TCR-TRANSLATE: CONDITIONAL GENERATION OF REAL ANTIGEN-SPECIFIC T-CELL RECEPTOR SEQUENCES

Dhuvarakesh Karthikeyan [*,†,‡]
Bioinformatics and Computational Biology Program
University of North Carolina, Chapel Hill
dkarthikeyan1@unc.edu

Colin Raffel [§]
Department of Computer Science
University of Toronto
craffel@cs.toronto.edu

Benjamin Vincent [*,†,¶,‖]
Department of Medicine, Hematology Div.
University of North Carolina, Chapel Hill
benjamin_vincent@unc.edu

Alexander Rubinsteyn [*,†,¶,‖]
Department of Genetics
University of North Carolina, Chapel Hill
alex.rubinsteyn@unc.edu

ABSTRACT

The paradoxical nature of T-cell receptor (TCR) specificity, which requires both precise recognition and adequate coverage of antigenic peptide-MHCs (pMHCs), poses a fundamental challenge in immunology. Efforts at modeling this complex many-to-many mapping have focused on the detection of reactive TCR-pMHC pairs as a binary classification task, with little success on unseen epitopes. Here, we present TCR-TRANSLATE, a framework that adapts low-resource machine translation techniques including semi-synthetic data augmentation and multi-task objectives to generate target-conditioned CDR3β sequences for unseen input pMHCs. We examine twelve model variants derived from the BART and T5 model architectures on a target-rich validation set of well-studied antigens, finding an optimal model, TCRT5, that samples known and native-like CDR3β sequences for unseen epitopes. Our findings highlight both the potential and limitations of sequence-to-sequence modeling in rapidly generating antigen-specific TCR repertoires, emphasizing the need for experimental validation to bridge the gaps between predictions, metrics, and functional capacity.

1 INTRODUCTION

T-cell receptors (TCRs) are highly specific, stochastically generated pattern recognition receptors that enable the immune system's T cells to recognize nonself cells such as those that are infected or malignant. These TCRs interact with intracellular peptides presented on major histocompatibility complexes (MHCs) at the surface of most somatic cells, giving rise to a network of TCR:peptide-MHC (pMHC) specificities capable of self-nonself discrimination with single-amino acid precision (Kalergis et al., 1999). T cell based therapies including CAR-T, engineered TCRs, and TCR bispecifics have shown durable treatment across a variety of indications (Tzannou et al., 2017; Chung et al., 2024; Harrison, 2024) but are bottlenecked by laborious and low-yield *in-vitro* TCR discovery platforms for identifying specific and self-tolerant TCRs (Liu et al., 2022). *In-silico* methods to decipher the TCR::pMHC mapping have the potential to transform the field of precision immunotherapy by operationalizing a potent mechanism of targeting cells at the sub-protein resolution.

[*]Personalized Immunotherapy Research Lab
[†]Computational Medicine Program, UNC School of Medicine
[‡]Molecular and Cellular Biophysics Program, UNC School of Medicine
[§]Vector Institute
[¶]Bioinformatics and Computational Biology, UNC School of Medicine
[‖]Lineberger Comprehensive Cancer Center

Experimentally, individual TCRs have been shown to recognize up to 10^6 unique peptides (Wooldridge et al., 2011), and vice versa (Bentzen et al., 2018; Sewell, 2012). The resulting many-to-many mapping is further confounded by sparse and biased paired experimental data, with most antigen-specific TCRs being identified in the context of only a few, well-studied diseases (Hudson et al., 2023). Current approaches in modeling antigen-specificity revolve around the CDR3β loop of the TCR, a highly variable portion of the TCR that plays a strong role in determining antigen specificity, and frame the problem as a binary classification task (Gao et al., 2023; Nielsen et al., 2024; Zhang et al., 2024b)-with limited utility in TCR design (Wu et al., 2021). Prior work on generative models focus on unconditional generation of TCR sequences that recapitulate repertoire level statistics in the aggregate (Davidsen et al., 2019; Isacchini et al., 2021). More recent work on conditional generation in Fast et al. (2023) demonstrates the potential for sequence based generation strategies using a 1-D CNN encoder and LSTM decoder to sample CDR3β sequences against a known antigen. Though transformer architectures for this task were introduced in Yang et al. (2023) and Karthikeyan et al. (2023), and related research has since emerged in Zhou et al. (2024) and Lin et al. (2024), a deep understanding of the real-world utility and limitations of conditional TCR sequence generation remains elusive.

In this work, we adopt the framework from Karthikeyan et al. (2023) (See B.1) and systematically trained twelve sequence-to-sequence (seq2seq) (Figure 1a) model variants of the BART (Lewis et al., 2019) and T5 (Raffel et al., 2020) architectures (Figure 1b) using techniques derived low-resourced machine translation (Haddow et al., 2022). Specifically, we sought to leverage the reflexive nature of sequence co-dependencies between source-target pairs by jointly learning a bidirectional mapping (Niu et al., 2018; Yang et al., 2019), sharing representations and aligning latent spaces across both sequences (Ding et al., 2021). To the best of our knowledge, these approaches have not been applied to the functional protein design domain.

We deeply characterize the performance of our models under optimal conditions of our validation dataset comprising the top-20 pMHCs with the most known cognate TCRs (Figure 1c-d, B.2). By forfeiting their inclusion in the training data to maximize exact sequence matches during evaluation, we find some methods increased performance at the cost of diversity, driven by the generation of polyspecific TCRs observed to bind multiple unrelated pMHCs. This along with the conserved model performance across pMHCs, highlight the critical role of training data composition and its relationship to the validation set. Finally, in order to evaluate performance of our model in a simulated real-world scenario, we demonstrate better-than-random performance of our flagship model on a subset of test antigens from last year's IMMREP2023 TCR specificity challenge not seen during training or model selection. Furthermore, these epitopes are highly dissimilar to epitopes seen during training with a minimum edit distance of 5. Our results emphasize the potential and qualify the drawbacks of sequence-to-sequence modeling for sampling real antigen-specific TCR sequences in a severely data-constrained setting.

2 RESULTS

For our *in-silico* exploration and validation experiments, we considered three different training schemes for the BART and T5 models each, stratified by their pre-training status for a total of twelve model variants (See B.3). All models were evaluated on CDR3β sequence generation (Figure 2a). Our baseline models (TCRBART-0 and TCRT5-0) were trained on the pMHC $\rightarrow$ TCR direction with no pre-training. The bidirectional models (TCRBART-0 (B) and TCRT5-0 (B)) were trained on conditional sequence generation in both directions (pMHC $\leftrightarrow$ TCR). Finally the multi-task models were trained on both directions as well a masked language modeling loss term for both TCR and pMHC sequences (TCRBART-0 (M) and TCRT5-0 (M)). Similarly, six models were pre-trained and then finetuned using the same learning tasks to add TCRBART-FT, TCRBART-FT (B), TCRBART-FT (M), TCRT5-FT, TCRT5-FT (B), and TCRT5-FT (M) (Figure 2b). Model checkpoints were chosen according to B.6, to compare training paradigms by their best representative model performances.

We first calibrate our metrics by benchmarking conditional models $P(TCR|pMHC)$ against unconditional generation $P(TCR)$. We evaluated our baseline conditional models TCRBART-0 and TCRT5-0 on a reduced set of metrics to determine the advantage of target-conditioning. As our unconditional baseline, we used soNNia's `'Ppost'` (Isacchini et al., 2021), a generative model that extends V(D)J recombination to include thymic selection, sampling a TCR distribution closer

to what is observed in the periphery. In addition, to investigate our training set composition's impact on validation performance, we evaluated sequences from TCRBART-0 and TCRT5-0 derived in an input-free manner (TCRBART-Unconditional, TCRT5-Unconditional). As expected, we found that the conditional models outperformed unconditional approaches across all metrics except global diversity (Figure 1e). Surprisingly, TCRT5-Unconditional achieved non-zero F1s, revealing high-likelihood training CDR3β sequences in the validation set.

2.1 Multi-task training increases accuracy metrics while decreasing diversity metrics of generated sequences

Under our evaluation framework, no models outperformed the others on all metrics across all pMHCs, making identifying an optimal model and training mode difficult (Table 1). For example, while mean and median sequence recoveries increased for bidirectional and multi-task variants, their Char-BLEU scores decreased (Figure 2c). On F1 score, some models excelled on a small subset of examples and others showed marginal improvements across a broader set, observed as a divergence in the mean and median scores. Reassuringly, however, all training procedures maintained or improved F1 performance for over 50% of validation pMHCs over baseline (Figure 2d). Using mean average precision (mAP) to assess calibration, we found that the bidirectional models outperformed baseline, and both outperformed multi-task variants (Figure 2e). Diversity metrics, however, revealed a decline in unique sequences generated across pMHCs going from the baseline models to the bidirectional and multi-task ones. This was most evident for TCRBART-0 (M), which had strong accuracy metrics despite a drop of over 80% in unique sequences generated (Figure 2f), highlighting the importance of using both accuracy and diversity between pMHCs to represent model performance and demonstrate learned input sensitivity.

Therefore, to holistically characterize the models on accuracy and diversity, we visualized performance on a biaxial plot of average F1 score and mean pairwise Jaccard dissimilarity. We chose the average F1 score to summarize model accuracy for its sensitivity to capturing positive outliers, a useful property given the sparsity of data availability. Viewed through this lens, we found that while pre-training and finetuning pushed the diversity/accuracy pareto front for the TCRT5-FT variants, we observed the complete opposite effect of degradation in both model performance and diversity for the TCRBART-FT models (Figure 2g). Since both TCRBART-FT and TCRT5-0 generated less than 10% of the maximum number of unique sequences, with average Jaccard dissimilarities of less than 0.5, we fix the TCRBART-0 and TCRT5-FT variants as the best BART and T5 models, and restrict further analyses to these models. However, the differences between the baseline, bidirectional, and multi-task models of TCRBART-0 and TCRT5-FT were less obvious. Crucially, remained the fact that the bidirectional and multi-task model variants generated fewer sequences and still improved performance. When we examined the generated sequences, we saw an enrichment for empirically de-risked CDR3β sequences, which we suspect yielded low loss as a binder to many training pMHCs and incidentally multiple validation examples (Figure S3).

2.2 Multi-task models preferentially sample polyspecific CDR3β sequences via training set statistics

While some level of TCR cross-reactivity is essential to TCR function, distinctly 'polyspecific' TCRs (Wucherpfennig et al., 2007) bind sufficiently unrelated pMHCs (Figure 3a). While Quiniou et al. (2023) lay out the criteria for a polyspecific TCR as: (i) possessing higher probabilities of generation (ii) sampling particular V and J genes at a higher rate (iii) shared CDR3 sequences between individuals, and (iv) activation by multiple unrelated peptides, we identify an 'ML-centric' notion of polyspecificity and identify CDR3β sequences found in multiple disease conditions and bound more than two epitopes (n=915 CDR3β sequences). To understand competitive performance at a fraction of the diversity we qualify the translations' polyspecificity status. Between TCRBART-0 and TCRT5-FT, we found that the multi-task models generated more polyspecific CDR3β sequences for both TCRBART-0 ($p_{bidxn} = 0.048$, $p_{multi} < 0.0001$) and TCRT5-FT ($p_{bidxn} = 0.002$, $p_{multi} = 0.009$) and their mean polyspecificity increased as well (Figure 3b-c). We observed an inverse correlation between the number of polyspecific TCRs and unique sequences generated (Pearson's r: -.957).

Table 1: Performance Metrics on Top-20 Validation pMHCs. Mean/Median values are reported where applicable. Best in class metric is highlighted in bold.

Model	Char-BLEU (↑)	F1@100 (↑)	%Rec. (↑)	mAP (↑)	N_{unique} (↑)
TCRBART-0	**93.5**	**.057**/.010	81.7/81.5	0.163	1276
TCRBART-0 (B)	93.0	.055/.015	81.9/83.9	**0.196**	**1302**
TCRBART-0 (M)	91.7	.042/**.030**	**84.3/85.9**	0.140	240
TCRBART-FT	87.4	.013/0.00	84.1/85.1	0.049	**240**
TCRBART-FT (B)	**89.8**	.014/**.010**	83.5/84.6	**0.062**	185
TCRBART-FT (M)	86.1	**.023/.010**	**84.5/85.0**	0.048	127
TCRT5-0	**89.5**	.040/.030	83.2/84.7	0.142	129
TCRT5-0 (B)	34.8	**.054/.035**	84.5/**85.9**	**0.167**	139
TCRT5-0 (M)	37.8	**.054**/.020	**84.7/85.9**	0.129	**177**
TCRT5-FT	**96.4**	**.091/.020**	**84.9**/85.0	0.246	**1300**
TCRT5-FT (B)	93.5	.083/.015	84.6/**85.1**	**0.279**	933
TCRT5-FT (M)	94.4	.082/**.020**	**84.7**/84.6	0.180	833

However, given our dataset's deduplication step, these polyspecific CDR3β sequences were also more represented than those with fewer known binders. To determine if the models were parroting sequences seen most during training at similar frequencies, we examined the translations' rank order against potential explanatory variables such as polyspecificity, number of cognate epitopes/alleles, and training set incidence. We found that while the highly ranked sequences were more common in the training set, they also had more dissimilar known cognate epitopes, suggesting robustness in capturing polyspecificity (Figure S4a-b). Regression analysis of occurrence across validation pMHCs and training frequency showed that this relationship was surprisingly attenuated in the multi-task models with Pearson's correlation coefficients of 0.41, 0.42, and 0.3 for TCRT5-FT, TCRT5-FT (B), and TCRT5-FT (M), respectively and 0.43, 0.35, and 0.2 for TCRBART-FT, TCRBART-FT (B), and TCRBART-FT (M) as well (Figure S4c). These results suggest that the multi-task models are sampling more polyspecific CDR3β sequences, somewhat independent of training frequency.

Since our validation set is comprised of highly immunogenic viral peptides known to be the targets of polyspecific TCRs, we checked if our F1 performance could be explained solely by polyspecific generations. Although many of the models' translations across pMHCs were polyspecific with high model likelihoods, we found that the baseline models sampled both more non-polyspecific and non-polyspecific true positive binders than the bidirectional and multi-task models (Figure 3e). Given our desire for a model that generates CDR3β sequences to rare epitopes, we find polyspecific TCR generation a potentially misleading avenue for metric hacking, misrepresenting true usefulness. Thus, while the bidirectional and multi-task models show promise in increasing accuracy through self-consistency for the receptor:ligand design problem, we note that their utility may be limited given the current asymmetrically sampled TCR:pMHC landscape. We therefore select TCRT5-FT as our flagship model and henceforth refer to it simply as TCRT5.

2.3 TCRT5 GENERATES REAL UNSEEN ANTIGEN-SPECIFIC CDR3β SEQUENCES

Having selected TCRT5 for its superior accuracy, diversity, and minimal reliance on polyspecific TCRs, we proceeded to assess model usefulness in a more qualitative manner. First, to evaluate how well TCRT5's translations reflected the global statistics of our reference set, we examined the distributions of CDR3β lengths and generation probabilities (Figure 4a). TCRT5 captures CDR3β lengths with a slight decrease in spread (mean: 14.6, sd: 1.2) compared to the reference set (mean: 14.5, sd: 2.0). However, its generations had a significantly higher $\log p_{gen}$ (mean: -7.04, sd: 0.85) than the reference set (mean: -9.83, sd: 2.356), indicating TCRT5 was missing lower probability sequences. This reduction in repertoire diversity was also captured by various sequence embedding models (Figure S5a-c) (Jiang et al., 2023; Zhang et al., 2024a). To determine whether this effect stemmed from our choice of decoding algorithm, we compared p_{gen}'s from beam search and ancestral sampling against reference CDR3βs and found beam search shifted the distribution towards

sequences of higher biological likelihood than ancestral and reference (Figure S6a). Interestingly, these p_{gen} values correlated with model log-likelihood scores (Figure S6b).

We then assessed the intra-sequence diversity of the individual translations for the validation pMHCs. Sequence logo plots of the cognate sequences for well-known canonical epitopes revealed a decrease in diversity, particularly near the start of the sequence (Figure 4b), a loss of entropy quantified using the positional Δentropy value (Figure 4c), likely due to the bias of starting sequences with the 'CASS' motif. Additionally, we used the Jensen-Shannon divergence to assess shifts in k-mer frequencies between the reference and translation sequences. SoNNia generated sequences demonstrated divergence from reference k-mer usage at lower values of k, while TCRT5 recovers at longer k-mer lengths and both converge at large k (Figure 4d). We found consistent results at k=1,000 (Figure S7). To determine TCRT5's input-specificity, we computed the Jaccard index to assess overlap between translation sequences across pMHCs (Figure 4e) and found sequences with high similarity clustered together, such as melanoma antigens EAAGIGILTV and ELAGIGILTV, though more data is required to determine to what extent this is true. Finally, we compared generation probabilities, polyspecificity, and training set frequency and found higher generation probability sequences were more frequently sampled, with no clear correlation between training frequency and increased sampling (Figure 4f).

Finally, to test whether TCRT5 could generate known binders not seen during training, we stratified the validity (OLGA $p_{gen} > 0$), known specificity, and training set membership of each of the 100 x 20 = 2,000 input:translation pairs and found that of the 2,000 generations, 1,996 of them had nonzero generation probabilities, 181 were known binders, and 7 were TCRs that were not seen during the supervised training (Figure 4g). Notably, one of these seven was not found in the pre-training set, indicating a real potential for sampling *de novo* TCRs spanning multiple pMHCs: KLGGALQAK (A*03:01), LLWNGPMAV (A*02:01), YLQPRTFLL (A*02:01), and YVLDHLIVV (A*02:01), demonstrating that the performance wasn't localized to a single pMHC.

2.4 TCRT5 ACHIEVES NON-RANDOM PERFORMANCE ON SPARSELY VALIDATED EPITOPES

The goal of a TCR design model like TCRT5 is to sample TCRs against rare epitopes not seen during training, especially when few or no known TCRs exist. As highlighted in the recent IMMREP2023 TCR specificity competition, models for binary prediction often struggle to outperform random predictors in this regime (Nielsen et al., 2024). We sought to evaluate TCRT5 in this context by generating CDR3β sequences for unseen 'private' epitopes from the IMMREP dataset that were absent our training and validation set (FTDALGIDEY, SALPTNADLY, TSDACMMTMY, TDLGQNLLY-all presented on HLA-A*01:01) (Figure 5a, See Table S2). While there are no publicly available models that condition on peptide-MHC information, we compare our model against the unconditional soNNia (Isacchini et al., 2021) as well as GRATCR (Zhou et al., 2024) and ER-BERT (Yang et al., 2023), which both sample CDR3β sequences from epitopes in an MHC-agnostic manner. For these analyses we generated 1,000 generations per pMHC and observed only TCRT5 yielded a known sequence, one of the twelve known binders for FTDALGIDEY at rank 514. We observed that while TCRT5 and GRATCR achieve comparable performance, ER-BERT had consistently lower metrics (Figure 5b). GRATCR as implemented (See B.7.1), generated a significant number of repeated sequences within each epitope compared to TCRT5 and ER-BERT (Figure S8). We compared calculated sequence recovery rates found that TCRT5 and GRATCR achieved a higher mean sequence recovery across all peptides as compared to soNNia and ER-BERT (Figure 5c). Interestingly, we found ER-BERT did not consistently generate the N-terminus cysteine or the C-terminus phenylalanine, which we suspect explains its reduced performance on sequence recovery based metrics. However, when we compared the conditional likelihoods of the known binders against the synthetic negatives from the IMMREP dataset, we found the performance between TCRT5 and ER-BERT to be comparable (Figure S9).

While mean sequence recovery of generations serves as a preliminary sanity check for antigen-specificity, we sought to further assay whether both models generated more useful individual sequences than an unconditional generator (soNNia). Since we found the dynamic range for nonzero F1 scores to occur when sequence recovery was greater than 90% (Figure S2c-e), we checked to see if any of the models generated more sequences with $\geq 90\%$ sequence identity to the true cognate sequences compared to random. We found that only TCRT5 generated sequences within this threshold for the test peptides. To determine the significance of this, we established two independent, but

complementary null distributions of our test statistic computed using: 1) soNNia generations with real reference sequences 2) TCRT5 generations with a randomized size-matched reference set. We found that for all of the test epitopes except one, TCRT5 generations showed statistical significant at a threshold of α=0.05 (Figure 5d, See B.8). For the soNNia random generations null, we observe p_1-values of $< .0001$, $.079$, $< .0001$, and $.004$ for FTDALGIDEY, SALPTNADLY, TDLGQNLLY, and TSDACMMTMY, respectively. For the random references null, we observe p_2-values of .006, .254, .028, and .024. These results indicate that while TCRT5 generates sequences closer to ground truth than random, there may be some degree of non-specific generation as as well.

3 DISCUSSION

One of the most fundamental questions in immunology is how T-cell receptors (TCRs) achieve precise recognition of "nonself" pMHCs through a complex interaction network that has posed a significant challenge to modeling efforts. Addressing this challenge would significantly advance our understanding of T cell biology, offering new insights into adaptive immune receptor specificity and providing a foundation for diverse applications in cutting-edge therapeutic modalities. Our work demonstrates the potential of generative models to sample antigen-specific repertoires with high fidelity and diversity in a data-sparse domain. This capability would have profound implications for cellular therapies, enabling the rapid generation of TCRs that can be screened for cross-reactivity with self-epitopes, streamlining the traditional labor-intensive TCR discovery process. While our study reveals the advantages and limitations of sequence-to-sequence models in capturing TCR specificity, it also underscores the need for further validation to bridge the gap between computational predictions and functional relevance. As more data becomes available, both model performance and evaluation metrics are expected to improve, moving the field closer to scalable, high-precision TCR design for precision immunotherapy.

REFERENCES

Amalie K Bentzen, Lina Such, Kamilla K Jensen, Andrea M Marquard, Leon E Jessen, Natalie J Miller, Candice D Church, Rikke Lyngaa, David M Koelle, Jürgen C Becker, Carsten Linnemann, Ton N M Schumacher, Paolo Marcatili, Paul Nghiem, Morten Nielsen, and Sine R Hadrup. T cell receptor fingerprinting enables in-depth characterization of the interactions governing recognition of peptide-mhc complexes. *Nature biotechnology*, November 2018. ISSN 1087-0156. doi: 10.1038/nbt.4303. URL https://europepmc.org/articles/PMC9452375.

Si-Yi Chen, Tao Yue, Qian Lei, and An-Yuan Guo. TCRdb: a comprehensive database for T-cell receptor sequences with powerful search function. *Nucleic Acids Research*, 49(D1):D468–D474, 09 2020. ISSN 0305-1048. doi: 10.1093/nar/gkaa796. URL https://doi.org/10.1093/nar/gkaa796.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation, 2014.

James B. Chung, Jennifer N. Brudno, Dominic Borie, and James N. Kochenderfer. Chimeric antigen receptor t cell therapy for autoimmune disease. *Nature Reviews Immunology*, 24(11):830–845, Nov 2024. ISSN 1474-1741. doi: 10.1038/s41577-024-01035-3. URL https://doi.org/10.1038/s41577-024-01035-3.

Asa Cooper Stickland, Xian Li, and Marjan Ghazvininejad. Recipes for adapting pre-trained monolingual and multilingual models to machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.eacl-main.301. URL http://dx.doi.org/10.18653/v1/2021.eacl-main.301.

Kristian Davidsen, Branden J Olson, III DeWitt, William S, Jean Feng, Elias Harkins, Philip Bradley, and IV Matsen, Frederick A. Deep generative models for t cell receptor protein sequences. *eLife*, 8:e46935, sep 2019. ISSN 2050-084X. doi: 10.7554/eLife.46935. URL https://doi.org/10.7554/eLife.46935.

Jennifer N. Dines, Thomas J. Manley, Emily Svejnoha, Heidi M. Simmons, Ruth Taniguchi, Mark Klinger, Lance Baldo, and Harlan Robins. The immunerace study: A prospective multicohort study of immune response action to covid-19 events with the immunecode™ open access database. *medRxiv*, 2020. doi: 10.1101/2020.08.17.20175158. URL https://www.medrxiv.org/content/early/2020/08/21/2020.08.17.20175158.1.

Liang Ding, Di Wu, and Dacheng Tao. Improving neural machine translation by bidirectional training. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3278–3284, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.263. URL https://aclanthology.org/2021.emnlp-main.263.

Edo Dotan, Gal Jaschek, Tal Pupko, and Yonatan Belinkov. Effect of tokenization on transformers for biological sequences. *bioRxiv*, 2023. doi: 10.1101/2023.08.15.553415. URL https://www.biorxiv.org/content/early/2023/08/17/2023.08.15.553415.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rihawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Debsindhu Bhowmik, and Burkhard Rost. Prottrans: Towards cracking the language of life's code through self-supervised deep learning and high performance computing. *bioRxiv*, 2020. doi: 10.1101/2020.07.12.199554. URL https://www.biorxiv.org/content/early/2020/07/12/2020.07.12.199554.

Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, and Samy Bengio. Why does unsupervised pre-training help deep learning? *J. Mach. Learn. Res.*, 11: 625–660, March 2010. ISSN 1532-4435.

Ethan Fast, Manjima Dhar, and Binbin Chen. Tapir: a t-cell receptor language model for predicting rare and novel targets. *bioRxiv*, 2023. doi: 10.1101/2023.09.12.557285. URL https://www.biorxiv.org/content/early/2023/09/15/2023.09.12.557285.

Yicheng Gao, Yuli Gao, Yuxiao Fan, Chengyu Zhu, Zhiting Wei, Chi Zhou, Guohui Chuai, Qinchang Chen, He Zhang, and Qi Liu. Pan-peptide meta learning for t-cell receptor–antigen binding recognition. *Nature Machine Intelligence*, 5:236–249, 2023.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. Survey of Low-Resource Machine Translation. *Computational Linguistics*, 48(3):673–732, 09 2022. ISSN 0891-2017. doi: 10.1162/coli_a_00446. URL https://doi.org/10.1162/coli_a_00446.

Charlotte Harrison. Tcr cell therapies vanquish solid tumors — finally. *Nature Biotechnology*, 42 (10):1477–1479, Oct 2024. ISSN 1546-1696. doi: 10.1038/s41587-024-02435-5. URL https://doi.org/10.1038/s41587-024-02435-5.

James Henderson, Yuta Nagano, Martina Milighetti, and Andreas Tiffeau-Mayer. Limits on inferring t-cell specificity from partial information, 2024. URL https://arxiv.org/abs/2404.12565.

Ilka Hoof, Bjoern Peters, John Sidney, Lasse Eggers Pedersen, Alessandro Sette, Ole Lund, Søren Buus, and Morten Nielsen. Netmhcpan, a method for mhc class i binding prediction beyond humans. *Immunogenetics*, 61:1–13, 2009.

Dan Hudson, Ricardo A Fernandes, Mark Basham, Graham Ogg, and Hashem Koohy. Can we predict t cell specificity with digital biology and machine learning? *Nature Reviews Immunology*, pp. 1–11, 2023.

Giulio Isacchini, Aleksandra M. Walczak, Thierry Mora, and Armita Nourmohammad. Deep generative selection models of t and b cell receptor repertoires with sonnia. *Proceedings of the National Academy of Sciences*, 118(14):e2023141118, 2021. doi: 10.1073/pnas.2023141118. URL https://www.pnas.org/doi/abs/10.1073/pnas.2023141118.

Yuepeng Jiang, Miaozhe Huo, Pingping Zhang, Yiping Zou, and Shuai Cheng Li. Tcr2vec: a deep representation learning framework of t-cell receptor sequence and function. *bioRxiv*, 2023. doi: 10.1101/2023.03.31.535142. URL https://www.biorxiv.org/content/early/2023/04/02/2023.03.31.535142.

Alexis M. Kalergis, Toshiro Ono, Fuming Wang, Teresa P. DiLorenzo, Shinichiro Honda, and Stanley G. Nathenson. Single Amino Acid Replacements in an Antigenic Peptide Are Sufficient to Alter the TCR Vb Repertoire of the Responding CD8+ Cytotoxic Lymphocyte Population1. *The Journal of Immunology*, 162(12):7263–7270, 06 1999. ISSN 0022-1767. doi: 10.4049/jimmunol.162.12.7263. URL https://doi.org/10.4049/jimmunol.162.12.7263.

Dhuvarakesh Karthikeyan, Colin Raffel, Benjamin Vincent, and Alex Rubinsteyn. Conditional generation of antigen specific t-cell receptor sequences. In *NeurIPS 2023 Generative AI and Biology (GenBio) Workshop*, 2023. URL https://openreview.net/forum?id=SckdgVW3Kq.

Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining, 2019.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better, 2022. URL https://arxiv.org/abs/2107.06499.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, 2019.

Yicheng Lin, Dandan Zhang, and Yun Liu. Tcr-gpt: Integrating autoregressive model and reinforcement learning for t-cell receptor repertoires generation, 2024. URL https://arxiv.org/abs/2408.01156.

Yating Liu, Xin Yan, Fan Zhang, Xiaoxia Zhang, Futian Tang, Zhijian Han, and Yumin Li. Tcr-t immunotherapy: The challenges and solutions. *Frontiers in Oncology*, 11, 2022. ISSN 2234-943X. doi: 10.3389/fonc.2021.794183. URL https://www.frontiersin.org/articles/10.3389/fonc.2021.794183.

Yuta Nagano and Benjamin Chain. tidytcells: standardizer for tr/mh nomenclature. *Frontiers in Immunology*, 14, 2023. ISSN 1664-3224. doi: 10.3389/fimmu.2023.1276106. URL https://www.frontiersin.org/journals/immunology/articles/10.3389/fimmu.2023.1276106.

Morten Nielsen, Anne Eugster, Mathias Fynbo Jensen, Manisha Goel, Andreas Tiffeau-Mayer, Aurelien Pelissier, Sebastiaan Valkiers, María Rodríguez Martínez, Barthélémy Meynard-Piganeeau, Victor Greiff, Thierry Mora, Aleksandra M. Walczak, Giancarlo Croce, Dana L. Moreno, David Gfeller, Pieter Meysman, and Justin Barton. Lessons learned from the immrep23 tcr-epitope prediction challenge. *ImmunoInformatics*, 16, Dec 2024. ISSN 2667-1190. doi: 10.1016/j.immuno.2024.100045. URL https://doi.org/10.1016/j.immuno.2024.100045.

Xing Niu, Michael Denkowski, and Marine Carpuat. Bi-directional neural machine translation with synthetic parallel data. In Alexandra Birch, Andrew Finch, Thang Luong, Graham Neubig, and Yusuke Oda (eds.), *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pp. 84–91, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-2710. URL https://aclanthology.org/W18-2710.

Timothy J. O'Donnell, Alex Rubinsteyn, and Uri Laserson. 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. doi: https://doi.org/10.1016/j.cels.2020.06.010. URL https://www.sciencedirect.com/science/article/pii/S2405471220302398.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pp. 311–318, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://doi.org/10.3115/1073083.1073135.

Valentin Quiniou, Pierre Barennes, Vanessa Mhanna, Paul Stys, Helene Vantomme, Zhicheng Zhou, Federica Martina, Nicolas Coatnoan, Michele Barbie, Hang-Phuong Pham, Béatrice Clémenceau, Henri Vie, Mikhail Shugay, Adrien Six, Barbara Brandao, Roberto Mallone, Encarnita Mariotti-Ferrandiz, and David Klatzmann. Human thymopoiesis produces polyspecific CD8+ α/β T cells responding to multiple viral antigens. *Elife*, 12, March 2023.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2020.

Anirban Sarkar, Ziqi Tang, Chris Zhao, and Peter K Koo. Designing dna with tunable regulatory activity using discrete diffusion. *bioRxiv*, 2024. doi: 10.1101/2024.05.23.595630. URL https://www.biorxiv.org/content/early/2024/05/24/2024.05.23.595630.

Zachary Sethna, Yuval Elhanati, Jr Callan, Curtis G, Aleksandra M Walczak, and Thierry Mora. OLGA: fast computation of generation probabilities of B- and T-cell receptor amino acid sequences and motifs. *Bioinformatics*, 35(17):2974–2981, 01 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz035. URL https://doi.org/10.1093/bioinformatics/btz035.

Andrew Sewell. Why must t cells be cross-reactive? *Nature reviews. Immunology*, 12:669–77, 08 2012. doi: 10.1038/nri3279.

Mikhail Shugay, Dmitriy V Bagaev, Ivan V Zvyagin, Renske M Vroomans, Jeremy Chase Crawford, Garry Dolton, Ekaterina A Komech, Anastasiya L Sycheva, Anna E Koneva, Evgeniy S Egorov, Alexey V Eliseev, Ewald Van Dyk, Pradyot Dash, Meriem Attaf, Cristina Rius, Kristin Ladell, James E McLaren, Katherine K Matthews, E Bridie Clemens, Daniel C Douek, Fabio Luciani,

Debbie van Baarle, Katherine Kedzierska, Can Kesmir, Paul G Thomas, David A Price, Andrew K Sewell, and Dmitriy M Chudakov. VDJdb: a curated database of T-cell receptor sequences with known antigen specificity. *Nucleic Acids Research*, 46(D1):D419–D427, 09 2017. ISSN 0305-1048. doi: 10.1093/nar/gkx760. URL https://doi.org/10.1093/nar/gkx760.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mass: Masked sequence to sequence pre-training for language generation, 2019.

Ido Springer, Nili Tickotsky, and Yoram Louzoun. Contribution of t cell receptor alpha and beta cdr3, mhc typing, v and j genes to peptide binding prediction. *Frontiers in Immunology*, 12, 2021. ISSN 1664-3224. doi: 10.3389/fimmu.2021.664514. URL https://www.frontiersin.org/articles/10.3389/fimmu.2021.664514.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks, 2014.

Nili Tickotsky, Tal Sagiv, Jaime Prilusky, Eric Shifrut, and Nir Friedman. McPAS-TCR: a manually curated catalogue of pathology-associated T cell receptor sequences. *Bioinformatics*, 33(18): 2924–2929, 05 2017. ISSN 1367-4803. doi: 10.1093/bioinformatics/btx286. URL https://doi.org/10.1093/bioinformatics/btx286.

Ifigeneia Tzannou, Anastasia Papadopoulou, Swati Naik, Kathryn Leung, Caridad A. Martinez, Carlos A. Ramos, George Carrum, Ghadir Sasa, Premal Lulla, Ayumi Watanabe, Manik Kuvalekar, Adrian P. Gee, Meng-Fen Wu, Hao Liu, Bambi J. Grilley, Robert A. Krance, Stephen Gottschalk, Malcolm K. Brenner, Cliona M. Rooney, Helen E. Heslop, Ann M. Leen, and Bilal Omer. Off-the-shelf virus-specific t cells to treat bk virus, human herpesvirus 6, cytomegalovirus, epstein-barr virus, and adenovirus infections after allogeneic hematopoietic stem-cell transplantation. *Journal of Clinical Oncology*, 35(31):3547–3557, 2017. doi: 10.1200/JCO.2017.73.0655. URL https://doi.org/10.1200/JCO.2017.73.0655. PMID: 28783452.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

Randi Vita, Swapnil Mahajan, James A Overton, Sandeep Kumar Dhanda, Sheridan Martini, Jason R Cantrell, Daniel K Wheeler, Alessandro Sette, and Bjoern Peters. The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Research*, 47(D1):D339–D343, 10 2018. ISSN 0305-1048. doi: 10.1093/nar/gky1006. URL https://doi.org/10.1093/nar/gky1006.

Ekeruche-Makinde Wooldridge, Skowera van den Berg, Tan Miles, Clement Dolton, Price Llewellyn-Lacey, and Sewell Peakman. A single autoimmune t cell receptor recognizes more than a million different peptides. *Journal of Biological Chemistry*, 287(92):1168–1177, 2011. doi: 10.1074/jbc.M111.289488. URL https://www.sciencedirect.com/science/article/pii/S0167569998012997.

Kevin Wu, Kathryn E. Yost, Bence Daniel, Julia A. Belk, Yu Xia, Takeshi Egawa, Ansuman Satpathy, Howard Y. Chang, and James Zou. Tcr-bert: learning the grammar of t-cell receptors for flexible antigen-xbinding analyses. *bioRxiv*, 2021. doi: 10.1101/2021.11.18.469186. URL https://www.biorxiv.org/content/early/2021/11/20/2021.11.18.469186.

Kai W Wucherpfennig, Paul M Allen, Franco Celada, Irun R Cohen, Rob De Boer, K Christopher Garcia, Byron Goldstein, Ralph Greenspan, David Hafler, Philip Hodgkin, Erik S Huseby, David C Krakauer, David Nemazee, Alan S Perelson, Clemencia Pinilla, Roland K Strong, and Eli E Sercarz. Polyspecificity of T cell and B cell receptor recognition. *Semin. Immunol.*, 19(4): 216–224, August 2007.

Hsiu-Wei Yang, Yanyan Zou, Peng Shi, Wei Lu, Jimmy J. Lin, and Xu Sun. Aligning cross-lingual entities with multi-aspect information. In *Conference on Empirical Methods in Natural Language Processing*, 2019. URL https://api.semanticscholar.org/CorpusID:202121966.

Jiannan Yang, Bing He, Yu Zhao, Feng Jiang, Zhonghuang Wang, Yixin Guo, Zhimeng Xu, Bo Yuan, Jiangning Song, Qingpeng Zhang, and Jianhua Yao. De novo generation of t-cell receptors with desired epitope-binding property by leveraging a pre-trained large language model. 10 2023. doi: 10.1101/2023.10.18.562845.

Kun Yu, Ji Shi, Dan Lu, and Qiong Yang. Comparative analysis of CDR3 regions in paired human $\alpha\beta$ CD8 T cells. *FEBS Open Bio*, 9(8):1450–1459, August 2019.

Pengfei Zhang, Seojin Bang, Michael Cai, and Heewook Lee. Context-aware amino acid embedding advances analysis of tcr-epitope interactions. *bioRxiv*, 2024a. doi: 10.1101/2023.04.12.536635. URL https://www.biorxiv.org/content/early/2024/02/28/2023.04.12.536635.

Yumeng Zhang, Zhikang Wang, Yunzhe Jiang, Dene R. Littler, Mark Gerstein, Anthony W. Purcell, Jamie Rossjohn, Hong-Yu Ou, and Jiangning Song. Epitope-anchored contrastive transfer learning for paired cd8+ t cell receptor–antigen recognition. *Nature Machine Intelligence*, Oct 2024b. ISSN 2522-5839. doi: 10.1038/s42256-024-00913-8. URL https://doi.org/10.1038/s42256-024-00913-8.

Zhenghong Zhou, Junwei Chen, Shenggeng Lin, Liang Hong, Dong-Qing Wei, and Yi Xiong. Gratcr: epitope-specific t cell receptor sequence generation with data-efficient pre-trained models. *bioRxiv*, 2024. doi: 10.1101/2024.07.21.604503. URL https://www.biorxiv.org/content/early/2024/07/23/2024.07.21.604503.

A REFLECTION

In this study we set out to investigate the potential of seq2seq models to design antigen-specific TCR sequences for unseen epitopes and characterize these generations beyond simple metrics. Building on our previous framing of TCR design as a sequence-to-sequence task, we explored low-resource machine translation techniques to address the scarcity of labeled TCR-epitope data. To this end, we introduced joint pre-training, bidirectional training, and semi-synthetic data augmentation, and evaluated these approaches on a target-rich dataset of well-studied pMHCs. Remarkably, the models demonstrated an ability to generalize outside a skewed training distribution, achieving exact sequence matches to reference CDR3β sequences of viral epitopes and neoantigens. This learned input sensitivity, while not perfect, suggests robustness to data imbalance, which can help inform *in-vitro* data generation strategies on depth vs. breadth-first discovery of antigen-specific TCRs for rare pMHCs.

After confirming the utility of target-conditioning, we leveraged low resource machine translation methods including pre-training and bidirectional training to learn the intra- and inter- sequence co-dependencies of TCR and pMHC. Interestingly, we found that pre-training had opposite effects for TCRBART, which used BERT(Vaswani et al., 2023)-style token masking, and TCRT5, pre-trained on T5-style span masking. Since both used a masking rate of 15%, we suspect that the higher order k-mers learned by span masking may be better suited for CDR3β sequences, though more work is necessary to confirm this. For both TCRBART and TCRT5, the bidirectional and multi-task models achieved higher sequence recovery and median F1 scores across validation pMHCs, driven by a bias towards sampling polyspecific TCRs. Unsurprisingly, the performance of all models correlated with the density of reference data for each pMHC. Notably, we found a significant increase in F1 sensitivity above a 90% sequence identity threshold, suggesting this metric may be a more meaningful indicator than mean sequence recovery for assessing antigen-specific repertoire quality.

Putting it all together, we selected TCRT5 as the most well-rounded model for its high accuracy, diversity, and attenuated reliance on polyspecific generations. We rigorously benchmarked the model, highlighting a reduction in repertoire diversity driven by preferential sampling of sequences with high V(D)J generation probabilities via beam search. Still, we show TCRT5's utility by generating validated antigen-specific CDR3β sequences not encountered during training. We further demonstrate its greater than random performance on the IMMREP2023 unseen antigens, possessing substantially fewer known cognate TCRs, providing a pathway for generating functionally relevant TCR repertoires for sparsely sampled epitopes likely to be encountered in a real therapeutic scenario. Additionally we benchmark ER-BERT and GRATCR, and show performance comparable to TCRT5 on individual metrics while TCRT5 scores well across all metrics, especially sequence recovery above 90%. Together, these results highlight the exciting potential of seq2seq models while underscoring the importance of carefully considering metrics in prioritizing models and predictions.

The current iteration of our study has many limitations, stemming from both our approach and an innate scarcity of available data. First and foremost, is our focus on the CDR3β loop of the TCR, even though the α chain and V and J genes have been shown to play an important role in determining specificity (Springer et al., 2021; Henderson et al., 2024). In its current state, our model would require template TCRs for which the CDR3β can be designed. Second, given the sparse nature of data, we risk high variance across recall-based metrics. To mitigate this, we leveraged on our target-rich dataset for model/checkpoint selection and to thoroughly characterize the models' behavior. We understand and accept that this introduces leakage through model selection. However, we argue for its necessity given the severe data sparsity to evaluate pMHCs stemming from multiple disease contexts. Importantly, TCRT5 demonstrates consistent, monotonic improvement in performance across training checkpoints, suggesting that our final model demonstrates real learning rather than random fluctuations to its parameters. Additionally, we test TCRT5 on the IMMREP2023 epitopes that were not used for training or in validation with a

B METHODS

B.1 SEQUENCE REPRESENTATION

Following our prior work (Karthikeyan et al., 2023), we adopted the same sequence-to-sequence (seq2seq) framework, relaxing the direction of pMHC $\rightarrow$ TCR source-target pairs to train on both directions, but evaluate on the former. To represent the TCR:pMHC trimeric complex, comprised of three sub-interactions (TCR-peptide, TCR-MHC, peptide-MHC) as a source-target sequence pair, we made a few simplifying assumptions that allowed for a more straightforward problem formulation. First, we assume a stable pMHC complex, reducing the problem to a dimeric interaction between TCR and pMHC. Second, we focus on the amino acid residues at the binding interface. For the TCR, we use the CDR3β loop, a contiguous span of 8-20 amino acids that typically make the most contact with the peptide (Yu et al., 2019). Similarly, for the pMHC, we use the whole peptide and the MHC pseudo-sequence, defined in (Hoof et al., 2009) as a reduced, noncontiguous, string containing the polymorphic amino acids within 4.0 Å of the peptide. We opt for a single character amino-acid level tokenization, primarily for its interpretibility (Dotan et al., 2023). In addition to the 20 canonical amino acids, we use standard special tokens including the start/end of sequence, masking, padding, and a separator token to delineate the boundary between the concatenated peptide and pseudo-sequence. For TCRT5, we additionally employ the use of sequence type tokens, retained from T5's use of task prefixes (Raffel et al., 2020), to designate translation direction:

TCRBART:

[SOS]EPITOPE[SEP]PSEUDOSEQUENCE[EOS] $\leftrightarrow$ [SOS]CDR3BSEQ[EOS]

TCRT5:

[PMHC]EPITOPE[SEP]PSEUDOSEQUENCE[EOS] $\leftrightarrow$ [TCR]CDR3BSEQ[EOS]

B.2 DATASET CONSTRUCTION

B.2.1 *Parallel Corpus*

Our parallel corpus comprised experimentally validated immunogenic TCR:pMHC pairs taken from publicly available databases (McPAS (Tickotsky et al., 2017), VDJdb (Shugay et al., 2017), and IEDB (Vita et al., 2018)). Additionally, we used a large sample of partially-labeled data derived from the MIRA (Dines et al., 2020) dataset, which contained CDR3β and peptide sequences, but contained MHC information at the haplotype resolution instead of the actual presenting MHC allele. Therefore, the presenting MHC allele was inferred from the individual's haplotype using MHCFlurry2.0's (O'Donnell et al., 2020) top-ranked presentation score. Of importance, these semi-synthetic examples were not used in evaluation. To aggregate the data spanning various sources, formats, and nomenclature, we mapped the columns from each individual dataset to a common consensus schema and concatenated the data along the consensus columns. Missing values were reasonably imputed according to other information for that data instance. To keep only the cytotoxic (CD8+) T cells, we filtered the instances where the cell-type was provided or where the HLA-Allele was of MHC-class I. Where the granularity of the HLA-information or TR genes was at the serotype level, we inferred the canonical gene/allele by starting off with the subgroup '*01' and incremented it until a matching IMGT gene was found. This step has the potential of introducing minor differences between the unknown ground truth and the imputed pseudo-sequence, as it is well conserved within serotype. Once the data was aggregated and values were imputed, we applied the following column-level standardization for each source of information:

- **Complementarity Determining Region (CDR3β), Epitope, and MHC Pseudo-sequence**: All amino acid representations were normalized used the `'tidytcells.aa.standardise'` function found in the TidyTcells python package (Nagano & Chain, 2023).
- **TR Genes**: The TidyTcells package (Nagano & Chain, 2023) was once again used to standardize the nomenclature surrounding the T-Cell Receptor genes (e.g. TRB-V and TRB-J).

- **HLA-Allele**: HLA alleles were imputed where allele level information when necessary and then normalized using the MHCgnomes package to the standard HLA-[A,B,C]*XX:YY format.

B.2.2 *Training/Validation Split*

To accurately assess the capacity of the models to sample antigen-specific sequences on unseen epitopes, we held out a validation set of the top-20 most target-rich pMHCs, which collectively account for over 80% of the labeled TCR sequences. We trained on the remaining data, further removing the occurrences of the held-out epitopes bound alternate MHCs to ensure a clean validation split (Figure 1c). We retained training sequences with a low edit distance to the validation pMHCs to better understand their influence on performance. The degree to which these sequences exhibit training set similarity is reflected in (Table S1). The parallel corpus was subsequently deduplicated to remove near duplicates (peptides with the same allele and a ¿= 6-mer overlap) which we found to marginally help overall performance, in accordance with (Lee et al., 2022). This resulted in a final dataset split of ≈330k training sequence pairs (N=6989 pMHCs) and 68k validation sequence pairs (N=20 pMHCs). A key limitation of this dataset is its highly skewed bias towards mainly viral epitopes and a very narrow HLA distribution towards well studied alleles (Figure 1d).

B.2.3 *Unlabeled 'Monolingual' Data*

We hypothesized that pre-training the encoder:decoder model using self-supervised methods on pMHC and TCR sequences could help boost the translation performance of the model by learning better representations for source and target sequences as in (Cooper Stickland et al., 2021), which crucially has been shown to improve performance in the low-resource setting (Haddow et al., 2022). For the unlabeled pMHC sequences, we used the positive MHC ligand binding assay data from IEDB (N≈740K) (Vita et al., 2018). For the TCR sequences, we used around (N≈14M) sequences from TCRdb (Chen et al., 2020) of which around 7M CDR3β sequences were unique. For this dataset we chose to retain duplicate CDR3β sequences as the TCRdb was amassed over multiple studies and populations, so we felt that the inclusion of duplicate CDR3βs was reflective of convergent evolution in the true unconditional TCR distribution.

B.3 Model Training

B.3.1 *Pre-Training*

TCRBART was pre-trained using masked amino acid modeling (BERT-style (Elnaggar et al., 2020)), while TCRT5 utilized masked span reconstruction, learning to fill in randomly dropped spans with lengths between 1 and 3. Of importance, neither model was trained on complete sequence reconstruction to reduce the possibility of memorization during pre-training. Both models were trained on unlabeled CDR3β and peptide::pseudo-sequences, simultaneously pre-training the encoder and decoder, inspired by the MASS/XLM approach (Song et al., 2019; Lample & Conneau, 2019). Unlike MASS/XLM, we omitted learned language embeddings, allowing the model to learn from the size differences between CDR3β and pMHC sequences. To address the imbalance in sequence types, we upsampled sequences for a 70/30 TCR to pMHC split.

B.3.2 Direct Training/Finetuning

For the parallel data, we used the same three training regimes (baseline, bidirectional, multi-task) for direct training from random initialization as well as finetuning from a pretrained model. This was done by extending the standard categorical cross entropy loss function (Equation 1), favored in seq2seq tasks for its desired effect of maximizing the conditional likelihoods over target sequences (Sutskever et al., 2014; Cho et al., 2014). For the baseline training, we used the canonical form of the cross entropy loss, as shown below:

$$\begin{aligned}\mathcal{L} = CE(\mathbf{y}, \hat{\mathbf{y}}) &= -\sum_{i=1}^{n} \mathbf{y}_i \log \hat{\mathbf{y}}_i \\ &= -\sum_{i=1}^{n}\sum_{j-1}^{k} y_{ij} \log p_\theta(y_{ij}|\mathbf{x})\end{aligned} \tag{1}$$

The bidirectional and multi-task models were trained using mutli-term objectives, forming a linear combination of individual loss terms corresponding to the cross entropy loss of each task/direction.

$$\mathcal{L}_{bidxn} = \mathcal{L}_{pmhc \to tcr} + \mathcal{L}_{tcr \to pmhc} \tag{2}$$

$$\mathcal{L}_{multi} = \mathcal{L}_{mlm} + \mathcal{L}_{pmhc \to tcr} + \mathcal{L}_{tcr \to pmhc} \tag{3}$$

In order to mitigate effects of model forgetting with stacking single-task training epochs, we shuffled the tasks across the epoch using a simple batch processing algorithm (Algorithm 1). After the batch was sampled, it was rearranged into one of four sequence-to-sequence mapping possibilities and trained on target reconstruction with the standard cross entropy loss, which was used for back-propagation. In this way, we could ensure that the model was simultaneously learning multiple tasks during training. For the bidirectional model, this was straightforward as we could swap the input and output tensors during training to get the individual loss contributions of the $\mathcal{L}_{pmhc \to tcr}and\mathcal{L}_{tcr \to pmhc}$ (Equation 2). For the multi-task model, the mapping possibilities are: 1) pMHC $\to$ TCR 2) TCR $\to$ pMHC 3) Corrupted pMHC* $\to$ pMHC 4) Corrupted TCR* $\to$ TCR, which combine to to form $\mathcal{L}_{multi}$ (Equation 3). These tasks and sequence mappings as seen by TCRBART and TCRT5 are summarized in Figure 2b.

Algorithm 1 Multi-Task Training Step

Batched Input: source pMHCs: **X**, target TCRs: **Y**
Sample $a \sim Bernoulli(0.5)$
if $a > 0.5$ **then**
  Swap **X** and **Y**
  Compute attention masks
end if
Sample $b \sim Bernoulli(0.5)$
if $b > 0.5$ **then**
  Set **X = X*** and **Y = X**
  Compute attention masks
end if
do Predict $\hat{\mathbf{Y}} = \phi(\mathbf{X})$ and gradient updates on CE(**y**, $\hat{\mathbf{y}}$)

B.4 Evaluation

To evaluate antigen-specificity, we build our framework around sampling exact CDR3β sequences from published experimental data on well-characterized validation epitopes not seen during training. This approach has an interpretable bias compared to black-box error profiles, at the cost of potentially under-representing actual performance. We calculate sequence similarity-based metrics beyond exact overlap to create a more robust evaluation framework, and characterize their concordances for future use on epitopes with fewer known cognate sequences. Broadly, our metrics can be summarized as evaluating the accuracy of the returned sequences, their diversity, or some combination of the two. They are summarized in brief below:

Accuracy Metrics

- **Char-BLEU**: Following BLEU-4 (Papineni et al., 2002), the character-level BLEU calculates the weighted n-gram precision against the $k = 20$ closest reference sequences to

abate unintended penalization of accurate predictions under a large reference set. We use the NLTK's `'sentence_bleu'` function to calculate a single translation's BLEU score and the `'corpus_bleu'` function to compute the BLEU score over an entire dataset.

- **Native Sequence Recovery**: We compute the index-matched sequence overlap with the closest known binder of the same sequence length, when available. This is the same as the length-normalized Hamming distance. The Levenshtein distance normalized to the length of closest reference was used for cases where a size-matched reference did not exist.
- **Mean Average Precision (mAP)**: Borrowed from information retrieval, mean average precision measures the average precision across the ranked model predictions. Here, we rank the generations by model log-likelihood scores and take the average of the precisions at the top-1, top-2, top-3, ... top-k ranked outputs. Then we take the mean over the various pMHCs' average precision (AP) values to get the mean average precision. This metric gauges both the accuracy of the model as well as the calibration of its sequence likelihoods.
- **Biological Likelihood**: To assess the plausibility of model outputs independent of antigen-specificity or labeled data, we compute generation probability of predictions using OLGA, a domain specific generative model that infers CDR3β sequence likelihood (Sethna et al., 2019).

Diversity Metrics

- **Total Unique Sequences**: As a measure of global diversity, we compute the number of total unique generations across the top-20 validation pMHCs as a diversity metric that captures model degeneracy and input specificity.
- **Jaccard Similarity/Dissimilarity Index**: The Jaccard Index or the Jaccard similarity score is used to measure the similarity of two sets and is calculated as the size of the intersection divided by the union of the two sets. Since the Jaccard Index is inversely proportional to diversity, one minus the Jaccard Index is often to use to represent diversity between two sets.
- **Positional ΔEntropy**: In order to quantify the change in diversity between the models' outputs and the reference distribution per CDR3β position, we report $H(q_i) - H(p_i)$ over the KL divergence to get a signed change in entropy between the amino acid usage distribution of reference distribution q and sample distribution p at position i.

Both

- **Precision, Recall, and F1@K**: Also taken from information retrieval, these metrics gauge precision, recall, and F1 by sampling $K = 100$ times, without rank, and measuring the exact sequence overlap with the reference sequences. In the case of beam search decoding, since we observed beam search to return unique sequences at our choice of decoding parameters, all of these metrics were equivalent and are simply represented by the F1 score.
- **K-mer Spectrum Shift**: As used in the DNA sequence design space (Sarkar et al., 2024), the k-mer spectrum shift measures the Jensen-Shannon (JS) divergence between the k-mer usage frequency distributions of two sets of sequences across different values of k. Here we compare the JS divergence between the distribution of k-mers derived from a pMHC's model generations and its reference set of sequences.

B.5 Hyperparameter Optimization

We first investigated the impact of course grained choices in the larger model such as the width and depth by sweeping over model architecture and training algorithm values. For the model architecture we varied the number of attention heads, batch size, d_{model}, feed forward layer dimension, and number of total layers. All models were trained using the cross entropy loss with the AdamW optimizer. For the optimizer, we varied the learning rate and weight decay parameters. To compare model parameters from both the zero-pre-training and pre-training+finetuning regimes, we ran a sweep on samples of the 'monolingual' (unlabeled TCR and pMHCs) and 'parallel' (paired TCR:pMHC) corpuses to tune performance on the pre-training and seq2seq task, respectively. Due to time and compute constraints, the sweeps were performed on a reduced sample of of 100k TCRs and 100k

pMHCs from the monolingual texts for pre-training. Interestingly, we found that while the optimal configurations for the BART models were relatively consistent, the optimal T5 configurations for pre-training was a deeper, more narrow network and the one for direct training was a wider more shallow network. To reconcile these we adopted the following heuristic: if the values were close, the higher parameter count was chosen and an intermediate value was chosen when the parameter values were far. Slight adjustments were performed at the layer count level to adjust for parameters and make TCRBART and TCRT5 comparable. The final TCRBART architecture uses 6 encoder and decoder layers at $d_{model} = 768$, totaling around 46 million parameters. The final TCRT5 implementation used $d_{model} = 256$ and 10 encoder and decoder layers for a total of 42M parameters.

B.6 Checkpoint Selection

In deciding to choose which checkpoints to use for each model, we observed a marked difference between the performance dynamics of the models with and without pre-training. This is most clearly observed when plotting the diversity and accuracy metrics for each of the model checkpoints, showing distinct training trajectories in the utility space (Figure S2a-b). The pre-trained models demonstrate asymptotically increasing model performance across checkpoints for both the F1@100 and native sequence recovery metrics while the models that were trained directly from random initialization showed signs of potential overfitting, as both metrics peaked early on during training and dropped over additional iterations (Figure S2c-d). This was observed on reduced learning rates as well, indicating a possible regularization effect from pre-training(Erhan et al., 2010). A distinct difference between TCRBART and TCRT5 variants was the effect of including pre-training. While TCRT5 showed a significant improvement given pre-training, TCRBART showed worse performance. However, the finetuning's performance dynamics proved to be more stable than the non-pre-trained version, complicating the benefit of adding pre-training to TCRBART. Additionally we examined the number of unique sequences for the models over the checkpoints and saw that the TCRBART-0 and TCRT5-FT showed increasing number of unique sequences over training steps (Figure S2e). For each of the models, the checkpoint with the best performance on F1 was chosen as representatives of the best performances for each training paradigm.

B.7 Benchmarking

B.7.1 GRATCR

For running GRATCR on the IMMREP epitopes, we followed the instructions provided by the GRATCR team (Zhou et al., 2024) at their hosted GitHub `https://github.com/zhzhou23/GRATCR`. We ran the beam search decoding as provided. The script to sample the finetuned GRATCR:

```
python GRA.py --data_path="./data/IMMREP_peptides.csv"
--tcr_vocab_path="./Data/vocab/total-beta.csv"
--pep_vocab_path="./Data/vocab/total-epitope.csv"
--model_path="./model/gra.pth" --bert_path="./model/bert_pretrain.pth"
--gpt_path="./model/gpt_pretrain.pth" --mode="generate"
--result_path="./results.csv" --batch_size=1 --beam=1000
```

B.7.2 ER-BERT

ER-BERT was run using the UAA (unique amino acid) model for a more direct comparison to TCRT5. We utilize the `seq_generate` method as described in their codebase with the default parameters as shown in `https://github.com/TencentAILabHealthcare/ER-BERT/` under `Code/evaluate_seq2seq_MIRA.py` as used by the ER-BERT team (Yang et al., 2023).

B.8 Statistics

For evaluating TCRT5 on the IMMREP epitopes, we constructed two independent nulls to evaluate the comparative performance of unconditional generation as well as test the conformity to an epitope-specific repertoire. The first null distribution was constructed by bootstrapping our test statistic using soNNia generated sequences, counting the number of instances where the soNNia

CDR3β sequences (out of 1000 sequences per simulation) achieved at least a 90% sequence recovery. The second null was constructed by calculating the sequence recovery rate of TCRT5 generations against a reference size-matched sample of random CDR3β sequences generated by soNNia to evaluate the specificity of TCRT5 to a particular epitope-specific repertoire over a random repertoire. We calculated empirical p-values based on the fraction of simulations where the null distributions matched or exceeded the observed TCRT5 performance with the ground truth sequences.

C FIGURES

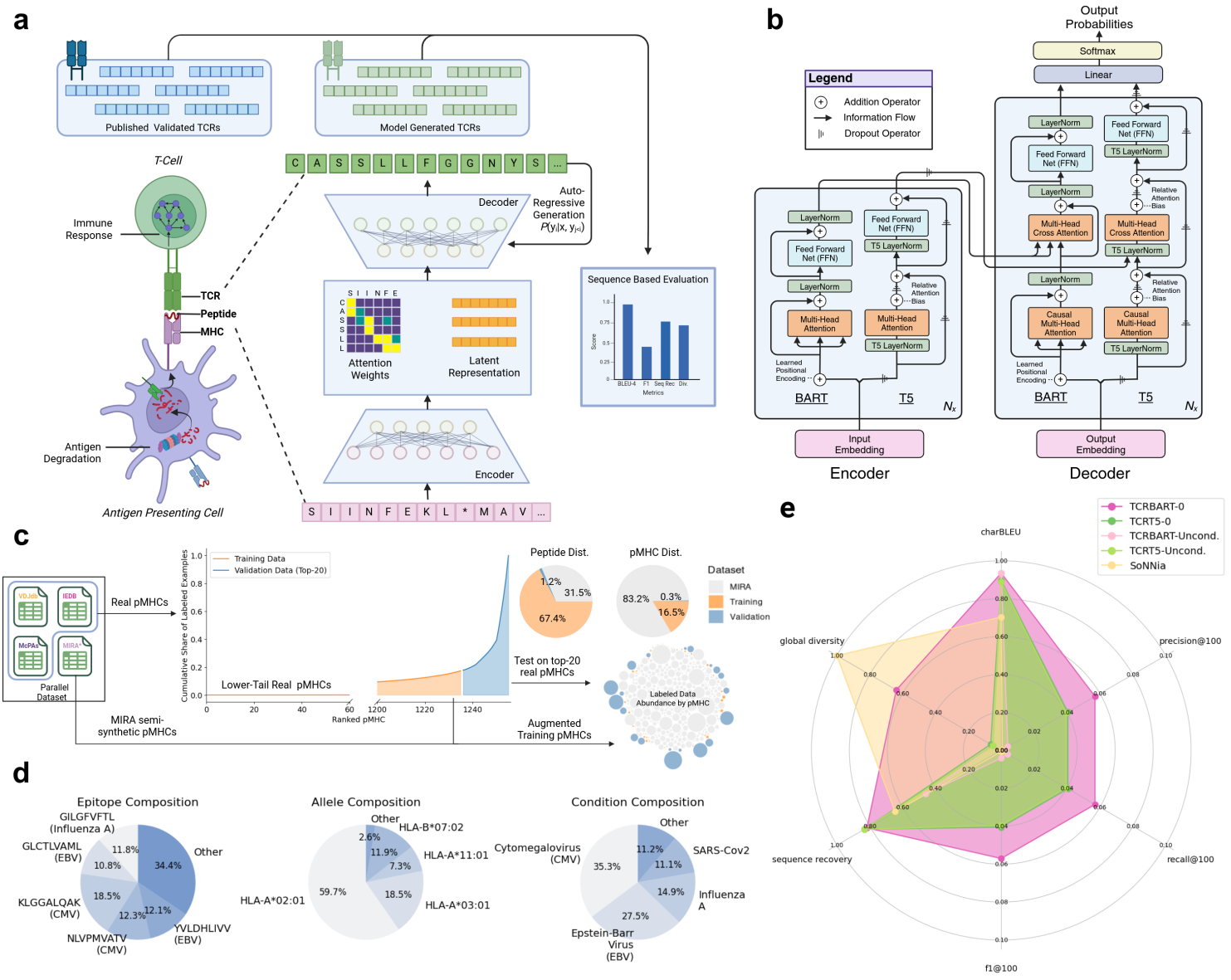


Figure 1: Overview of TCR-TRANSLATE. (a) Casting antigen-specific TCR design as a sequence-to-sequence (seq2seq) task. We make use of an encoder:decoder abstraction to process peptide:MHC sequence information and autoregressively sample target-conditioned CDR3β sequences. (b) Specific architecture of TCRBART and TCRT5. Transformer architecture juxtaposing BART and T5 encoder and decoder layers highlighting key operations to the residual stream, inspired by (Vaswani et al., 2017). (c) Dataset creation. Given severe data-sparsity, the top-20 pMHCs from IEDB, VDJdb, and McPAS in terms of known TCRs was witheld as validation, while the remainder was used for training with semi-synthetic pMHCs from MIRA. (d) Composition of validation set. Breakdown of epitopes, alleles, and disease contexts of the top-20 real pMHCs. (e) Conditional generation outperforms unconditional generation methods. Radar plot showing the performance of TCRBART and TCRT5 models trained without pre-training (TCRBART-0, TCRT5-0) evaluated against their unconditional generations (TCRBART-Unconditional, TCRT5-Unconditional) as well as the the averaged metrics over 1000 simulations of the statistical soNNia generative model.

Figure 2: Multi-task training increases accuracy and decreases diversity metrics. (a) Diagram outlining pre-training, seq2seq training/finetuning, and their common generation scheme (inference). (b) Sequence I/O representation of TCRBART and TCRT5 broken down by task. (c) TCR-TRANSLATE accuracy metrics. Swarm plots showing the median, quartile, and individual contributions of each of the validation pMHCs for Char-BLEU, F1@100, and native sequence recovery. (d) Fraction of pMHC F1@100 scores that remain equivalent to or greater than the baseline models (TCRBART-0, TCRT5-0). (e) Model calibration as measured by mean average precision (mAP) across pMHCs calculated using sequence likelihood based rank per model. (f) Barplot of global diversity calculated as the total number unique sequences across pMHCs (20 x 100=2000 max). (g) Scatterplot summarizing model performance on accuracy and diversity metrics. Accuracy is taken as the mean F1@100 score and diversity is shown both in terms of the total number of unique sequences generated (size of each data point) as well as the mean pairwise Jaccard dissimilarity scores across pMHCs.

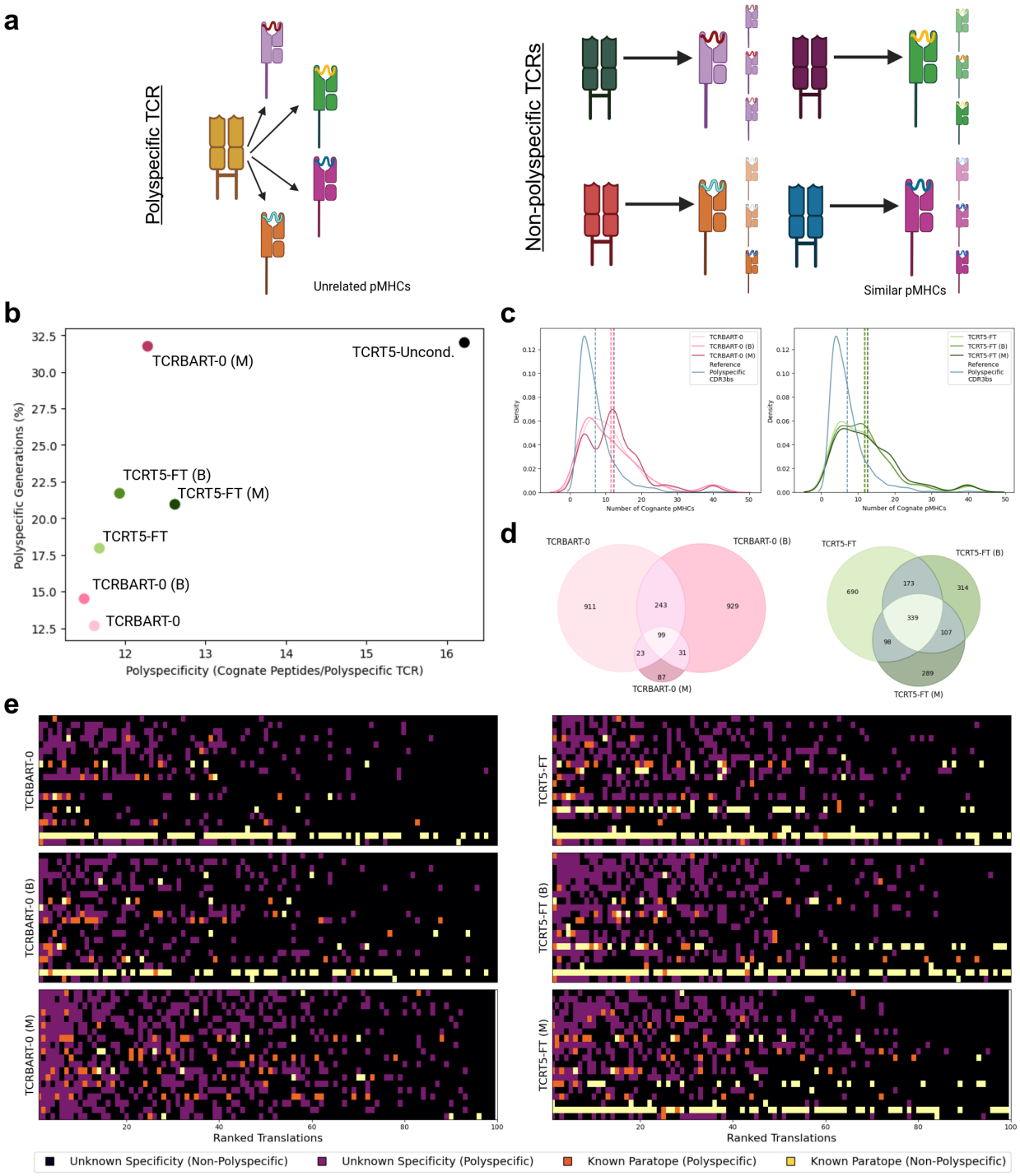


Figure 3: Multi-task training promotes degenerate sampling of polyspecific TCRs. (a) Diagram of showing polyspecific TCRs binding different, unrelated pMHCs juxtaposed against regular TCRs sharing a more conserved cross reactivity profile. (b) Scatterplot of the number of polyspecific generations as a percentage as well as the mean polyspecificity (number of distinct peptides) of the polyspecific TCRs per model is shown. (c) Distribution of TCR polyspecificity across the parallel data and model generations. Density plot of cognate peptide counts for polyspecific TCRs aggregated from the combined training and validation set (reference CDR3βs) and the model variants per class. (d) Venn diagrams of translation overlaps for TCRBART-0 and TCRT5-FT model variants. (e) TCRBART and TCRT5 sample polyspecific and known binders with higher sequence likelihoods than those of unknown specificities. Discrete heatmaps where rows indicate individual pMHCs, columns indicate ranked translation, and color indicates known binding and polyspecificity status are shown for TCRBART-0 and TCRT5-FT variants.

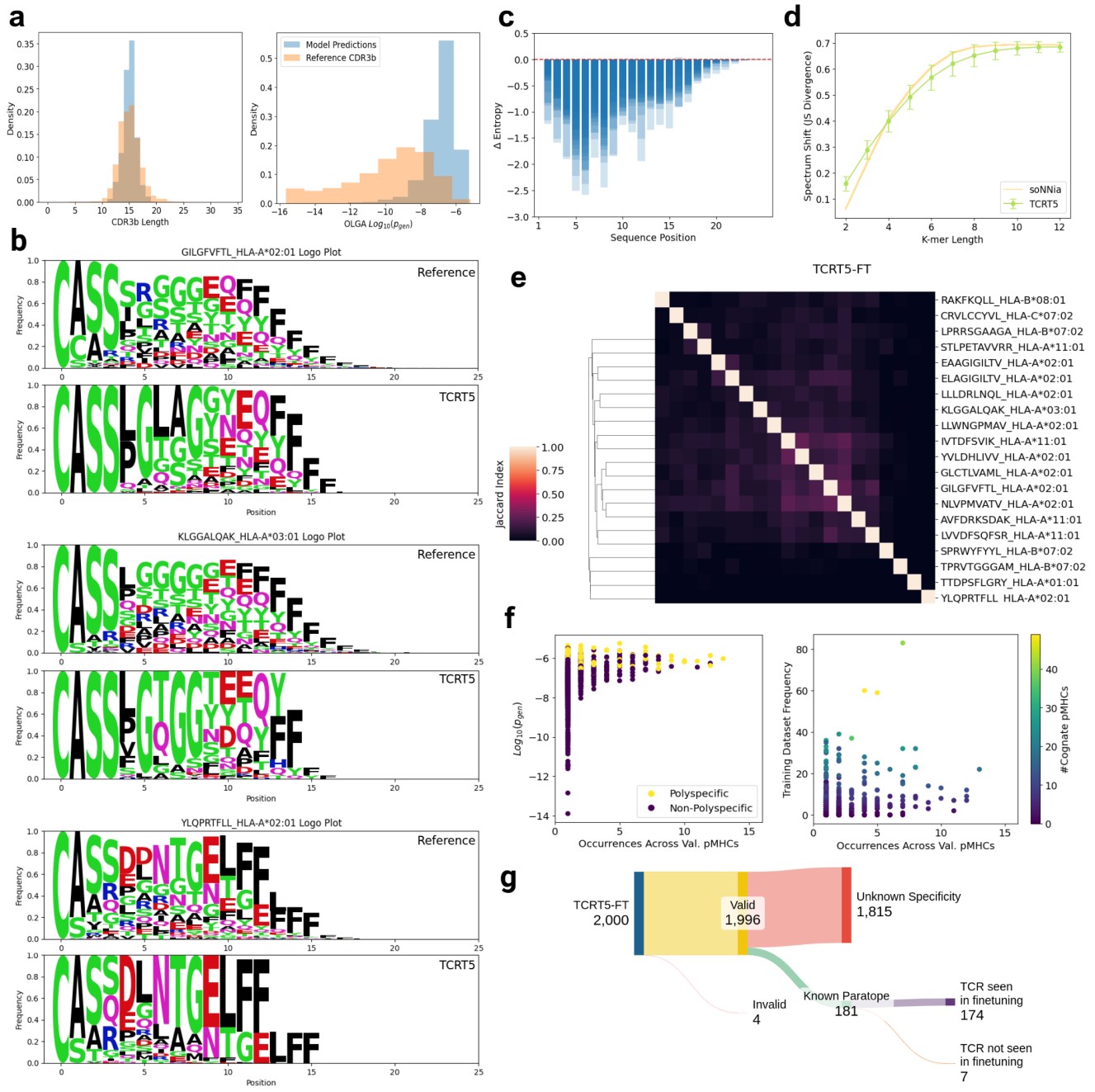


Figure 4: Qualitative assessment of TCRT5. (a) Repertoire-level features of reference and generated CDR3βs. TCRT5 captures the tails of the CDR3β length distribution but preferentially samples sequences at the right tail of OLGA generation probabilities. (b) Sequence logo plots showing the decrease in sequence diversity position across the generated and reference CDR3β sequences for three canonical pMHCs [GILGFVFTL (Influenza-A), KLGGALQAK (EBV), YLQPRTFLL (SARS-CoV2)]. (c) Generated sequences experience a decrease in Shannon entropy for nearly all positions compared to reference sequences across all pMHCs. Barplots for individual pMHCs are overlaid on one another. (d) K-mer spectrum shift plot showing the Jensen Shannon divergence between generated and reference sequences. Mean JS divergence for soNNia generations for 100 sequences sampled per pMHC across 100 simulations are shown for reference. Error bars mark the mean and 1-standard deviation across validation pMHCs. (e) Heatmap of Jaccard Index scores showing the generated sequence co-occurrence across different pMHC pairs. (f) TCRT5 repeats sequences across pMHCs in line with biological probabilities and is robust to training set abundance. Scatterplot visualizing the occurrence across pMHCs with the OLGA p_{gen}, polyspecificity, and training set frequency. (g) TCRT5 generates experimentally validated antigen-specific CDR3β sequences unseen during training. Sankey diagram showing the validity (non-zero OLGA p_{gen}), known antigen-specificity status, and training set membership of generated sequences across the validation pMHCs are shown.

a

	Origin	Protein/AA	# TCRs
FTDALGIDEY HLA-A*01:01	Human Herpes Virus - 1	UL48/479-488	12
SALPTNADLY HLA-A*01:01	Human Herpes Virus - 1	UL48/90-99	18
TDLGQNLLY HLA-A*01:01	Human Adenovirus C	Hexon/895-903	33
TSDACMMTMY HLA-A*01:01	Cytomegalovirus (HCMV)	IE1/290-299	11

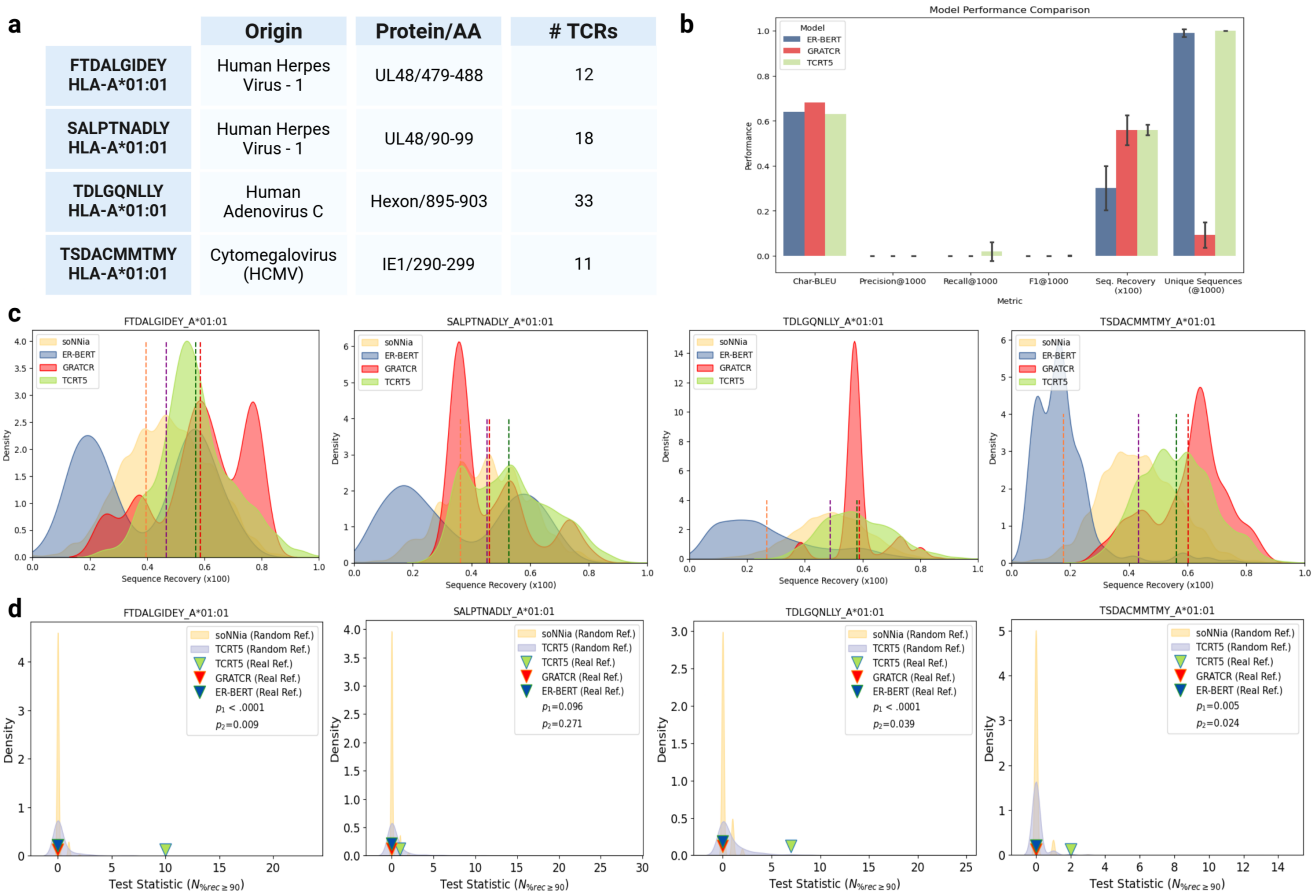
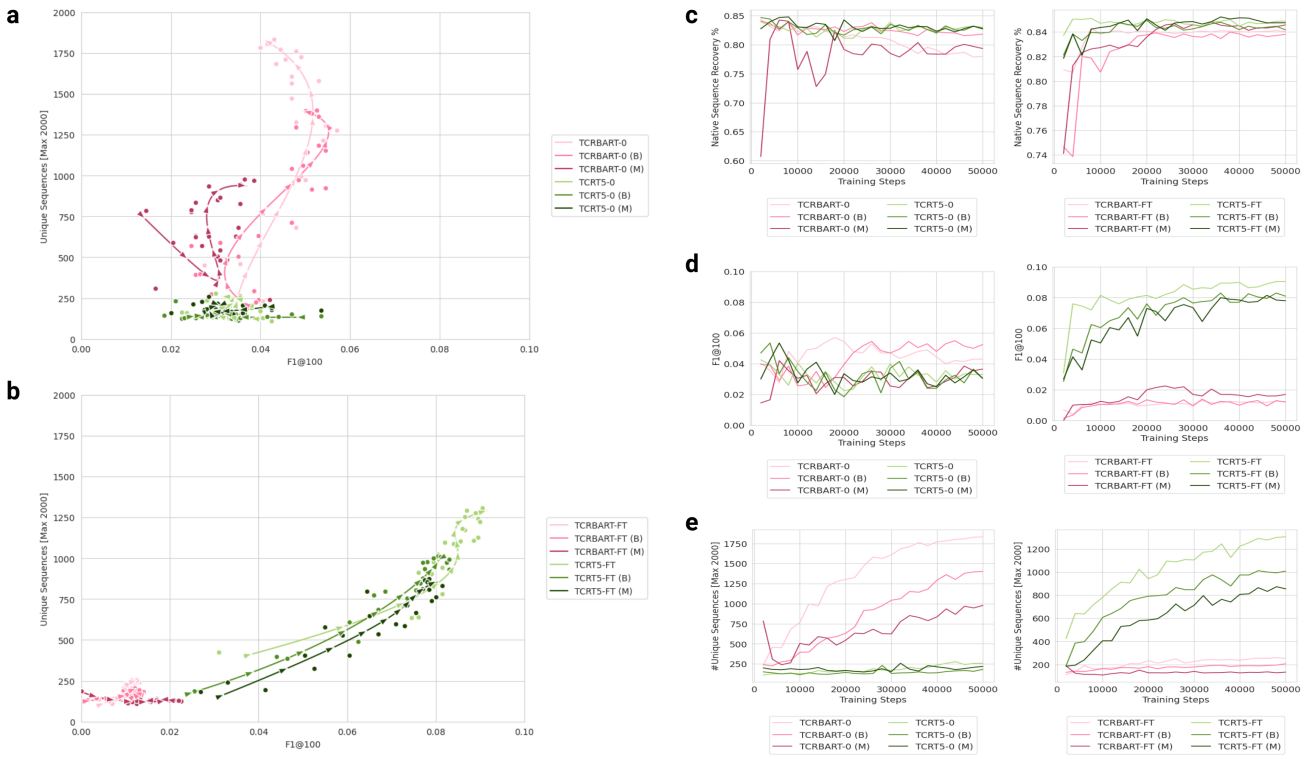
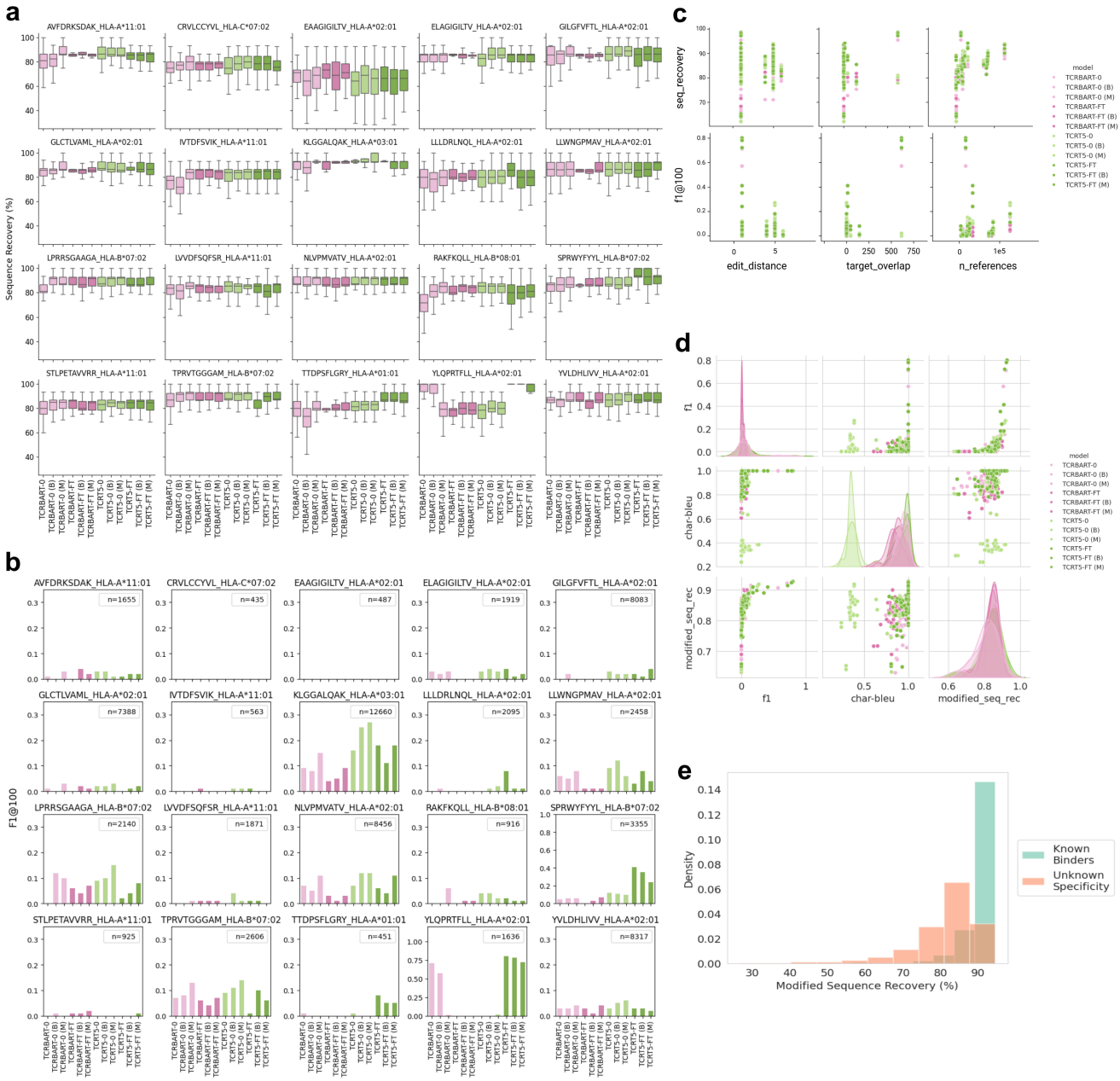


Figure 5: Benchmark Performance on IMMREP23 Unseen Epitopes. (a) Table of unseen pMHCs along with their species, originating proteins' amino acids, and the number of validated cognate TCRs. FTDALGIDEY, SALPTNADLY, TSDACMMTMY are designated "unseen" epitopes from IMMREP23 since they do not appear in IEDB and VDJdb. Here we include TDLGQNLLY in the unseen epitopes given its absence from our training set. (b) Metrics colored by model for k=1000 generations. Mean values are reported with error bars (1 SD). (c) Distribution of sequence recoveries are shown for 1000 conditional generations from TCRT5, GRATCR, and ER-BERT vs the unconditional soNNia `'-ppost'` generative model. (d) Bootstrapped test of significance. Our chosen test statistic is the number of generations that have at least a 90% sequence recovery rate for a fixed number of generations (k), in this case k=1000. We compare the observed test statistic using TCRT5's generations against the soNNia (yellow) and random reference (purple) null distributions. The observed statistic for GRATCR (red arrow) and ER-BERT (blue arrow) are reported as well. The soNNia null is constructed by sampling 1000 sequences from soNNia `'ppost'` and computing the sequence recoveries against the ground truth CDR3β sequences. The random reference null uses soNNia `'ppost'` to generate fake ground truth CDR3β sequences, equal in number to the ground truth TCRs per epitope, to use for sequence recovery calculations with the TCRT5 generations. Empirical p-values p_1 and p_2 are calculated as the number of cases from null distribution 1 and 2 respectively where the test statistics are greater than or equal to the observed statistic divided by the number of trials (1000). A p-value of 0.0 is binned as $< .0001$.

D EXTENDED DATA FIGURES



Extended Data Figure 1: Training dynamics highlight the robustness of pretrained models across checkpoints. Diversity vs. accuracy (F1) plotted for model checkpoints with smoothed interpolated splines and associated arrows showing the direction of model checkpoints through their training trajectory for: (a) Randomly initialized models (zero pre-training) (b) Pre-trained and finetuned models (c) Native sequence recovery for each checkpoint, colored by model, with panel split by pre-training status. (d) F1@100 for all checkpoints by pretraining status. (e) Number of unique generations for each checkpoint across training for all models with panel split by pre-training status. All models checkpoints were taken every 2000 steps across 20 epochs.



Extended Data Figure 2: Atomic Metrics. (a) Box and whisker plot of sequence recoveries split by individual pMHC and model. (b) Barplot showing F1@100 score per model and pMHC. Each subplot is demarcated with the number of reference CDR3βs in the top right corner. (c) Scatterplot showing relationship between accuracy metrics (sequence recovery and F1@100) and input features (edit distance to closest training pMHC, TCR overlap with closest training pMHC (by edit distance), and number of references (known TCR binders). (d) Correlation plot between sequence-derived metrics. Pairplot showing the pairwise relationships of F1@100, Char-BLEU, and (modified) sequence recovery, across model variants. Modified sequence recovery is calculated by first removing exact matches to the generated sequences from the reference sets and calculating sequence recovery to the closest sequence. (e) Histogram of modified sequence recovery values stratified by known binding status.

a

TCRBART-0 (M)

CDR3b	Cognate Validation Epitopes
CASSLGGNEQFF	[ELAGIGILTV, NLVPMVATV, RAKFKQLL, SPRWYFYYL]
CASSPGTGSYEQYF	[LLWNGPMAV, LPRRSGAAGA, RAKFKQLL, TPRVTGGGAM]
CASSLGTGGSYEQYF	[AVFDRKSDAK, KLGGALQAK, RAKFKQLL]
CASSLSGGSYNEQFF	[AVFDRKSDAK, KLGGALQAK, RAKFKQLL]
CASSLGGTDTQYF	[AVFDRKSDAK, GLCTLVAML, KLGGALQAK]
CASSPGQGYEQYF	[ELAGIGILTV, LLWNGPMAV, NLVPMVATV]
CASSPGTGGTDTQYF	[KLGGALQAK, LPRRSGAAGA, TPRVTGGGAM]
CASSLGQGNTEAFF	[KLGGALQAK, LPRRSGAAGA, TPRVTGGGAM]
CASSLLAGGTDTQYF	[LPRRSGAAGA, NLVPMVATV, TPRVTGGGAM]
CASSLGGSYEQYF	[LPRRSGAAGA, SPRWYFYYL, TPRVTGGGAM]
CASSLGGYEQYF	[GLCTLVAML, TPRVTGGGAM]
CASSLGTGSYEQYF	[KLGGALQAK, LLWNGPMAV]
CASSLGQGNQPQHF	[KLGGALQAK, LLWNGPMAV]
CASSLGGTGELFF	[KLGGALQAK, LLWNGPMAV]
CASSLGQGTDTQYF	[LPRRSGAAGA, TPRVTGGGAM]
CASSPGQGSTDTQYF	[LPRRSGAAGA, TPRVTGGGAM]
CASSLGQGGSYEQYF	[LPRRSGAAGA, TPRVTGGGAM]
CASSLGAGGTDTQYF	[LPRRSGAAGA, TPRVTGGGAM]
CASSLGTSGSYEQYF	[LPRRSGAAGA, TPRVTGGGAM]
CASSLAGGYEQYF	[NLVPMVATV, SPRWYFYYL]
CASSLGGTEAFF	[NLVPMVATV, TPRVTGGGAM]

TCRBART-0 (B)

CDR3b	Cognate Validation Epitopes
CASSLGGSYEQYF	[LPRRSGAAGA, SPRWYFYYL, TPRVTGGGAM]
CASSLGGTGELFF	[KLGGALQAK, LLWNGPMAV]
CASSPGTYEQYF	[LPRRSGAAGA, TPRVTGGGAM]
CASSLGYEQYF	[LPRRSGAAGA, TPRVTGGGAM]
CASSLAGSYEQYF	[LPRRSGAAGA, SPRWYFYYL]
CASSLGGYEQYF	[LPRRSGAAGA, TPRVTGGGAM]
CASSLGETQYF	[LPRRSGAAGA, NLVPMVATV]

TCRBART-0

CDR3b	Cognate Validation Epitopes
CASSLGQGNTEAFF	[KLGGALQAK, TPRVTGGGAM]
CASSLGQGNQPQHF	[KLGGALQAK, LLWNGPMAV]
CASSLGGTEAFF	[NLVPMVATV, TPRVTGGGAM]

b

TCRT5-FT (M)

CDR3b	Cognate Validation Epitopes
CASSLGQGNTEAFF	[KLGGALQAK, LPRRSGAAGA, NLVPMVATV, TPRVTGGGAM]
CASSPGTGSYEQYF	[LLWNGPMAV, LPRRSGAAGA, TPRVTGGGAM]
CASSLGGGTEAFF	[ELAGIGILTV, NLVPMVATV]
CASSPGQGYEQYF	[ELAGIGILTV, NLVPMVATV]
CASSLGTGELFF	[KLGGALQAK, NLVPMVATV]
CASSLGGTGELFF	[KLGGALQAK, LLWNGPMAV]
CASSLLAGGTDTQYF	[LPRRSGAAGA, NLVPMVATV]
CASSLGTGGNQPQHF	[LPRRSGAAGA, TPRVTGGGAM]
CASSPGTGGTDTQYF	[LPRRSGAAGA, TPRVTGGGAM]
CASSLGTSGSYEQYF	[LPRRSGAAGA, TPRVTGGGAM]
CASSLAGGYEQYF	[NLVPMVATV, SPRWYFYYL]

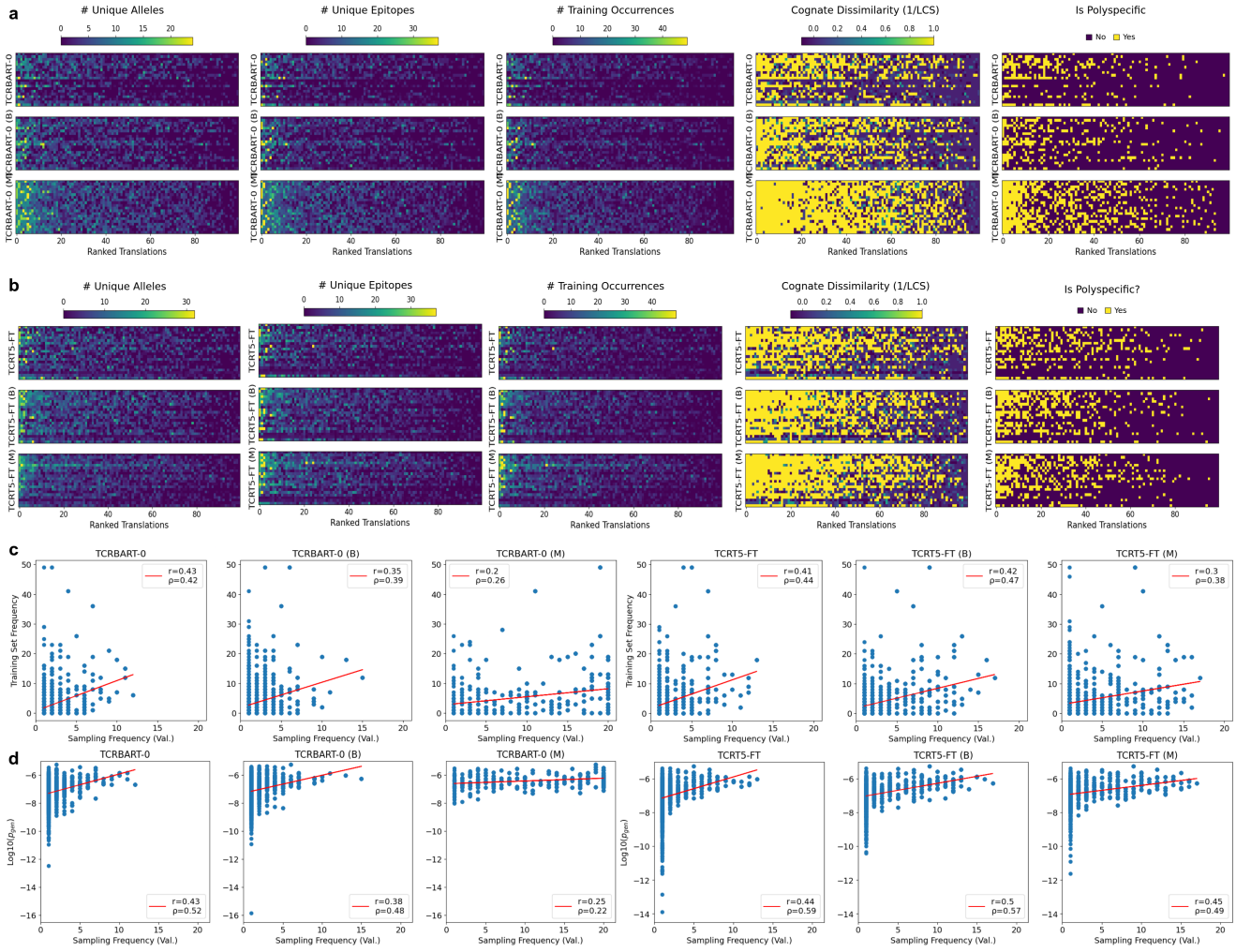
TCRT5-FT (B)

CDR3b	Cognate Validation Epitopes
CASSLGQGGTEAFF	[AVFDRKSDAK, KLGGALQAK]
CASSLGTGGNQPQHF	[GILGFVFTL, TPRVTGGGAM]
CASSLGGYEQYF	[GLCTLVAML, LPRRSGAAGA]
CASSLGTGELFF	[KLGGALQAK, NLVPMVATV]
CASSPGTGGTDTQYF	[KLGGALQAK, TPRVTGGGAM]
CASSLGSYEQYF	[LLLDRLNQL, YVLDHLIVV]
CASSPGTGSYEQYF	[LLWNGPMAV, TPRVTGGGAM]
CASSLAGGYEQYF	[NLVPMVATV, SPRWYFYYL]

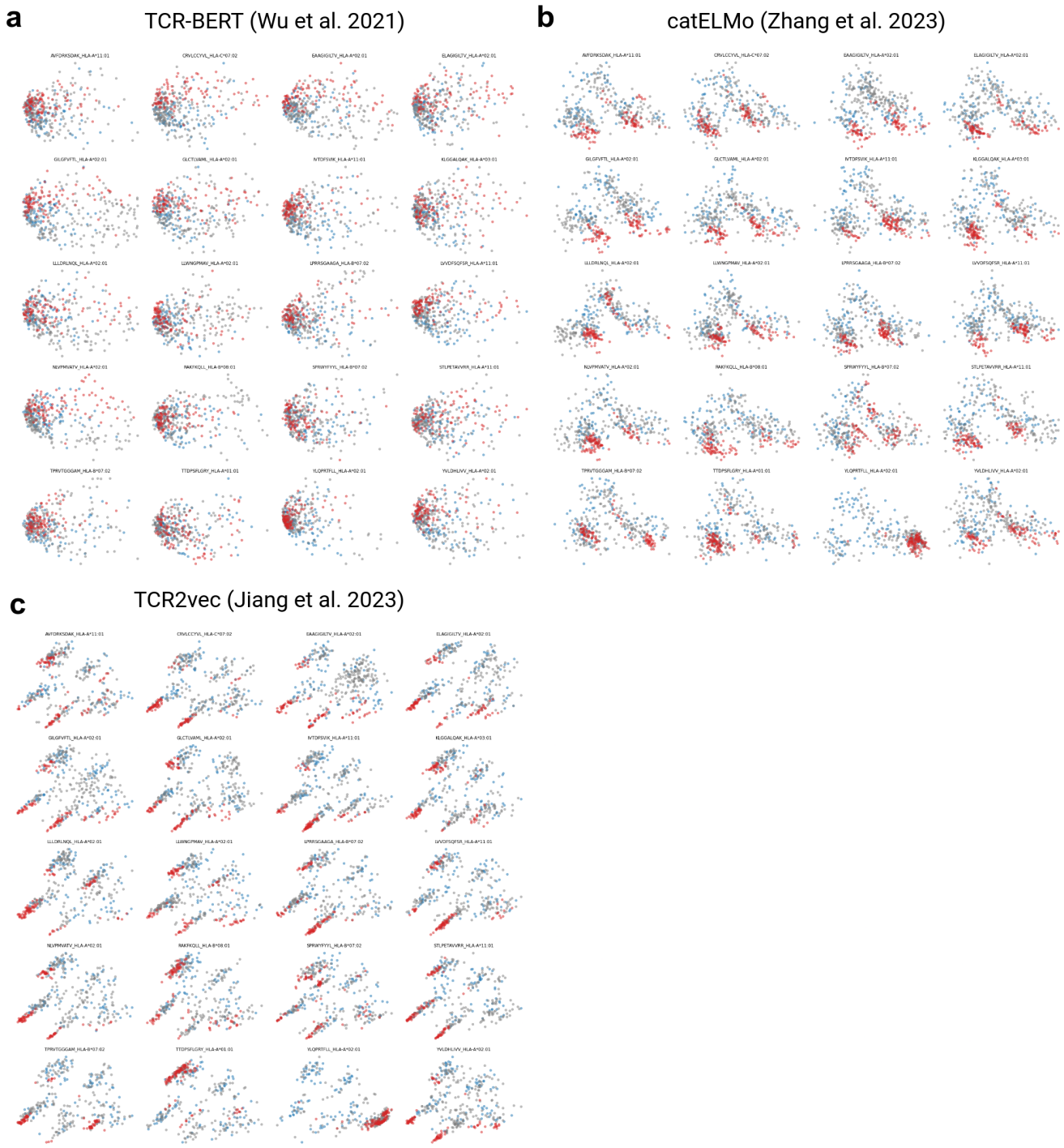
TCRT5-FT

CDR3b	Cognate Validation Epitopes
CASSLGETQYF	[ELAGIGILTV, NLVPMVATV]
CASSLGTGELFF	[KLGGALQAK, NLVPMVATV]
CASSLGTGGSYEQYF	[KLGGALQAK, RAKFKQLL]
CASSLGQGNTEAFF	[KLGGALQAK, TPRVTGGGAM]
CASSLGSYEQYF	[LLLDRLNQL, YVLDHLIVV]
CASSLLAGGTDTQYF	[LPRRSGAAGA, NLVPMVATV]
CASSLAGGYEQYF	[NLVPMVATV, SPRWYFYYL]

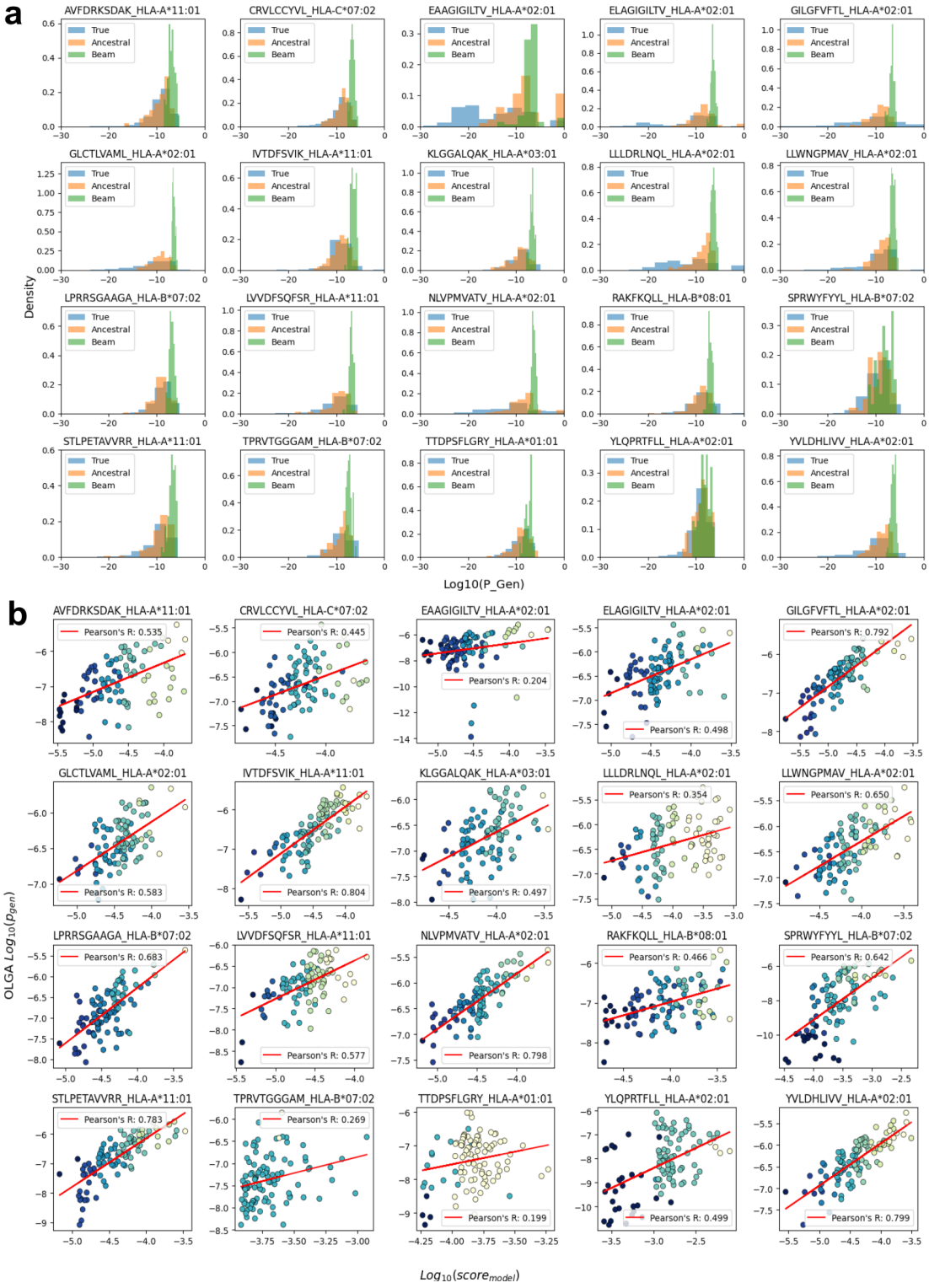
Extended Data Figure 3: Multi-task models sample more known validated polyspecific TCR sequences. (a) Subset of TCRBART-0 generations across model variants that are known binders to more than one validation pMHC (may be from the same disease context). (b) Subset of TCRT5-FT generations across model variants that are known binders to more than one validation pMHC (may be from the same disease context). Each row is an individual CDR3β sequence that was generated for and found in the experimentally validated set of reference TCRs for the listed validation pMHCs.



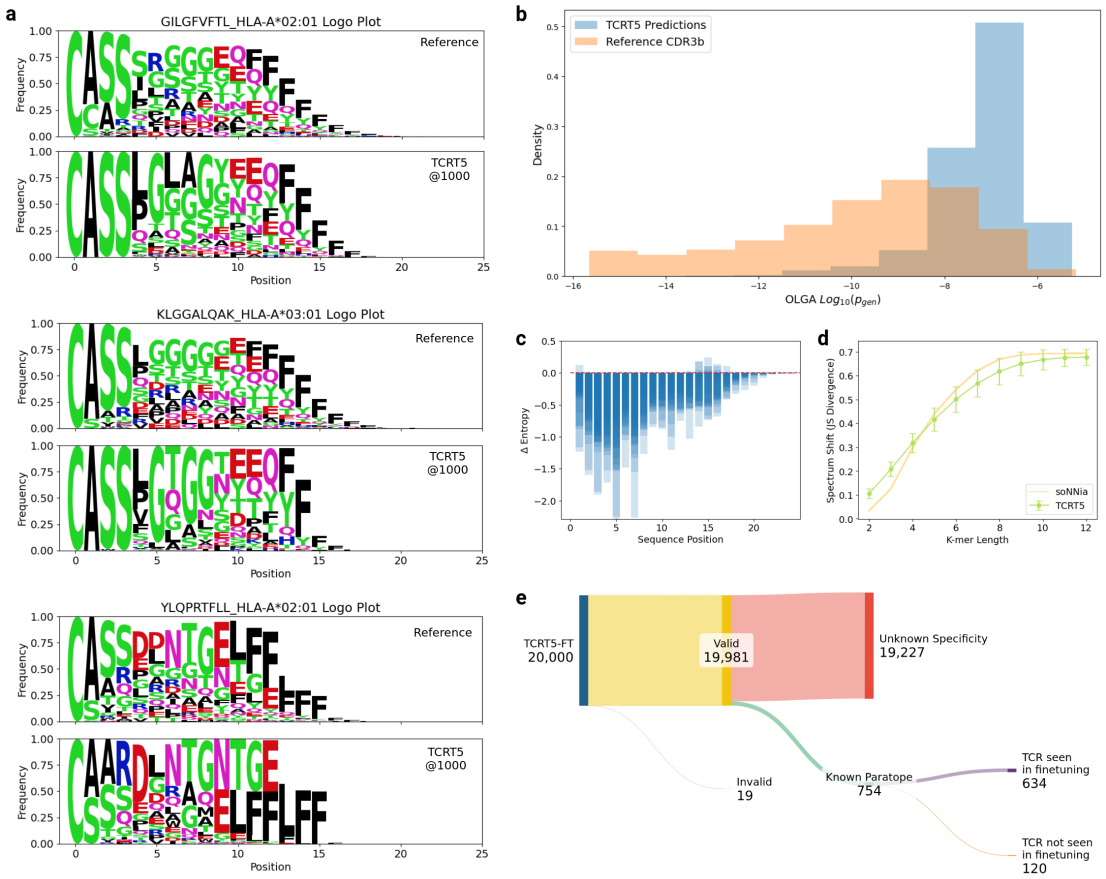
Extended Data Figure 4: Exploring polyspecificity vs. training set statistics across baseline, bidirectional, and multi-task model variants (a) Heatmap of ranked TCRBART-0 translations across pMHCs colored by number of known alleles, known epitopes, training set frequency, epitope dissimilarity, and membership status in the 915 polyspecific TCRs. (b) Analogous heatmap as panel 'a' but for TCRT5-FT generations. (c) Correlation plots for TCRBART-0 and TCRT5-FT model generations and training set occurrence. Line of best fit is shown in red. Pearson's r and Spearman's ρ are also provided for each model. (c) Correlation plots for TCRBART-0 and TCRT5-FT sampling frequency across epitopes and training OLGA p_{gen}. Line of best fit is shown in red. Pearson's r and Spearman's ρ shown at bottom right for each subplot.



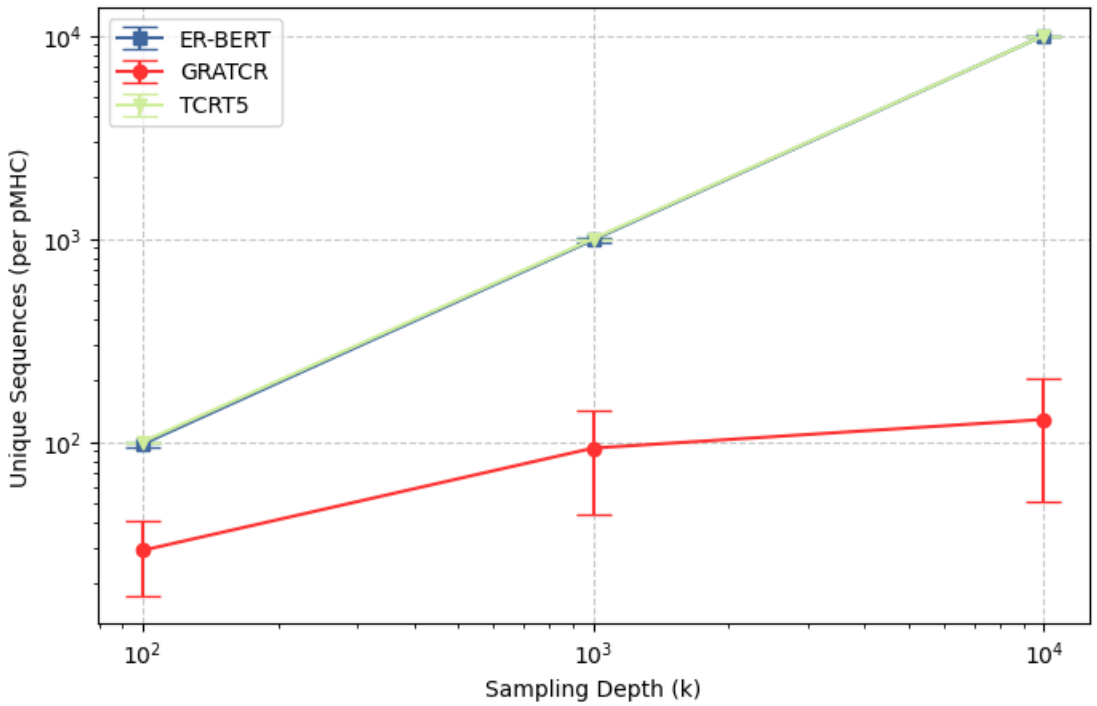
Extended Data Figure 5: CDR3β embeddings highlight reduction in sampled TCR space. PCA dimensionality reduction of embeddings generated by sequence based methods are shown for: (a) TCR-BERT (Wu et al., 2021) (b) catELMo (Zhang et al., 2024a) (c) TCR2vec (Jiang et al., 2023). Red points indicate sequences generated by TCRT5, gray corresponds to reference translations, and blue points are soNNia generated sequences. Reference TCRs are downsampled to 200 sequences and 100 background sequences are shown.



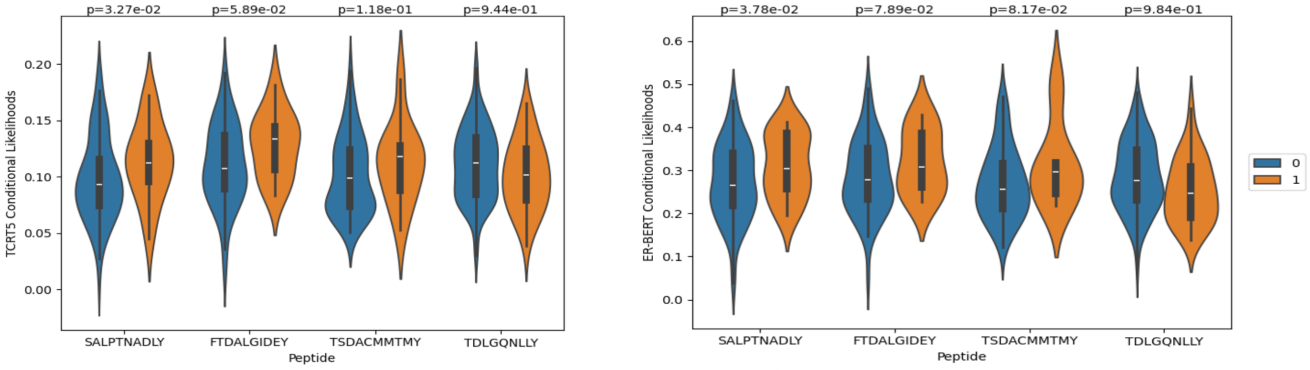
Extended Data Figure 6: TCRT5 Sequence Likelihoods. (a) Histograms showing the OLGA p_{gen} values for the reference CDR3βs as well as those generated by beam search and ancestral sampling methods. (b) Correlation plots showing the model scores (model sequence likelihoods) against the biophysical OLGA p_{gen}. Axes are log10 scaled. Red line is the best fit line with associated Pearson's R.



Extended Data Figure 7: TCRT5 Metrics @1000 (a) Sequence logo plot generated from TCRT5 for the canonical GILGFVFTL (Influenza A), KLGGALQAK (CMV), and YLQPRTFLL (SARS-CoV2) from 1000 generations instead of 100. (b) TCRT5@1000 with beam search still preferentially samples sequences at the right tail of OLGA generation probabilities. (c) Generated sequences experience a decrease in Shannon entropy across most positions, however, some pMHC examples exhibit an increase in entropy compared to reference sequences. Barplots for individual pMHCs are overlaid on one another. (d) K-mer spectrum shift plot showing the Jensen Shannon divergence between generated and reference sequences for TCRT5@1000. Error bars mark the mean and 1-standard deviation across validation pMHCs. Mean soNNia values are shown per simulated run, with 1000 generations per pMHC per run over 100 simulations. (e) Sankey diagram of TCRT5@1000 generations showing the validity as measured by nonzero generation probability, known binding status, and training set membership.



Extended Data Figure 8: Unique Sequences at Sampling Depth k Number of unique sequences returned per pMHC by beam search decoding as implemented in TCRT5, GRATCR, and ER-BERT. Markers indicate the mean number of unique sequences and error bars represent the standard deviation across the IMMREP2023 "private" antigens.



Extended Data Figure 9: Conditional Likelihoods of IMMREP23 Sequences Conditional likelihoods of true positive '1' and synthetic negative '0' CDR3β sequences from the IMMREP dataset were passed through TCRT5 (right) and ER-BERT (left) to get conditional likelihoods for each source-target pair. P-values were computed using the one-sided Mann-Whitney U test.

Extended Data Table 1: Characterization of Train/Val. Target Overlap

VAL. PEPTIDE	CLOSEST TRAIN PEPTIDE(S)	EDIT DISTANCE	CDR3β OVERLAP
AVFDRKSDAK	RLFRKSNLK	5	0/1655
AVFDRKSDAK	AAFKRSCLK	5	0/1655
AVFDRKSDAK	AVGVGKSAL	5	0/1655
CRVLCCYVL	PVTLACFVL	5	0/435
CRVLCCYVL	CFVECAPVC	5	0/435
CRVLCCYVL	WPVTLACFVL	5	4/435
EAAGIGILTV	AAGIGILTV	1	2/487
ELAGIGILTV	ELAGIGALTV	1	1/1919
ELAGIGILTV	ELAAIGILTV	1	1/1919
ELAGIGILTV	ELAGIGLTV	1	5/1919
GILGFVFTL	GILEFVFTL	1	1/8083
GILGFVFTL	GILGLVFTL	1	1/8083
GILGFVFTL	GIWGFVFTL	1	0/8083
GLCTLVAML	ALNTLVKQL	4	0/7388
IVTDFSVIK	IPTDFTISV	5	0/563
IVTDFSVIK	ITNFKSVLY	5	0/563
IVTDFSVIK	YTDFSSEII	5	0/563
IVTDFSVIK	HVTFFIYNK	5	0/563
KLGGALQAK	ALGGLLTMV	5	0/12660
KLGGALQAK	KLFAAETLK	5	0/12660
KLGGALQAK	CLGGLLTMV	5	1/12660
KLGGALQAK	MLWGYLQYV	5	0/12660
LLLDRLNQL	LLLLDRLNQL	1	146/2095
LLWNGPMAV	LLFGPVYV	4	0/2458
LLWNGPMAV	LLEWLAMAV	4	0/2458
LLWNGPMAV	LLFGYPVAV	4	0/2458
LPRRSGAAGA	LPSYAAFAT	5	0/2140
LPRRSGAAGA	LPSYAALAT	5	0/2140
LVVDFSQFSR	HLVDFQVTI	6	1/1871
LVVDFSQFSR	RVVVLSFEL	6	0/1871
LVVDFSQFSR	VVDSYYSLL	6	0/1871
LVVDFSQFSR	ALVYFLQSI	6	0/1871
LVVDFSQFSR	LLHGFSFYL	6	0/1871
LVVDFSQFSR	LVQSTQWSL	6	0/1871
LVVDFSQFSR	VLCNSQTSL	6	0/1871
NLVPMVATV	NLVPVVATV	1	1/8456
NLVPMVATV	NLVPQVATV	1	1/8456
NLVPMVATV	NLVPMVASV	1	1/8456
NLVPMVATV	NLVAMVATV	1	2/8456
NLVPMVATV	NLVGMVATV	1	1/8456
NLVPMVATV	ALVPMVATV	1	1/8456
NLVPMVATV	NLVPTVATV	1	1/8456
RAKFKQLL	RLSFKELLV	4	0/916
SPRWYFYYL	LPRWYFYYL	1	14/3355
STLPETAVVRR	GLPWNVVRI	6	0/925
TPRVTGGGAM	APRITFGGL	5	0/2606
TTDPSFLGRY	HTTDPSFLGRY	1	46/451
YLQPRTFLL	YLQPRTFL	1	606/1636
YLQPRTFLL	YLRPRTFLL	1	0/1636
YVLDHLIVV	KVLEYVIKV	5	1/8317
YVLDHLIVV	SVLLFLAFV	5	1/8317
YVLDHLIVV	TVYSHLLLV	5	2/8317
YVLDHLIVV	VLLFLAFVV	5	0/8317

Extended Data Table 2: Characterization of Train/IMMREP23 Target Overlap

TEST PEPTIDE	CLOSEST TRAIN PEPTIDE(S)	EDIT DISTANCE	CDR3β OVERLAP
FTDALGIDEY	ATDALMTGY	5	0/12
SALPTNADLY	SLYNTVATLY	5	0/16
SALPTNADLY	ALPETTADI	5	0/16
SALPTNADLY	SLFNTVATLY	5	0/16
SALPTNADLY	NLQSNHDLY	5	0/16
TDLGQNLLY	TALALLLLD	5	0/32
TDLGQNLLY	VSDGGPNLY	5	0/32
TDLGQNLLY	FLTENLLLY	5	0/32
TDLGQNLLY	TLYSLTLLY	5	0/32
TDLGQNLLY	GTDLEGNFY	5	0/32
TDLGQNLLY	TPSGTWLTY	5	0/32
TDLGQNLLY	TLSGTWLTY	5	0/32
TSDACMMTMY	TSAMHTMLF	5	0/11
TSDACMMTMY	TSAMQTMLF	5	0/11
TSDACMMTMY	ATDALMTGY	5	0/11
TSDACMMTMY	NSSTCMMCY	5	0/11
TSDACMMTMY	SSSTCMMCY	5	0/11