) and B-cell receptor (BCR). Within PIRD, the T and B cell Antigen database (TBAdb) is a manually curated dataset of TCRs and BCRs with known antigen specificity. The dataset contains the following columns:disease information (ICDname, Disease.name, Category), antigen details (Antigen, Antigen.sequence, HLA), receptor sequences (Locus, CDR3.alpha.aa, CDR3.beta.aa, CDR3.alpha.nt, CDR3.beta.nt), gene usage (Valpha, Jalpha, Vbeta, Dbeta, Jbeta), experimental methods (Seq.platform, Species, Origin, Nucleotide.type, Cell.subtype, Prepare.method, Evaluate.method), study design (Case.num, Control.type, Control.num, Filteration), publication information (Journal, Pubmed.id), data quality (grade), and label. We used the CDR3.beta.aa and Antigen.sequence columns to collect the binding pairs of TCRs and epitopes.

McPAS-TCR is a manually curated database of TCR sequences associated with various pathological conditions (including pathogen infections, cancer, and autoimmunity) and their respective antigens in humans and mice. The dataset contains comprehensive information organized into the following categories: TCR sequence data (CDR3.alpha.aa, CDR3.beta.aa, CDR3.alpha.nt, CDR3.beta.nt, TRAV, TRAJ, TRBV, TRBD, TRBJ, Reconstructed.J.annotation), study characteristics (Species, Category, Pathology, Pathology.Mesh.ID, Additional.study.details, Antigen.identification.method, Single.cell, NGS, PubMed.ID), antigen and epitope information (Antigen.protein, Protein.ID, Epitope.epitope, Epitope.ID, MHC), T cell properties (Tissue, T.Cell.Type, T.cell.characteristics), and additional information (Mouse.strain, Remarks, label).

B.2. Unpaired datasets

Four unpaired datasets (TCRdb (Chen et al., 2021), NetMHCPan v4.0 (Jurtz et al., 2017), MHCFlurry v2.0 (O'Donnell et al., 2020), and SysteMHC (Huang et al., 2023)) were utilized to collect large-scale unpaired CDR3β sequences and epitope sequences for semi-supervised learning. A subset of these datasets was also used for Triple Negative Sampling (TNS) to construct diverse negative samples for the training and testing of the Robust Affinity Predictor (RAP).

TCRdb, a comprehensive structured collection of TCR sequencing experiments, contains 131 TCR-seq projects, 8,265 TCR-seq samples, and 277,439,349 TCR CDR3 sequences as of August 2024. From the downloadable dataset (`https://guolab.wchscu.cn/TCRdb/#/download`), we obtained 7,331,478 unique CDR3β sequences after preprocessing. The dataset was split into TCRNetSet for training the Robust Affinity Predictor and TCRCandidateSet for BINDSEARCH.

NetMHCPan v4.0 is a predictive model for interactions between class I MHC alleles (represented as MHC pseudo sequences) and epitopes. The project's dataset (`https://services.healthtech.dtu.dk/suppl/immunology/NetMHCpan-4.0`) comprises two components: Binding Affinity (BA) and Eluted Ligand (EL) data. BA data, derived from *in vitro* binding assays, provide quantitative IC50 values that measure the binding strength between epitopes and MHC molecules. EL data, obtained through mass spectrometry (MS) experiments, provide information on naturally processed and

presented epitopes, capturing aspects of antigen processing and presentation beyond binding affinity.

MHCFlurry v2.0 is another predictive model for epitope-MHC interactions, which divided the task into antigen presentation prediction and binding affinity prediction. The study compiled datasets from affinity measurements and mass spectrometry experiments, establishing several benchmarks as detailed in the Methods section of the original paper (O'Donnell et al., 2020).

SysteMHC is an archive of MS-based immunopeptidomics data. Regarding the class I MHC interaction, the atlas includes more than 4,680 MS raw files, 154 allele types, 1 million epitopes, and 457,360 HLA-bound epitopes. The per-allele data in SpectraST format were downloaded from `https://systemhc.sjtu.edu.cn/download`.

The epitopes from NetMHCPan v4.0, MHCFlurry v2.0, and SysteMHC were merged and split into EpiNegSet for training the Robust Affinity Predictor and EpiCandidateSet for BINDSEARCH.

B.3. Pseudo-labeled datasets

These datasets were constructed by evaluating the binding affinities between TCRs from TCRCandidateSet ($I = 6,831,478$) and epitopes from EpiCandidateSet ($J = 20,000,000$). For each TCR, $\beta = 10,000$ epitopes were randomly sampled without replacement from EpiCandidateSet and ranked by binding affinity. The top $n_{\text{max\_tcr}} = 32$ epitopes with the highest binding affinities were selected for each TCR, yielding $I \times n_{\text{max\_tcr}} = 218,607,296$ pairs. Pairs containing redundant epitopes occurring more than $n_{\text{max\_epi}} = 100$ times in the dataset were excluded. While the application of the Antigen Category Filter decreased the size of the pseudo-labeled dataset, it was a necessary quality control step for applying EpitopeGen to repertoire-level datasets.

B.4. External test sets

The dataset released by Glanville et al. (Glanville et al., 2017) comprises TCR sequences from antigen-specific T cells isolated from multiple individuals. Their training set, which we used as an external test set, includes 2,068 unique TCRs of known specificity, derived from T cells sorted using eight different epitope-MHC tetramers in 33 donors.

The dataset released by Nolan et al. (Nolan et al., 2020) provides a large-scale collection of T-cell receptor beta (TCRβ) sequences and their binding associations to the SARS-CoV-2 epitopes. We specifically used their Multiplex Identification of Antigen-Specific T-Cell Receptors Assay (MIRA) dataset, which contains over 135,000 high-confidence SARS-CoV-2-specific TCRs. This dataset maps TCRs binding to SARS-CoV-2 virus epitopes from exposed subjects and naïve controls, offering a comprehensive view of TCR-epitope interactions in the context of COVID-19. We constructed a test set by sampling 2,000 pairs of TCR and epitope by processing `peptide-detail-ci.csv` of `ImmuneCODE-MIRA-Release002.1`.

B.5. 10x CD8$^+$ datasets

10x Genomics Inc. published many single-cell sequencing datasets using their platforms. We used a dataset containing CD8$^+$ T cells from a healthy donor (10x Genomics, 2022), obtained using the Single Cell Immune Profiling platform and analyzed using Cell Ranger 3.0.2. This dataset serves as a test set to simulate a real-world application scenario, distinct from those used in EpitopeGen's training. The model is challenged to generate epitopes for CD8$^+$ T cells based on single-cell TCR sequencing data from an individual patient with specific MHC alleles. This approach assesses the model's ability to generate diverse epitopes with a natural antigen distribution.

C. Justification for antigen target ratios in Antigen Category Filter

The distribution of antigen categories in Antigen Category Filter (ACF) significantly influences EpitopeGen's characteristics. Although the exact ratio of antigens recognized by $CD8^+$ T cells *in vivo* remains undefined, we can estimate this distribution based on the current immunological understanding.

$CD8^+$ T cells respond primarily to endogenous antigens presented via MHC class I molecules, with viruses representing the predominant category due to their intracellular replication cycle. This intracellular lifecycle leads to extensive endogenous epitope generation and presentation, making viruses the primary targets for $CD8^+$ T cell surveillance. Several studies support this viral dominance: Masopust et al. (Masopust et al., 2007) demonstrated that approximately 80% of splenic $CD8^+$ T cells recognized lymphocytic choriomeningitis virus (LCMV)-derived epitopes during peak primary infection. Moutaftsi et al. (Moutaftsi et al., 2006) found that 29.6% of $CD8^+$ T cells produced IFN-γ when exposed to vaccinia virus Western Reserve strain (VACV-WR)-infected cells. Furthermore, Addo et al. (Addo et al., 2003) reported robust responses of 10,640 spot-forming cells per million PBMC in untreated chronically infected individuals. Based on these references, we set the proportion of viral antigens to be P(virus) $\geq$ 50%.

Bacteria primarily elicit $CD4^+$ T cell and B cell responses due to their predominantly extracellular nature. However, certain bacterial species have evolved intracellular survival strategies. As Shepherd et al. (Shepherd & McLaren, 2020) documented, bacteria such as *Listeria monocytogenes* and *Shigella flexneri* directly target the host cell cytosol, while *Mycobacterium tuberculosis* and *Salmonella* persist in vacuolar compartments. Even traditional 'extracellular pathogens such as *Staphylococcus aureus* can invade intracellular spaces (Friot et al., 2023), although they represent a smaller proportion of targets of $CD8^+$ T cells compared to viruses. We set bacterial antigens to P(bacteria) $\leq$ P(virus) $\times$ 0.2.

$CD8^+$ T cells that respond to self-antigens and tumor-associated antigens occur at significantly lower frequencies compared to those that target external pathogens, mainly due to thymic selection mechanisms that maintain immune tolerance (Kenison et al., 2024). Rizzuto et al. (Rizzuto et al., 2009) demonstrated that self- / tumor antigen-specific T cells exist at frequencies significantly lower than those specific for foreign antigens, attributing this difference to negative selection in the thymus, a crucial process for preventing autoimmunity. Nelson et al. (Nelson et al., 2019) reported that initial self-reactive T cells are exceptionally rare and difficult to detect prior to antigenic boost, contrasting with the robust responses generated against viral antigens. In quantitative terms, Pittet et al. (Pittet et al., 1999) determined that approximately 1 in 2,500 naive $CD8^+$ T cells recognize tumor (melanoma) antigens, with remarkably similar proportions observed in healthy donors, suggesting that this may represent a baseline frequency for self-antigen recognition. We therefore constrained the self-antigen ratio in the range of 0.03-0.15 and the tumor antigen ratio in the range of 0.01-0.05.

Fungi and protozoa constitute a more limited set of potential antigens, despite some species causing intracellular infections. Mittal et al. (Mittal et al., 2019) described how *Histoplasma capsulatum* can infect human cells, while Stuckey et al. (Stuckey Peter V. & Santiago-Tirado Felipe H., 2023) estimated that among the 3–5 million fungal species thought to exist, only a few dozen regularly cause human infections. This led us to restrict their combined proportion to P(fungi + parasites) $\leq$ P(virus) $\times$ 0.1.

Notably, archaea, despite their presence as human commensals, have not demonstrated pathogenic capabilities and generally do not trigger $CD8^+$ T cell responses (Cavicchioli et al., 2003; Gill & Brinkman, 2011). We kept the proportion of archaeal antigens in our dataset very low (P(archaea)$\leq$0.01).

This evidence supports five key immunological principles that govern ACF design: (1) Viral Dominance, (2) Limited Bacterial Presence, (3) Endogenous Epitope Presence (Self-Antigen and Tumor-Associated Antigen), (4) Rare Fungi and Parasites, and (5) Absence of Pathogenic Archaea.