# Multi-Task Training Increases Native Sequence Recovery of Antigen-Specific T-cell Receptor Sequences

**Dhuvarakesh Karthikeyan** [1 2] **Benjamin Vincent** [1 3 4] **Alexander Rubinsteyn** [1 2 4]

## Abstract

T-cells are a critical component of the adaptive immune system that use specialized T-cell receptors (TCRs) to bind non-self peptide fragments presented by major histocompatibility complex (MHC) molecules on the surface of other cells. Given their importance, a foundation model of TCR specificity that is capable of reliably mapping between TCR sequences and their cognate peptide-MHC (pMHC) ligands remains an unmet need. This study presents a key step towards developing a comprehensive foundation model by exploring the bi-directional mapping of both pMHCs to their corresponding TCRs, and vice versa. While validation performance was significantly worse in the TCR to pMHC direction given the highly asymmetric distribution of pMHC data, we find that the bidirectionally trained model outperformed the model trained in a single pMHC to TCR direction, at the cost of diversity. We work through a rigorous evaluation using well characterized pMHCs and present our framework and findings as a potential direction towards a unified generative foundation model of TCR:pMHC cross-reactivity.

## 1. Introduction

The T-cell receptor (TCR) and peptide-MHC (pMHC) interaction is a fundamental immunological event that triggers our bodies' T-cell response against cancer, viruses, and even self-antigens. As such, there has been significant effort in developing a model of TCR specificity to map TCR sequences to their cognate epitopes, or design TCRs against antigens of interest (Hudson et al., 2023). Such a model would revolutionize cellular therapies (Cao et al., 2021; Tzannou et al., 2017; Ellebrecht et al., 2016; Poole et al., 2022) and our ability to contextualize the T-cell response against a specific pathogen. However, a foundation model that reliably goes between TCRs and their corresponding pMHC ligands remains blocked by issues of extreme data sparsity and noisy provenance coupled with a complicated cross-reactivity landscape (Wooldridge et al., 2011; Sewell, 2012).

In addition to efforts in the structural space (Ribeiro-Filho et al., 2024), recent work explored the explicit framing of the TCR reactivity problem as a sequence-to-sequence (seq2seq) task, demonstrating the capacity of sequence-based models to design TCR sequences against a specific pMHC (Karthikeyan et al., 2023; Fast et al., 2023). Sequence-to-sequence learning introduced the concept of an encoder:decoder model that was end-to-end trainable in the context of machine translation tasks (Sutskever et al., 2014; Kalchbrenner & Blunsom, 2013; Cho et al., 2014). However, this framework quickly became a paradigm-shifting method for training deep networks to maximize the conditional likelihood of target sequences given a source sequence for arbitrary source:target pairs such as question answering and text summarization, which we apply here to the TCR:pMHC.

A key consideration for encoder:decoder models is the amount of parallel data (source-target pairs) available, a requirement that scales with the complexity of the sequence mapping. In machine translation, a number of methods have been developed to specifically address the challenge maximizing the information usage of low-resource languages, when labeled data is limited and expensive to generate (Haddow et al., 2022). Techniques such as back-translation (Sennrich et al., 2016), self-training (He et al., 2020), transfer learning (Liu et al., 2020), and semi-supervised translation approaches that leverage monolingual data have all been proposed with varying degrees of success, dependent on the problem specifics. However these approaches have been shown to exacerbate overfitting to specific domain contexts or sequence distributions (Shen et al., 2020). To combat these effects, multi-task learning of bidirectional transla-

[1] Computational Medicine Program, University of North Carolina, Chapel Hill, USA [2] Curriculum of Bioinformatics and Computational Biology, University of North Carolina, Chapel Hill, USA [3] Department of Microbiology and Immunology, University of North Carolina, Chapel Hill, USA [4] Lineberger Comprehensive Care Center, Chapel Hill, USA. Correspondence to: Dhuvarakesh Karthikeyan <dhuvikarthikeyan@gmail.com>.

*Proceedings of the ICML 2024 Workshop on Accessible and Efficient Foundation Models for Biological Discovery*, Vienna, Austria. 2024. Copyright 2024 by the author(s).

tion models and mulitilingual translation models have been shown to significantly improve translation quality by sharing representations and aligning latent spaces across multiple languages (Niu et al., 2018; Ding et al., 2021). In this work, we explore the use of multi-task training to train on both directions of the TCR:pMHC specificity problem and evaluate its effect on the conditional generation of TCR sequences as well as its potential to simultaneously sample accurate cognate pMHCs using the same weights.

## 2. Methodology

### 2.1. Setup

For our setup, we largely follow the same seq2seq problem formulation as laid out in (Karthikeyan et al., 2023). Briefly, we adopt the amino acid level tokenization scheme and restrict the sequence space to the high entropy portions of the TCR:pMHC interface, namely the CDR3b, peptide, and MHC pseudo-sequence as defined in (Hoof et al., 2009). In addition we use the BART and T5 encoder:decoder transformer architectures (Figure S1a), with modified hyper-parameters to control for their total parameter count (Table 1). However, in order to introduce flexibility with the directionality, we leverage the concept of task prefixes introduced in (Raffel et al., 2020) as a special token for the T5 model architecture, using the BART model and its prefix-agnostic tokenization scheme for comparison (Figure S1b). Additionally, we retain the use of the '[SEP]' token to separate the peptide, which typically spans 8-11 amino acids, and the fixed length psuedosequence.

### 2.2. Dataset

For the generation of a parallel corpus, we used 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 injection of weakly-labeled data derived from the MIRA (Dines et al., 2020)) which contained CDR3b and peptide sequences along with the HLA-type of the individual, instead of the presenting MHC allele. MHC allele was inferred using MHCFlurry2.0's (O'Donnell et al., 2020) ranked presentation score metric among the HLA-type. Of importance, these pseudo-synthetic examples were not used in evaluation. More on the dataset standardization procedures can be found in the Supplementary Methods (A.2.1). The resulting dataset was subsequently deduplicated to remove near duplicates which we found to marginally help overall performance, in accordance with (Lee et al., 2022) (Figure S3, Figure S4).

In order to assess the capacity of the models to produce not only plausible TCR sequences, but specifically antigen-specific sequences, we pivot from the balanced split on allele strategy used previously (Karthikeyan et al., 2023), motivated by our analysis in (Appendix A.1). Instead, to explicitly evaluate sequence generations for their antigen-specificity, we curated a target-rich dataset by separating out the top-20 most represented real pMHCs for validation and evaluate our models on how well their generations overlap with the real validated sequences. We train on the remaining data, further removing the occurrences of the held-out epitopes bound alternate MHCs to ensure a validation set of unseen epitopes. We allowed a high degree of sequence overlap with training sequences both out of necessity, given the sparsity of well characterized pMHCs, but also to qualitatively characterize their performance. The degree in which these sequences exhibit training set similarity is reflected in (Table 3). This resulted in a final dataset split of 330k training sequences (N=6989 pMHCs) and 68k validation sequences (N=20 pMHCs). A key limitation of this dataset is its highly skewed HLA distribution towards well studied alleles (A*02:01, A*03:01, A*11:01, etc).

### 2.3. Model Training

For our experiments, we considered three model training regimes. The baseline model was trained on the pMHC → TCR direction, a bidirectional model was trained on both directions, and finally a multi-task model was trained on both directions as well a masked language modeling objective for both TCR and pMHC sequences. The models were trained using the 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).

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

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

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

The baseline models were trained using the above cross entropy objective whereas the bidirectional and multi-task models were trained on a mutli-term objective, comprising of a linear combination of individual loss terms corresponding to each task/direction. This was achieved using a batch processing algorithm (Algorithm 1), where each batch was rearranged into one of four sequence-to-sequence mapping possibilities with equal probability and the model was trained on target reconstruction. 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 $\mathcal{L}_{pmhc\rightarrow tcr}$and$\mathcal{L}_{tcr\rightarrow pmhc}$ (Equation 2). For the multi-task model, the mapping possibilities are: 1) pMHC → TCR 2) pMHC → TCR 3) Corrupted pMHC* → pMHC 4) Corrupted TCR* → TCR, which combine to to form $\mathcal{L}_{multi}$ (Equation 3). For the purposes of comparison between models, each of the models was trained for 20 epochs, from which the checkpoint with the highest combined F1 and sequence recovery was chosen. Interestingly, we observed that training for longer resulted in worse validation performance and the convergence dynamics between TCRBART and TCRT5 were notably dissimilar. We chose this approach to better capture the models' real world utility, as opposed to training for a fixed number of steps and evaluating the learning objective for its accuracy and efficiency.

**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($\mathbf{y}$, $\hat{\mathbf{y}}$)
```

### 2.4. Conditional Sampling

Given its marked performance gains over other sampling methods in the TCR:pMHC sequence space (Karthikeyan et al., 2023), we fix beam-search as our chosen method of sequence generation for all generation-based metrics except for the Char-BLEU score, for which we used single generations via greedy decoding. Both decoding methods and the broader class of mode-seeking methods, aim to maximize for the highest probability conditional sequence. Recently, however, mode-seeking algorithms have come under scrutiny for sampling only a small portion of the true target distribution, as noted in (Eikema & Aziz, 2020). Instead of sampling tokens directly from the whole conditional distribution: $y_t \sim P(y_t|y_{<t}, x, \theta)$. Given source sequence $x$ and model parameters $\theta$, they try to approximate the $y^{MAP} = \arg\max_{y\in Y} \log p(y|x, \theta)$ which explores the conditional distribution about the mode. However, combined with the ability to assess and maintain longer high probability subsequences, the use of beam search results in a powerful method for sampling native sequence-like predictions.

### 2.5. Evaluation

*In-silico* evaluation of TCR:epitope specificity can be challenging since unlike antibodies, binding affinity and structural fit alone do not predict functional response (Singh et al., 2017) of TCRs. Thus, motivated by a need to evaluate model performance on antigen-specificity in an unequivocal manner, we build our evaluation around sampling exact CDR3b sequences from known experimental data on the unseen epitopes with a large enough sample of cognate TCRs (Appendix A.1). However, since generation of new data is both expensive and time consuming, we include metrics based on sequence similarity to known binders and characterize their concordance with F1 performance to determine their faithfulness and utility for future use on unseen epitopes with orders of magnitude fewer known cognate sequences. Our choice of metrics and their intuitions can be summarized in brief:

- **Char-BLEU**: Character-level weighted n-gram precision calculated on the greedy decoding sequences against the $k = 20$ closest reference sequences. We use NLTK's 'sentence_bleu' and 'corpus_bleu' functions to compute the single translation and whole dataset BLEU scores, respectively. (Papineni et al., 2002).

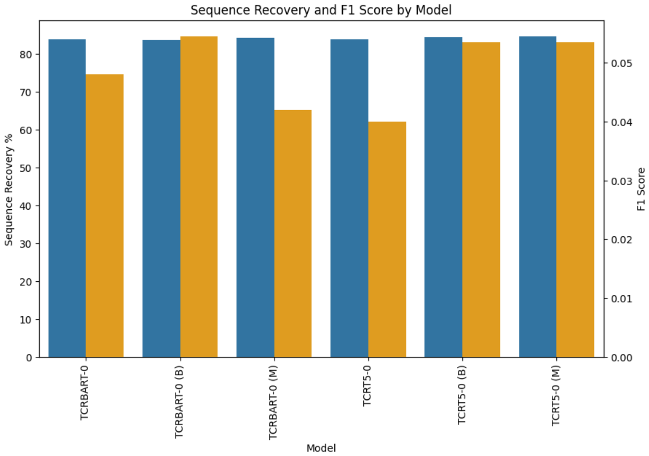
- **Perplexity**: Perplexity as a standard measure of language model performance, using the cross entropy loss calculated over the validation corpus.

- **Precision, Recall, and F1@K**: Precision, recall, and F1 measures exact whole sequence recovery, computed after sampling $K$ times (not necessarily $K$ separate sequences).

- **Native Sequence Recovery**: In the *de novo* protein design space, native sequence recovery is a useful indicator of model performance. Here, for each model prediction, the index-matched exact amino acid recovery with the closest match of the same sequence length is calculated.

## 3. Results

As shown in Figure 1, we first evaluate the dataset-level performance of the models trained in the various methods on TCR generations against the holdout dataset comprising the top-20 most well represented, *real* pMHCs. Of the models tested, the top performing models per architecture were the BART multi-task (TCRBART (M)) and T5 bidirectional model (TCRT5 (B)) with the former achieving an average F1@100 score of .042 and mean sequence recovery of 84.3%. TCRT5 (B) performed better on the F1 metric with a score of .054 and a sequence recovery rate of 84.4%. While the perplexity and CharBLEU scores were highly similar for the most part, we noticed a steep drop in CharBLEU performance for the T5 models with auxiliary tasks. We attribute this phenomena to the earlier checkpoints used, which for tasks of bidirectional and multi-task translation, may not have fully resolved for the single direction in which we are evaluating it. However for all of the models, adding the additional task directions increased both the mean and median F1 and sequence recovery.



| MODEL | CHARBLEU | PPL | F1@100 | % REC. |
|---|---|---|---|---|
| BART | 99.2 | **2.49** | .048/.01 | 83.9/85.0 |
| BART (B) | 98.7 | 2.53 | **.055**/.015 | 83.7/85.4 |
| BART (M) | **99.4** | **2.49** | .042/**.030** | **84.3/86.0** |
| T5 | **97.9** | **2.44** | .040/.01 | 83.9/85.3 |
| T5 (B) | 36.7 | 2.49 | **.054/.035** | 84.4/**85.9** |
| T5 (M) | 40.5 | 2.50 | **.054**/.02 | **84.7**/85.8 |

Figure 1: *Model Performance across Top-20 Dataset* a) Multiple bar chart showing the native sequence recovery and F1@100 score of the benchmarked models. b) Table of values showing dataset-level metrics for CharBLEU and Perplexity as well as mean/median values for F1 and Sequence Recovery. Results computed using 100 sequences generated using the beam search algorithm (beams=300).

To further understand the models' performance on individual pMHCs and their contribution to the global statistics, we computed the same metrics and stratified them by pMHC (Table 4). When independently evaluating the different pMHCs, we observe that the performances of the models are largely consistent across models within each pMHC for both sequence recovery as well as F1 score (Figure S5, Figure S6). The pMHC EAAGIGILTV-A*02:01 was the worst performing pMHC of the 20, showing poor model performance on both sequence recovery and the F1 score across all models. In contrast, KLGGALQAK-A*03:01 and NLVPMVATV-A*02:01, among others, stand out as examples where all models performed well. Given the cross-model consistency in performance, we sought to better understand the 20 epitopes in the context of the metrics and the training and validation data (Table 3).

Since the metrics all implicitly or explicitly rely on held-out sequences, we first checked for a correlation between the number of TCRs for a held-out example and its performance. Unsurprisingly, we observed a positive trend between the number of reference CDR3b sequences a model has and all of our performance metrics. Specifically, we note that a soft threshold where pMHCs with less 1000 reference CDR3bs have an F1 score of 0, while those with higher CDR3b counts had enough TCRs to where at least one coincided with the sampled sequences (Figure S6). Next, we sought to evaluate the concordance between the F1 score and the other sequence-based metrics. While the CharBLEU and F1 were loosely related, we see a strong relationship between the sequence recovery and F1 (Figure S7), indicating the utility of sequence recovery in assessing CDR3bs with lower known CDR3b abundancies. Building on this, we evaluate the sequence of recovery of random CDR3bs sampled from (Chen et al., 2020) and compare them to the reference CDR3bs for each antigen. We found that even random CDR3bs possess greater than 60% sequence identity with known antigen specific CDR3bs, which was significantly higher than random noise but lower than the median performance of sequences generated by the models (Figure S8), indicating our conditionally generated sequences capture signals that confer antigen specificity.

Qualitatively, we can begin to contextualize performance as driven by both apparent cases of memorization and generalization. For example, the SARS-Cov-2 epitope YLQPRTFLL had the highest performance of the pMHCs, with a sequence recovery of over 90% and an F1 score nearly 10x greater than the mean, which we find can be explained by the presence of an epitope in the training set that shares over 600 CDR3b sequences and has $d_{\text{edit}} = 1$ (Table 3). However, a low edit distance and high overlap do not necessarily guarantee outstanding performance, as evidenced by another SARS-Cov-2 epitope LLLDRLNQL. Remarkably, the second-best performing epitope was KLGGALQAK, with a $>$90% sequence recovery rate, despite the closest training sequence being five edits away and having an overlap of one. This may be due to the epitope having over 12,000 reference CDR3bs, so the likelihood of finding a matching or similar CDR3b here is far greater than the other epitopes.

In addition to model accuracy, we look at the diversity of sequences generated by these models, a known trade-off in the space of conditional generation (Vijayakumar et al., 2018). We find that while the multi-task models increase the performance of these models, they are prone to sample degenerate solutions by generating the same high probability TCRs across unrelated pMHCs (Figure S9) and also sample more sequences that were seen during training (Figure

S10). While these results highlight the models' capability to generate sequences with validated antigen-specificities, they also underscore the importance of contextualizing the accuracy in the greater context of model utility. The mixed results suggest that both memorization and generalization play roles in model performance and while learning the bi-directional mapping increases performance in the TCR generation space, the implicit self-consistency may be detrimental to model utility given the current sparsity in balanced pMHC data.

Finally, with regards to the reverse direction (TCR deorphanization), we found that both bidirectional and multi-task models fared substantially worse than the TCR generation direction. Interestingly, while the peptide-only Char-BLEU scores was near 0, the models generated plausible-looking MHC pseudosequences. We believe that the disparity in performance between tasks is due to combination of lack of tuning for generating longer sequences as well as the asymmetric number of pMHCs than TCRs in the training data. As such we hold off on reporting the standard metrics for this iteration of our study.

## 4. Discussion

In this study we set out to investigate the effect of multi-task training in improving the mapping quality between TCR and their corresponding pMHCs. We showed the improvement in model performance by including auxiliary tasks was matched with a decrease in diversity. To do so we constructed a dataset in which we could probabilistically expect to see exact sequence matches with the conditional generations. We find that the TCRBART and TCRT5 models showed similar performance, and that this performance is conserved across pMHCs and tied directly to the number of available reference TCRs. The performance of the multi-task models in generating accurate TCR sequences suggests that jointly learning both directions of the TCR-pMHC sequence mapping, along with the unconditional distributions of both modalities, enhances the ability to find motifs that are conserved across a large number of TCRs derived from different disease contexts. However, more analysis is required to determine if there is any biological relevance in these polyspecific motifs. The observed variability in performance across different pMHC epitopes highlights the critical role of training data composition and its relationship to the test set. Future work will explore leveraging 'monolingual' data for iterative back-translation and other semi-supervised techniques, to further improve model robustness, and engineering splits across more diverse TCR-pMHC interactions.

## References

Cao, X., Liu, G., Zhang, J., Zhao, Y., Chen, H., Zheng, H., Rui, W., Jia, L., Zhao, X., Lin, X., and Lu, P. A Novel CMV-Specific TCR-T Cell Therapy Is Effective and Safe for Refractory CMV Infection after Allogeneic Hematopoietic Stem Cell Transplantation. *Blood*, 138(Supplement 1):3848–3848, 11 2021. ISSN 0006-4971. doi: 10.1182/blood-2021-146446. URL https://doi.org/10.1182/blood-2021-146446.

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 Research*, 49(D1):D468–D474, 09 2020. ISSN 0305-1048. doi: 10.1093/nar/gkaa796. URL https://doi.org/10.1093/nar/gkaa796.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. Learning phrase representations using rnn encoder-decoder for statistical machine translation, 2014.

Dines, J. N., Manley, T. J., Svejnoha, E., Simmons, H. M., Taniguchi, R., Klinger, M., Baldo, L., and Robins, H. 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.

Ding, L., Wu, D., and Tao, D. Improving neural machine translation by bidirectional training. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t. (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.

Eikema, B. and Aziz, W. Is map decoding all you need? the inadequacy of the mode in neural machine translation, 2020.

Ellebrecht, C. T., Bhoj, V. G., Nace, A., Choi, E. J., Mao, X., Cho, M. J., Zenzo, G. D., Lanzavecchia, A., Seykora, J. T., Cotsarelis, G., Milone, M. C., and Payne, A. S. Reengineering chimeric antigen receptor t cells for targeted therapy of autoimmune disease. *Science*, 353(6295):179–184, 2016. doi: 10.1126/science.aaf6756. URL https://www.science.org/doi/abs/10.1126/science.aaf6756.

Fast, E., Dhar, M., and Chen, B. 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.

Haddow, B., Bawden, R., Barone, A. V. M., Helcl, J., and Birch, A. 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.

He, J., Gu, J., Shen, J., and Ranzato, M. Revisiting self-training for neural sequence generation, 2020.

Hoof, I., Peters, B., Sidney, J., Pedersen, L. E., Sette, A., Lund, O., Buus, S., and Nielsen, M. Netmhcpan, a method for mhc class i binding prediction beyond humans. *Immunogenetics*, 61:1–13, 2009.

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*, pp. 1–11, 2023.

Kalchbrenner, N. and Blunsom, P. Recurrent continuous translation models. In *Conference on Empirical Methods in Natural Language Processing*, 2013. URL https://api.semanticscholar.org/CorpusID:12639289.

Karthikeyan, D., Raffel, C., Vincent, B., and Rubinsteyn, A. 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.

Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C., and Carlini, N. Deduplicating training data makes language models better, 2022. URL https://arxiv.org/abs/2107.06499.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. Multilingual denoising pre-training for neural machine translation, 2020.

Mason, D. A very high level of crossreactivity is an essential feature of the t-cell receptor. *Immunology Today*, 19(9):395–404, 1998. ISSN 0167-5699. doi: https://doi.org/10.1016/S0167-5699(98)01299-7. URL https://www.sciencedirect.com/science/article/pii/S0167569998012997.

Nagano, Y. and Chain, B. 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.

Niu, X., Denkowski, M., and Carpuat, M. Bi-directional neural machine translation with synthetic parallel data. In Birch, A., Finch, A., Luong, T., Neubig, G., and Oda, Y. (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.

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. doi: https://doi.org/10.1016/j.cels.2020.06.010. URL https://www.sciencedirect.com/science/article/pii/S2405471220302398.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. 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.

Poole, A., Karuppiah, V., Hartt, A., Haidar, J. N., Moureau, S., Dobrzycki, T., Hayes, C., Rowley, C., Dias, J., Harper, S., Barnbrook, K., Hock, M., Coles, C., Yang, W., Aleksic, M., Lin, A. B., Robinson, R., Dukes, J. D., Liddy, N., Van der Kamp, M., Plowman, G. D., Vuidepot, A., Cole, D. K., Whale, A. D., and Chillakuri, C. Therapeutic high affinity t cell receptor targeting a krasg12d cancer neoantigen. *Nature Communications*, 13(1):5333, Sep 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-32811-1. URL https://doi.org/10.1038/s41467-022-32811-1.

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, 2020.

Ribeiro-Filho, H. V., Jara, G. E., Guerra, J. V. S., Cheung, M., Felbinger, N. R., Pereira, J. G. C., Pierce, B. G., and de Oliveira, P. S. L. Exploring the potential of structure-based deep learning approaches for t cell receptor design. *bioRxiv*, 2024. doi: 10.1101/2024.04.19.590222. URL https://www.biorxiv.org/content/early/2024/04/24/2024.04.19.590222.

Sennrich, R., Haddow, B., and Birch, A. Improving neural machine translation models with monolingual data, 2016.

Sethna, Z., Elhanati, Y., Callan, Curtis G, J., Walczak, A. M., and Mora, T. 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.

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

Shen, J., Chen, P.-J., Le, M., He, J., Gu, J., Ott, M., Auli, M., and Ranzato, M. The source-target domain mismatch problem in machine translation, 2020.

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., 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 Research*, 46(D1):D419–D427, 09 2017. ISSN 0305-1048. doi: 10.1093/nar/gkx760. URL https://doi.org/10.1093/nar/gkx760.

Singh, N. K., Riley, T. P., Baker, S. C. B., Borrman, T., Weng, Z., and Baker, B. M. Emerging concepts in tcr specificity: Rationalizing and (maybe) predicting outcomes. *The Journal of Immunology*, 199:2203 – 2213, 2017. URL https://api.semanticscholar.org/CorpusID:5375575.

Sutskever, I., Vinyals, O., and Le, Q. V. Sequence to sequence learning with neural networks, 2014.

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, 05 2017. ISSN 1367-4803. doi: 10.1093/bioinformatics/btx286. URL https://doi.org/10.1093/bioinformatics/btx286.

Tzannou, I., Papadopoulou, A., Naik, S., Leung, K., Martinez, C. A., Ramos, C. A., Carrum, G., Sasa, G., Lulla, P., Watanabe, A., Kuvalekar, M., Gee, A. P., Wu, M.-F., Liu, H., Grilley, B. J., Krance, R. A., Gottschalk, S., Brenner, M. K., Rooney, C. M., Heslop, H. E., Leen, A. M., and Omer, B. 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.

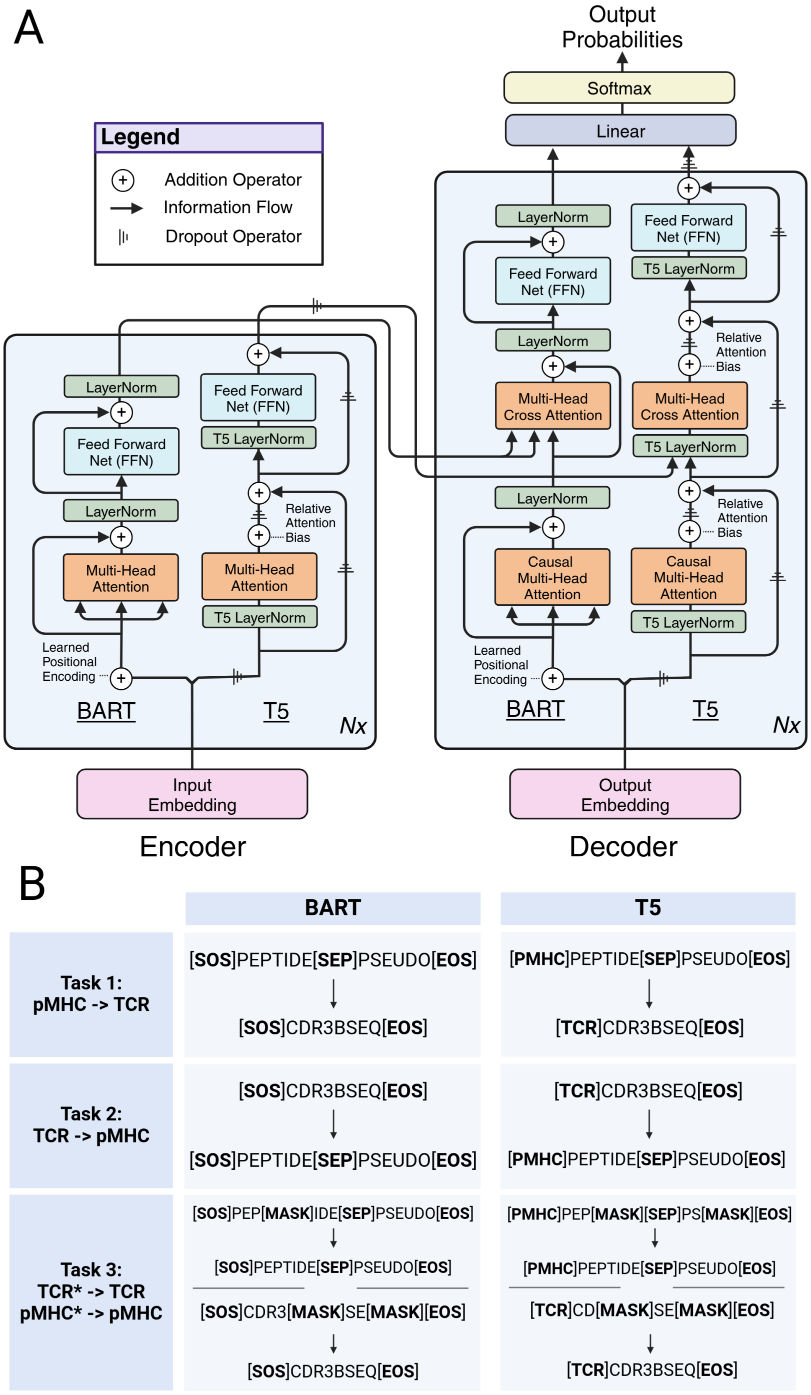
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need, 2023.

Vijayakumar, A. K., Cogswell, M., Selvaraju, R. R., Sun, Q., Lee, S., Crandall, D., and Batra, D. Diverse beam search: Decoding diverse solutions from neural sequence models, 2018.

Vita, R., Mahajan, S., Overton, J. A., Dhanda, S. K., Martini, S., Cantrell, J. R., Wheeler, D. K., Sette, A., and Peters, B. 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.

Wooldridge, E.-M., van den Berg, S., Miles, T., Dolton, C., Llewellyn-Lacey, P., and Peakman, S. 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.

## A. Appendix



| | BART | T5 |
|---|---|---|
| **Task 1: pMHC -> TCR** | [**SOS**]PEPTIDE[**SEP**]PSEUDO[**EOS**]<br>↓<br>[**SOS**]CDR3BSEQ[**EOS**] | [**PMHC**]PEPTIDE[**SEP**]PSEUDO[**EOS**]<br>↓<br>[**TCR**]CDR3BSEQ[**EOS**] |
| **Task 2: TCR -> pMHC** | [**SOS**]CDR3BSEQ[**EOS**]<br>↓<br>[**SOS**]PEPTIDE[**SEP**]PSEUDO[**EOS**] | [**TCR**]CDR3BSEQ[**EOS**]<br>↓<br>[**PMHC**]PEPTIDE[**SEP**]PSEUDO[**EOS**] |
| **Task 3: TCR* -> TCR pMHC* -> pMHC** | [**SOS**]PEP[**MASK**]IDE[**SEP**]PSEUDO[**EOS**]<br>↓<br>[**SOS**]PEPTIDE[**SEP**]PSEUDO[**EOS**]<br>[**SOS**]CDR3[**MASK**]SE[**MASK**][**EOS**]<br>↓<br>[**SOS**]CDR3BSEQ[**EOS**] | [**PMHC**]PEP[**MASK**][**SEP**]PS[**MASK**][**EOS**]<br>↓<br>[**PMHC**]PEPTIDE[**SEP**]PSEUDO[**EOS**]<br>[**TCR**]CD[**MASK**]SE[**MASK**][**EOS**]<br>↓<br>[**TCR**]CDR3BSEQ[**EOS**] |

Supplementary Figure 1: *Comparing TCRBART and TCRT5.* A) Transformer architecture juxtaposing BART and T5 encoder and decoder layers, inspired by Figure 1 of (Vaswani et al., 2023). While BART follows the traditional Transformer architecture, T5 introduces the following key changes: removing the additive bias from LayerNorm, moving the LayerNorm before the feature processing blocks, introducing relative attention bias instead of learned positional encoding, and implementing dropout throughout the network. B) Differentiating between sequence representations encountered in the three tasks by each model (BART vs. T5).

### A.1. Quantitative Break Down of Data Constraints on Model Performance

For all metrics, Seq2Seq performance is predicated on a representative number of reference target sequences. Given that we are evaluating generations where even the most represented pMHC has on the order of $\approx$ 10,000 observed sequences out of a theoretical max of $10^6$ (Mason, 1998; Sewell, 2012) (1% of the total diversity), we are in a regime where model performance is the lower bound on performance. This is because we are evaluating the model not only on its ability to generate correct target sequences, but are inadvertently asking it to generate target sequences that resemble those that have been experimentally validated, comletely unrelated to sequence veracity (Figure S2A). We distill this intuition into a probabilistic framework to contextualize the limits on recall-based metrics (i.e. F1 score) for model evaluation, given the amount of data that currently exists:

Assume we have a held-out pMHC ($pMHC_i$) with a theoretical set $C$ of cognate CDR3b sequences, of which a subset of sequences have been experimentally validated (observed). We can model the likelihood of the generated sequences belonging to the observed set using a composite distribution linking the binomial and hypergeometric distributions. Given model that samples $n$ CDR3b sequences conditioned on that pMHC, we can define Z to be an unobservable binomially distributed random variable representing the number of correct (but not necessarily observed) generated CDR3b sequences.

$$\Pr(Z = z; \theta_i) = \binom{n}{z} \theta_i^z (1-\theta_i)^{n-z}$$

where: Z = Random Variable: unobservable number of correct sequences that are in the reference set
n = Number of generated translations
$\theta_i$ =True model accuracy for a given $pMHC_i$

Then we can construct a conditional distribution of number of correct and observed sequences $Y|Z$ according to:

$$\Pr(Y = y|Z = z) = \frac{\binom{K}{y}\binom{N-K}{z-y}}{\binom{N}{z}}$$

where: Y = Random Variable: observed number of correct sequences that have been experimentally validated.
N = Number of total cognate sequences (ground truth, partially observed)
K = Number of experimentally validated cognate sequences
n = Number of generated sequences
z = Sample size (number of correct generated sequences, observed through Y)

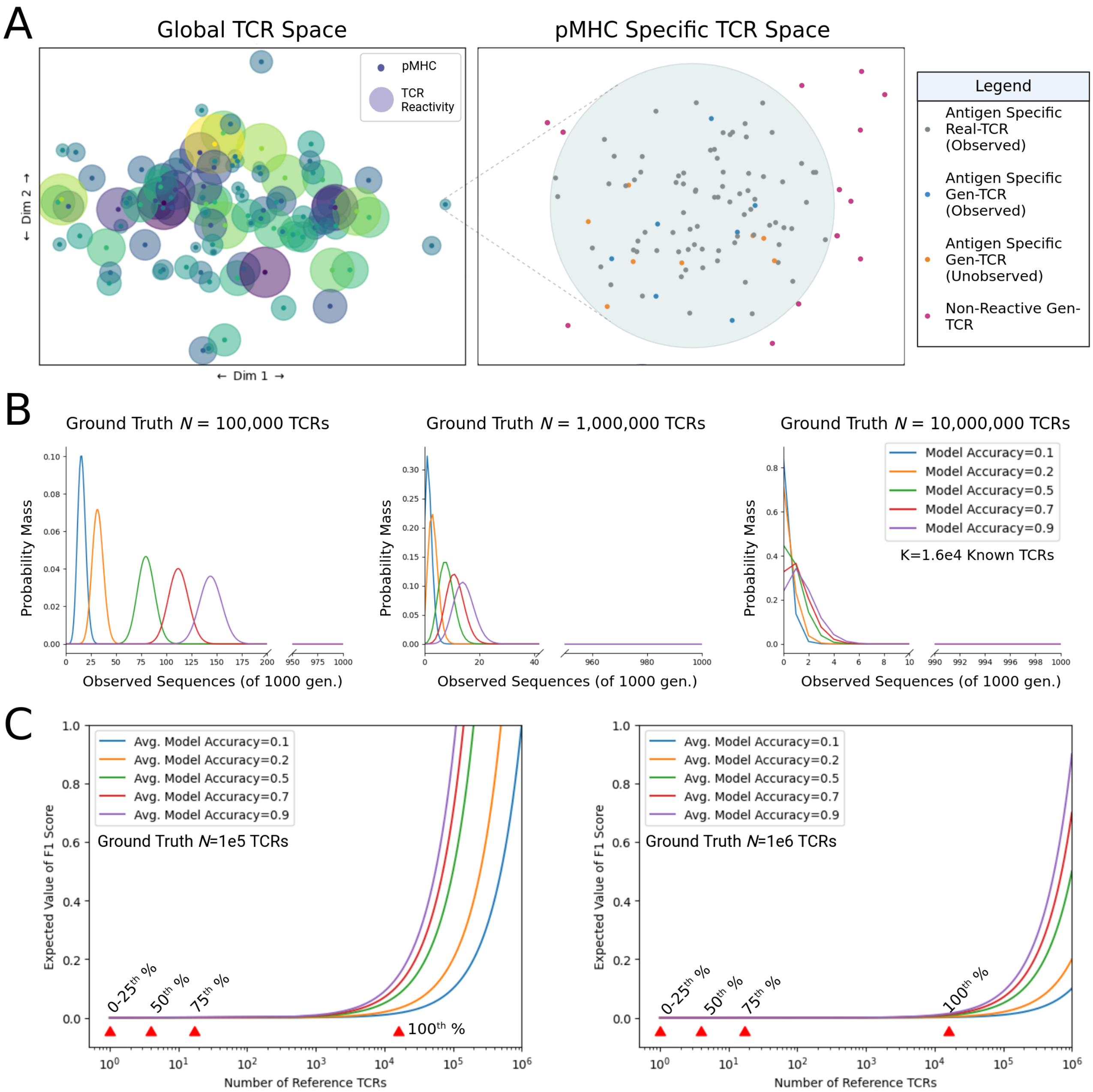
This gives a joint distribution:

$$\begin{aligned} \Pr(Y = y, Z = z) &= Pr(Y = y|Z = z)Pr(Z = z) \\ &= \frac{\binom{K}{y}\binom{N-K}{z-y}}{\binom{N}{y}} \binom{n}{z} \theta_i^z (1-\theta_i)^{n-z} \end{aligned} \tag{4}$$

By marginalizing on Y we get the following equation, whose PMF we plot in Figure S2B:

$$\begin{aligned} \Pr(Y = y) &= \sum_{z}^{n} Pr(Y = y|Z = z)Pr(Z = z) \\ \Pr(Y = y; N, K, n, \theta_i) &= \sum_{z=y}^{n} \frac{\binom{K}{y}\binom{N-K}{z-y}}{\binom{N}{y}} \binom{n}{z} \theta_i^z (1-\theta_i)^{n-z} \end{aligned} \tag{5}$$

We posit that this framework may be useful in characterizing model performance via estimating the parameter $\theta_i$ for pMHCs using Bayesian methods or jointly estimating $Z_i$ and $\theta_i$ through the Expectation Maximization algorithm. We leave this for future exploration and use the above for contextualizing evaluation performance in the current data regime.



Supplementary Figure 2: *Data Constraints on Model Performance*. A) Illustrative diagram showing the global and local TCR context. Global TCR space is represented by 2D projections of various TCR sequences, where antigen-specific TCRs are radially distributed about a pMHC and overlap designates TCR cross reactivity. The local pMHC-specific TCR space shows the distinction between observed TCRs and ground truth TCRs in relation to model generations. B) Probability mass function (PMF) of Equation plotted for different values of ground truth TCRs and different model accuracies. $K$ is fixed to be the current maximum of known TCRs for a given antigen $\approx 16,000$. C) Expected value of the F1 score is plotted for different average model accuracies given $\theta$ (the PMF simplifies to the conditional hypergeometric distribution). Red arrows indicate the percentiles of reference TCR counts ($K$) from the real data.

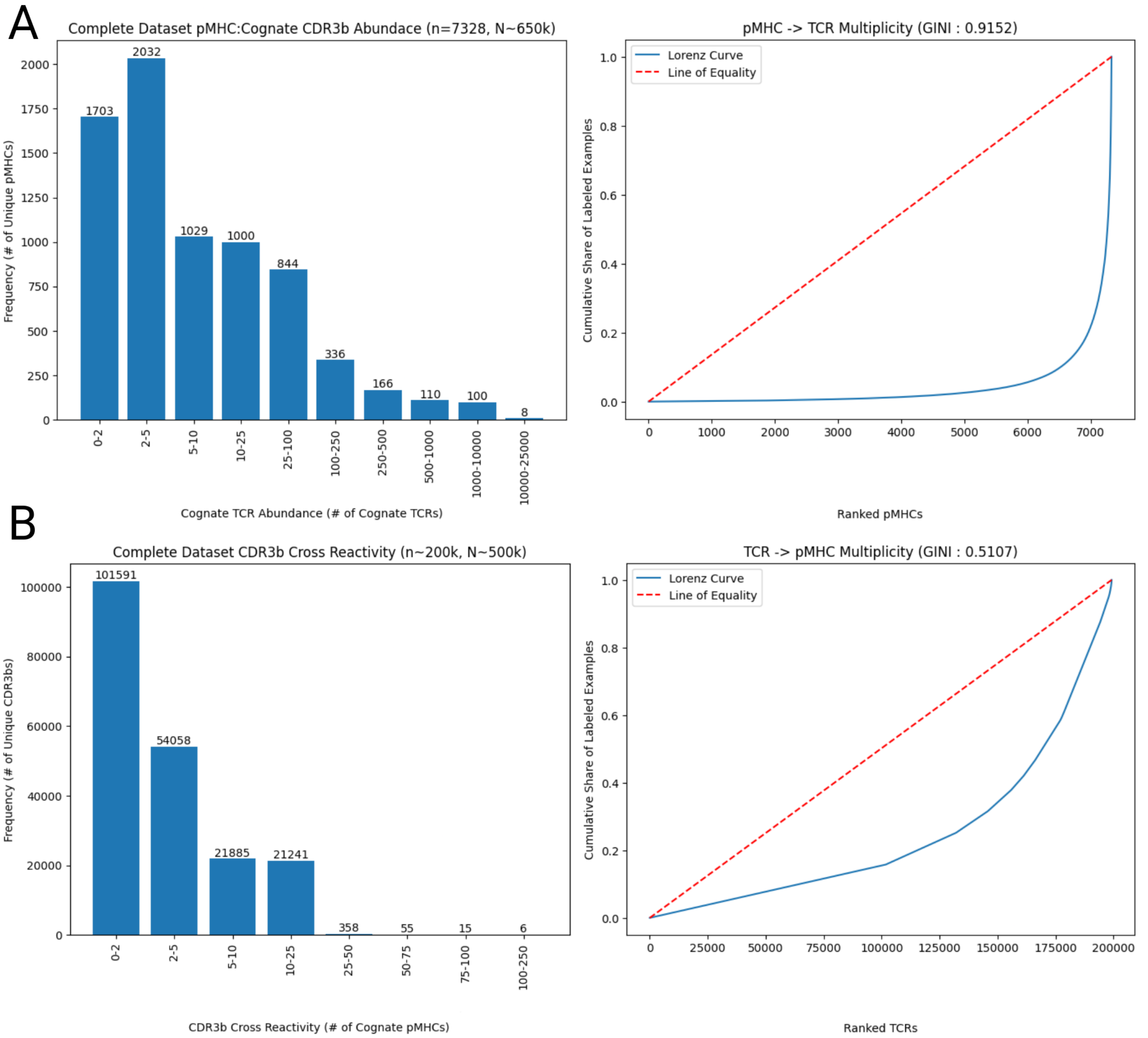
## A.2. Supplementary Methods

### A.2.1. Dataset Construction

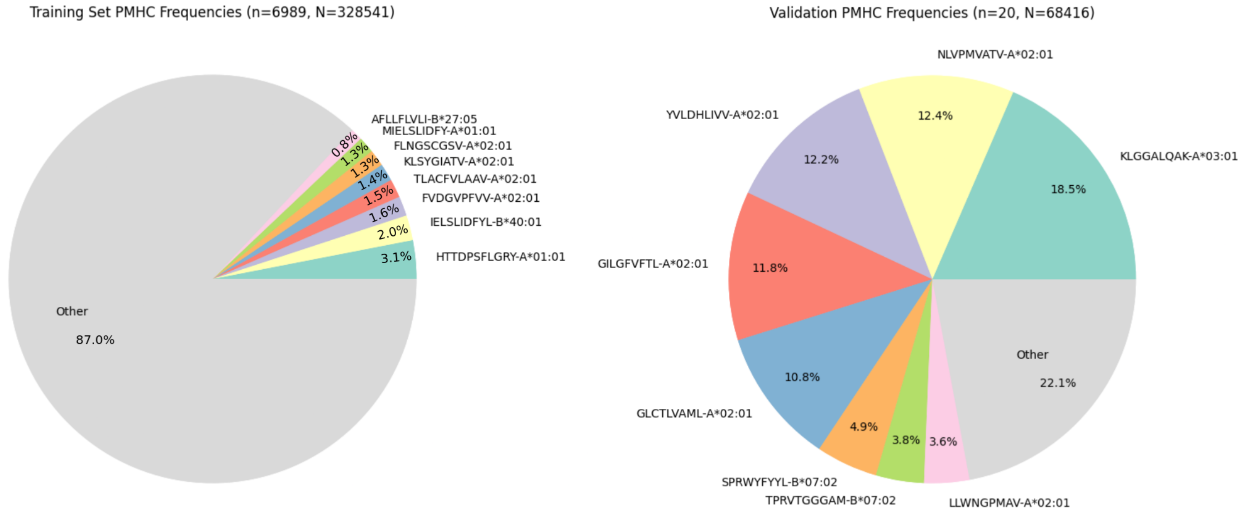
First, to aggregate the data spanning various sources, formats, and nomenclature, we mapped the columns from each individual dataset to a common consensus schema and concatenated the data along the consensus columns. In the interest of data retention, missing values were reasonably imputed according to other information for that data instance. To keep only the cytotoxic (CD8+) T-cells, we filtered the instances wherever the cell-type was provided or where the HLA-Allele was of MHC-class I. In cases where an HLA-haplotype was provided instead of the specific HLA-allele, as was the case for the MIRA data, we used MHCFlurry2.0 (O'Donnell et al., 2020) to predict the best presenting allele for the given epitope among the potential options from the haplotype information. This data augmentation step resulted in a 5-fold expansion of our training data. An key caveat to note is that the additional training examples are all derived from a single disease context (SARS-Cov-2), skewing the training data's distribution. Additionally there is room for slight error given the peptide-MHC assignment in-silico and not validated experimentally. However, given the merits of including examples with thousands of TCRs against an epitope, we argue for its inclusion. Where the granularity of the HLA-information or TR genes was at the serotype level, we inferred the canoncial 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 the pseudosequence 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 (CDR3b), 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.

Finally, given the importance of training LLMs on non-redundant data (Lee et al., 2022) a de-duplication step was performed to consolidate training examples where examples pMHCs that bound to the same CDR3b and shared a k-mer overlap of at least 6 were clustered together and a single representative pair was chosen from each cluster. The allocation of pairs from each cluster was done to balance pMHC representation.



Supplementary Figure 3: *Parallel Dataset Imbalance*. A) pMHC → TCR. Bar chart and corresponding Lorenz curve demonstrating the phenomenon in which a few well characterized pMHCs account for a disproportionately large share of the labeled data. B) TCR → pMHC. Bar chart and corresponding Lorenz curve showing that most TCRs have both a smaller and more equally distributed range of cognate pMHCs, which is unfortunately confounded by a lack of pMHC diversity. This is due in large part experimental reasons, namely the ease in which T-cell stimulation assays in the presence of specific antigen reveal diverse TCRs upon sequencing while stimulation assays against various pMHCs given a specific TCR not as common given their low positivity rates.



Supplementary Figure 4: *Relative pMHC Abundances in De-Duplicated Training and Validation Set*. Relative pMHC abundances for the de-duplicated training set (left) and validation set (right) are shown. The top-8 pMHCs account for 13% and 77.9% of the training and validation set, respectively.

## A.3. Model/Training Hyperparameters

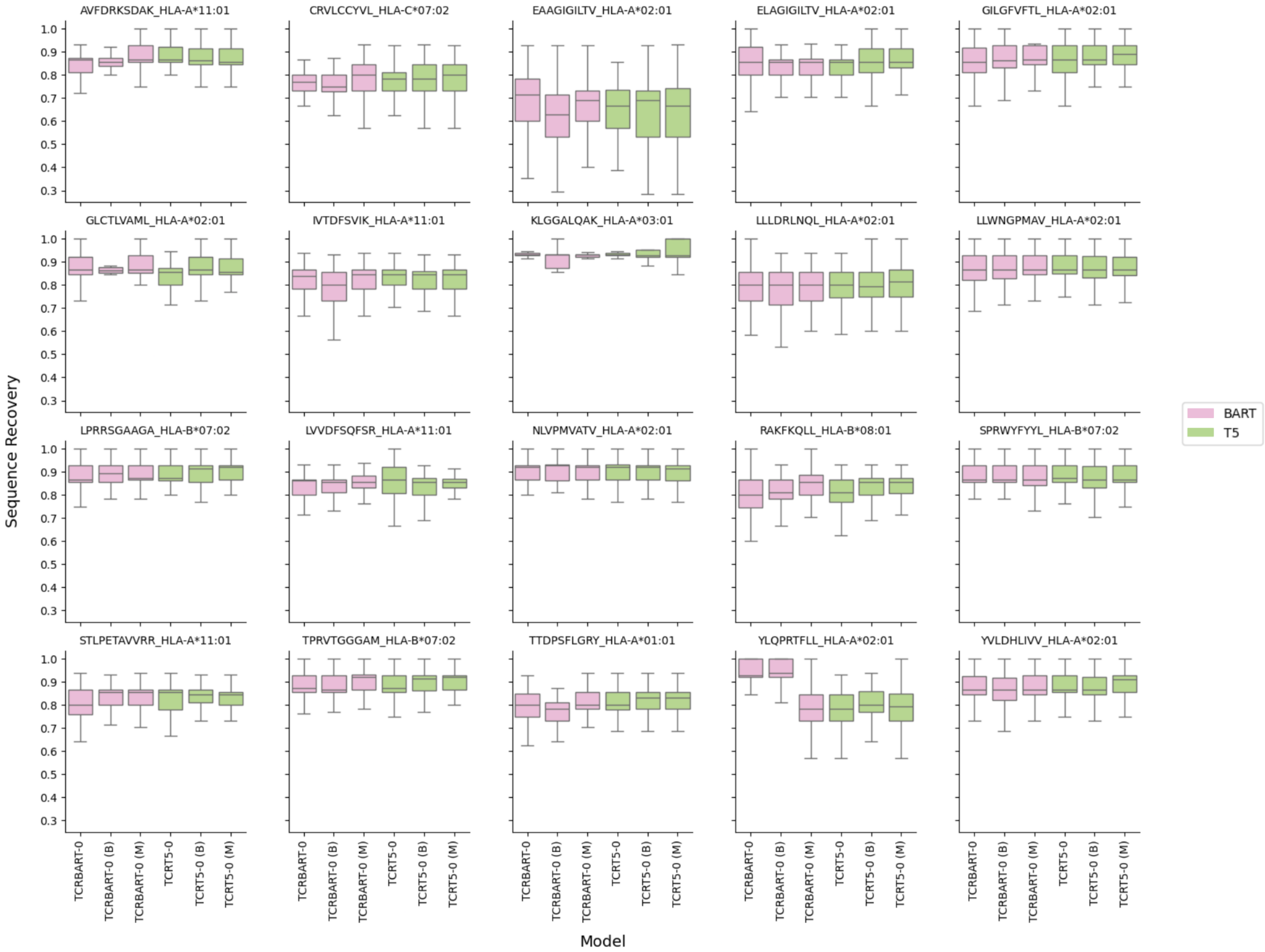
Table 1: Model Architecture Hyperparameters

| | TCRBART | TCRT5 |
|---|---|---|
| Parameters | 46M | 42M |
| $d_{model}$ | 768 | 256 |
| Vocab Size | 28 | 128 |
| Encoder Layers | 6 | 10 |
| Decoder Layers | 6 | 10 |
| Max Position Embedding | 512 | 512 |
| Attention Heads | 16 | 16 |
| Feed Forward Dim | 128 | 1024 |
| Cross Attention | ✓ | ✓ |

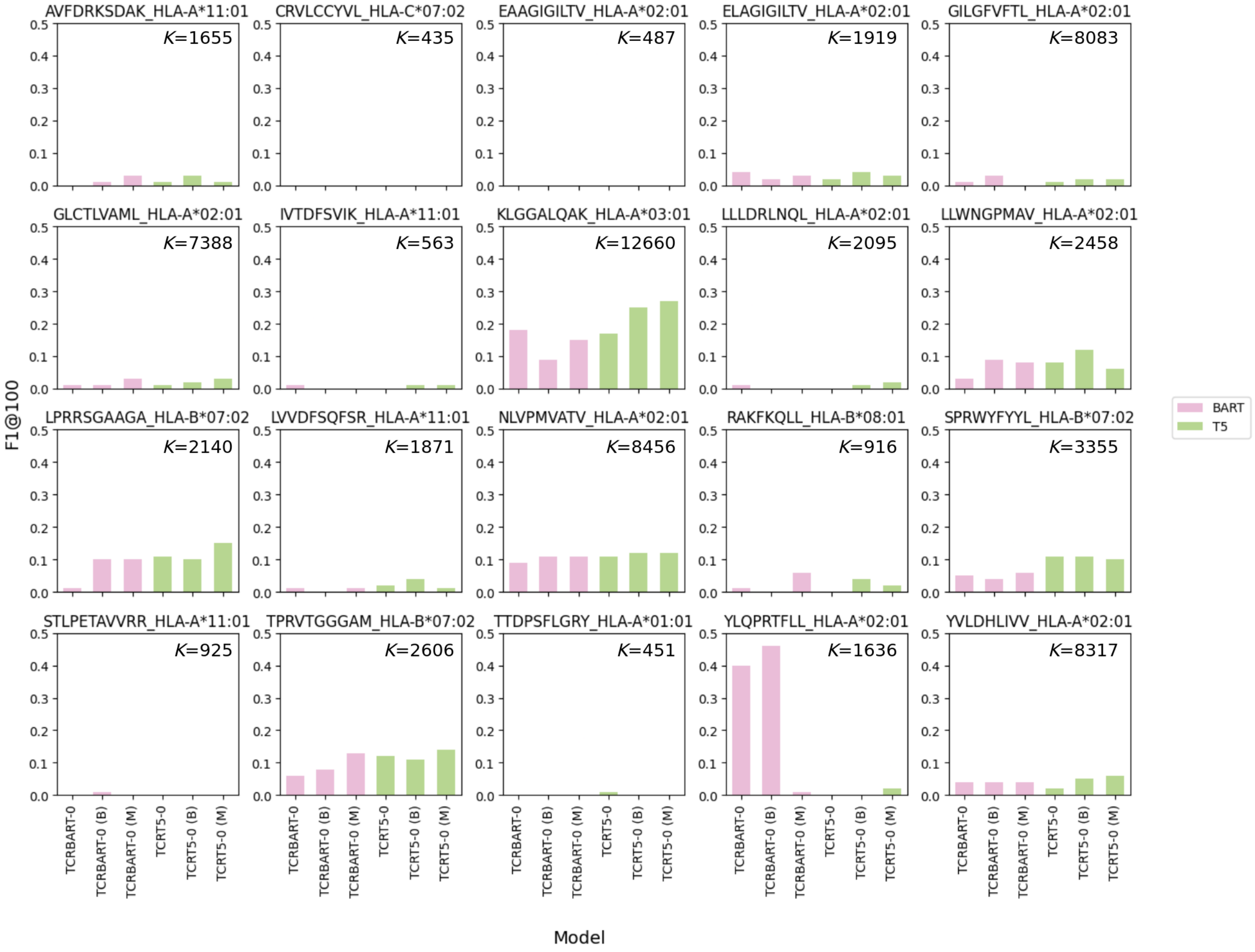
Table 2: Model Training Parameters

| | | BART | T5 |
|---|---|---|---|
| Baseline | Batch Size | 128 | 128 |
| | Learning Rate | 5e-05 | 3e-04 |
| | Weight Decay | 0.01 | 0.001 |
| | Optimizer | AdamW | AdamW |
| Bidirectional Training | Batch Size | 128 | 128 |
| | Learning Rate | 5e-05 | 3e-04 |
| | Weight Decay | 0.01 | 0.001 |
| | Optimizer | AdamW | AdamW |
| Multitask Training | Batch Size | 128 | 128 |
| | Learning Rate | 5e-05 | 3e-04 |
| | Weight Decay | 0.01 | 0.001 |
| | $p_{MLM}$ | 0.15 | 0.15 |
| | Optimizer | AdamW | AdamW |

### A.4. Extended Results



Supplementary Figure 5: *Sequence Recoveries by pMHC*. Box and whisker plots showing the range, median, and quartiles of sequence recoveries per pMHC split by model.



Supplementary Figure 6: *F1@100 Score by pMHC*. Bar plot of F1@100 scores per model shown for each pMHC. The number of reference CDR3b sequences, experimentally validated to bind that pMHC is shown at the top right corner of each subplot as $(K)$.

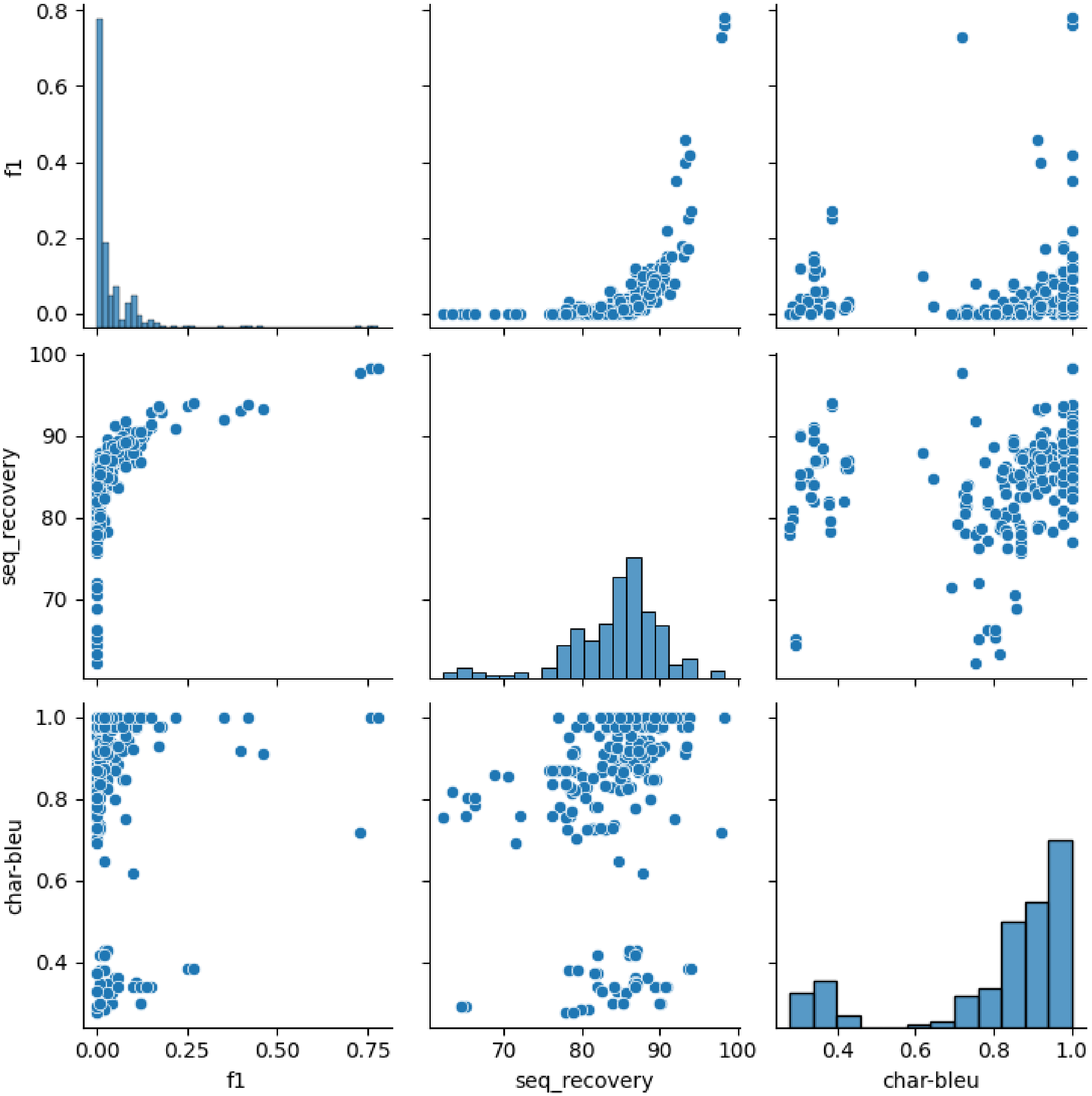
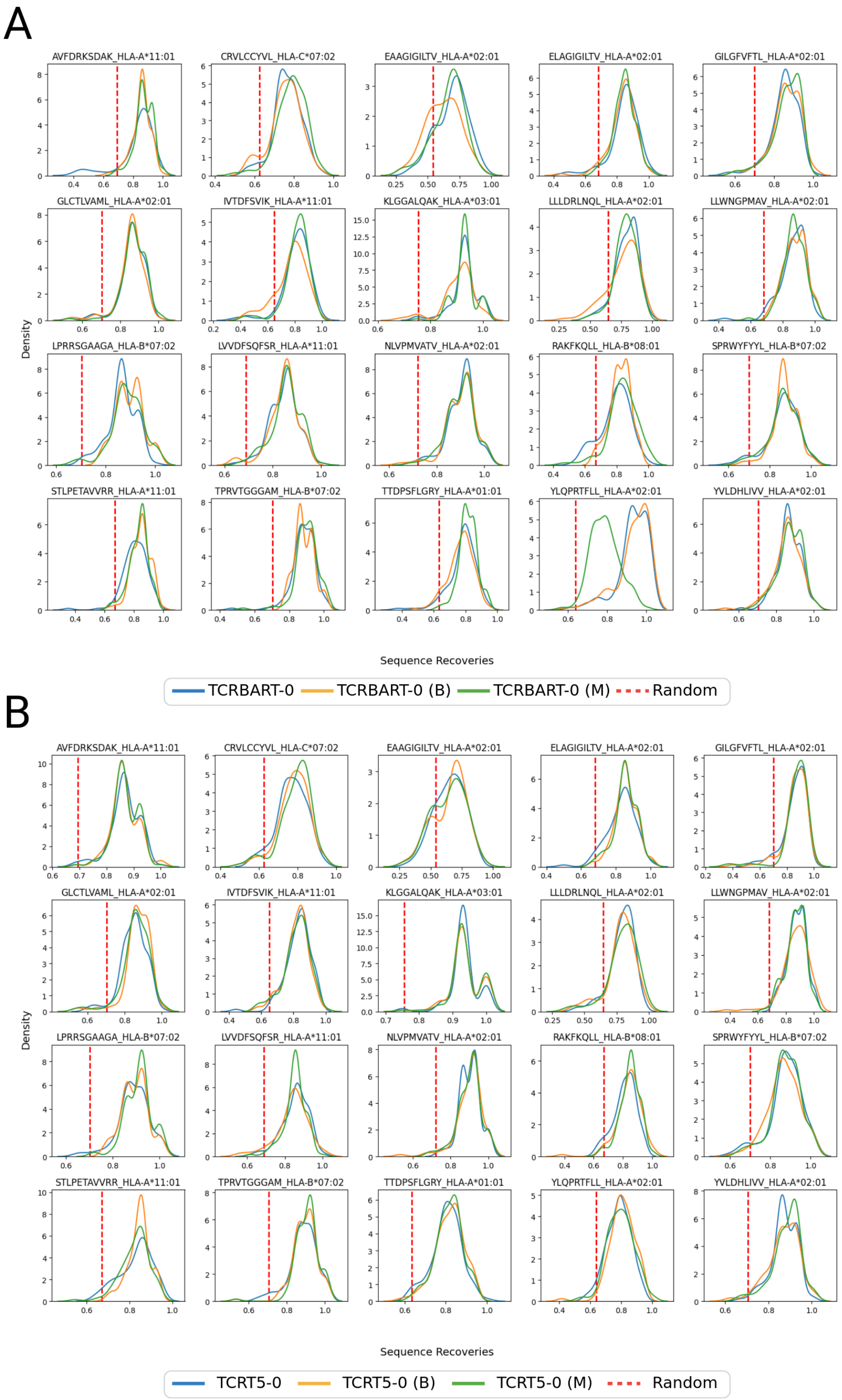
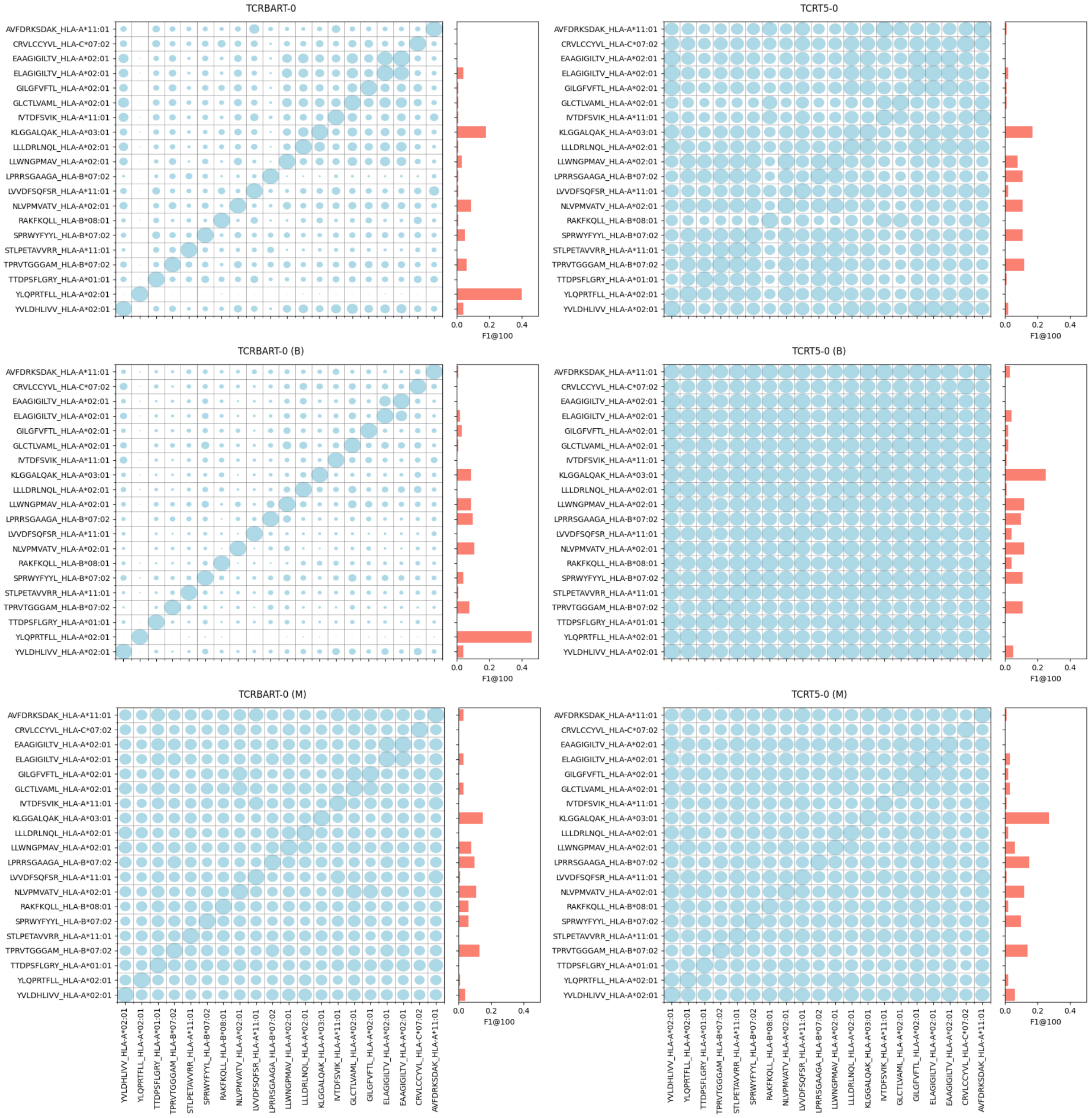


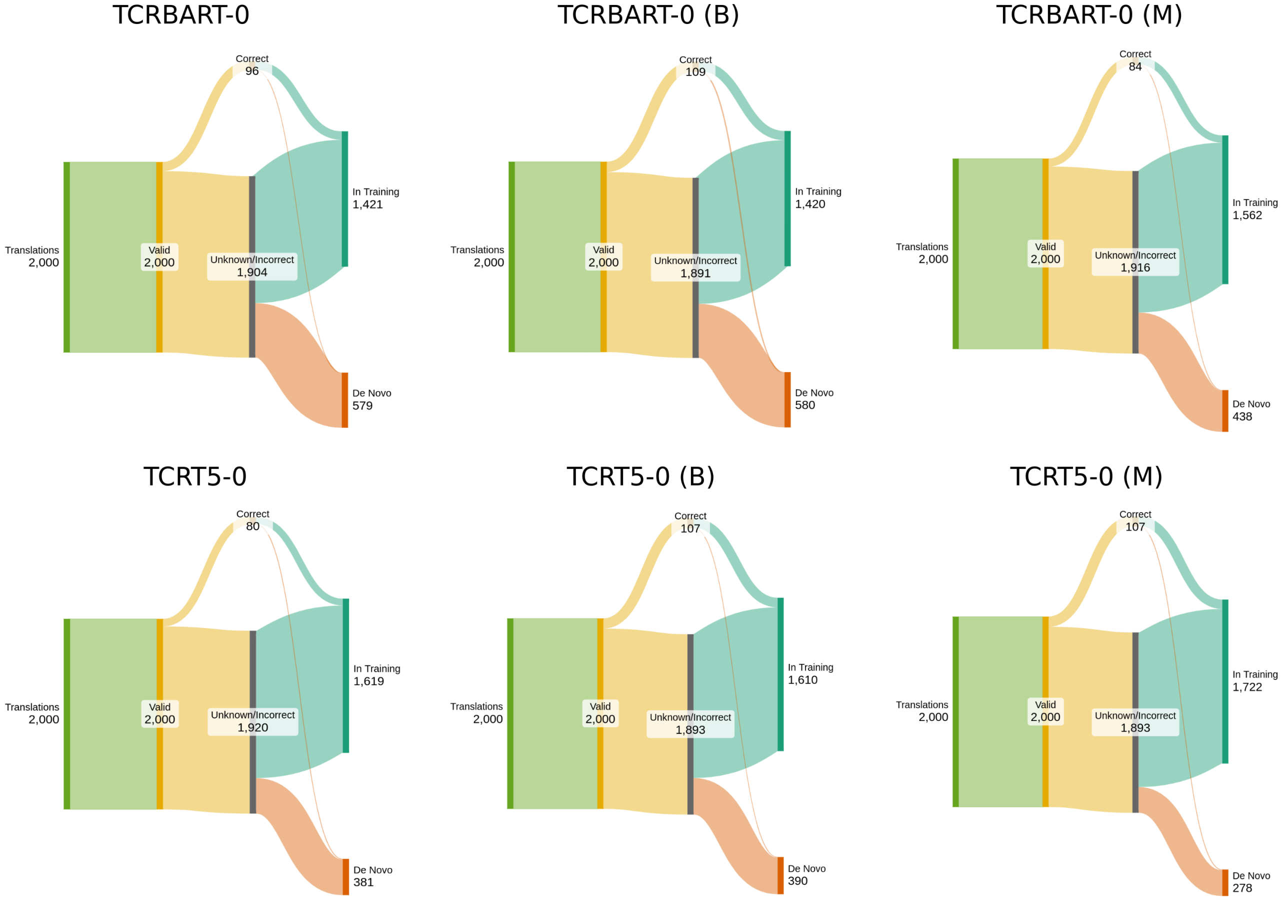
Figure 7: *Evaluation Metric Correlation Plot*. Pairwise scatterplots between performance metrics (CharBLEU, sequence recovery, F1@100) shown for all 20 pMHCs with the baseline and multi-task models (n=80 per plot).



Supplementary Figure 8: *Translation Quality of Model Generations vs. Random CDR3bs*. Density plot showing the sequence recoveries of translations derived from each model and the pMHC's known cognate TCRs vs. a set of 1000 random TCR sequences where known binders were filtered out. Simulation was ran 1000 iterations and the mean sequence recovery derived from random sequences is plotted as a red line.



Supplementary Figure 9: *Sequence Diversity and Accuracy of Generations*. Circle heatmap plot showing the pairwise overlap between generations conditioned on $pMHC_i$ and $pMHC_j$ as the size of the circle, colored by the row-wise F1@100 score of that generations conditioned on that pMHC.



Supplementary Figure 10: *Characterizing Validity, Accuracy, and Novelty of Generated Sequences*. Sankey plots showing the validity of translations (determined by OLGA (Sethna et al., 2019) $P_{gen} > 0$), accuracy (determined by sampling exact cognate TCRs) and novelty (membership in the training set) for each model.

Table 3: Characterization of Train/Test Target Overlap

| Test Peptide | Closest Train Peptide(s) | Edit Distance | CDR3b 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 |

Table 4: Individual Top-20 pMHC Performance Metrics

| | | Evaluation Metrics | | | | |
|---|---|---|---|---|---|---|
| | Model | Char-BLEU | P@100 | R@100 | F1@100 | % Recovery |
| AVFDRKSDAK-A*11:01 | TCRBART-0 | 0.865 | 0.00 | 0.00 | 0.00 | 82.6 |
| | TCRBART-0 (B) | 0.865 | 0.01 | 0.01 | 0.01 | 85.4 |
| | TCRBART-0 (M) | 0.865 | 0.03 | 0.03 | 0.03 | 87.3 |
| | TCRT5-0 | 0.777 | 0.01 | 0.01 | 0.01 | 86.8 |
| | TCRT5-0 (B) | 0.347 | 0.03 | 0.03 | 0.03 | 86.7 |
| | TCRT5-0 (M) | 0.347 | 0.01 | 0.01 | 0.01 | 86.9 |
| CRVLCCYVL-C*07:02 | TCRBART-0 | 0.869 | 0.00 | 0.00 | 0.00 | 76.7 |
| | TCRBART-0 (B) | 0.869 | 0.00 | 0.00 | 0.00 | 75.7 |
| | TCRBART-0 (M) | 0.869 | 0.00 | 0.00 | 0.00 | 78.7 |
| | TCRT5-0 | 0.869 | 0.00 | 0.00 | 0.00 | 76.9 |
| | TCRT5-0 (B) | 0.277 | 0.00 | 0.00 | 0.00 | 78.0 |
| | TCRT5-0 (M) | 0.277 | 0.00 | 0.00 | 0.00 | 78.9 |
| EAAGIGILTV-A*02:01 | TCRBART-0 | 0.859 | 0.00 | 0.00 | 0.00 | 68.9 |
| | TCRBART-0 (B) | 0.754 | 0.00 | 0.00 | 0.00 | 62.3 |
| | TCRBART-0 (M) | 0.784 | 0.00 | 0.00 | 0.00 | 66.2 |
| | TCRT5-0 | 0.760 | 0.00 | 0.00 | 0.00 | 65.1 |
| | TCRT5-0 (B) | 0.292 | 0.00 | 0.00 | 0.00 | 65.1 |
| | TCRT5-0 (M) | 0.292 | 0.00 | 0.00 | 0.00 | 64.5 |
| ELAGIGILTV-A*02:01 | TCRBART-0 | 1.000 | 0.04 | 0.04 | 0.04 | 84.6 |
| | TCRBART-0 (B) | 0.912 | 0.02 | 0.02 | 0.02 | 83.6 |
| | TCRBART-0 (M) | 0.951 | 0.03 | 0.03 | 0.03 | 84.6 |
| | TCRT5-0 | 0.974 | 0.02 | 0.02 | 0.02 | 83.2 |
| | TCRT5-0 (B) | 0.324 | 0.04 | 0.04 | 0.04 | 85.6 |
| | TCRT5-0 (M) | 0.324 | 0.03 | 0.03 | 0.03 | 85.6 |
| GILGFVFTL-A*02:01 | TCRBART-0 | 1.000 | 0.01 | 0.01 | 0.01 | 85.3 |
| | TCRBART-0 (B) | 0.912 | 0.03 | 0.03 | 0.03 | 86.2 |
| | TCRBART-0 (M) | 1.000 | 0.00 | 0.00 | 0.00 | 86.5 |
| | TCRT5-0 | 0.912 | 0.01 | 0.01 | 0.01 | 85.8 |
| | TCRT5-0 (B) | 0.420 | 0.02 | 0.02 | 0.02 | 86.1 |
| | TCRT5-0 (M) | 0.420 | 0.02 | 0.02 | 0.02 | 86.8 |
| GLCTLVAML-A*02:01 | TCRBART-0 | 1.000 | 0.01 | 0.01 | 0.01 | 86.3 |
| | TCRBART-0 (B) | 1.000 | 0.01 | 0.01 | 0.01 | 85.5 |
| | TCRBART-0 (M) | 1.000 | 0.03 | 0.03 | 0.03 | 87.1 |
| | TCRT5-0 | 0.851 | 0.01 | 0.01 | 0.01 | 84.9 |
| | TCRT5-0 (B) | 0.429 | 0.02 | 0.02 | 0.02 | 87.0 |
| | TCRT5-0 (M) | 0.429 | 0.03 | 0.03 | 0.03 | 86.0 |
| IVTDFSVIK-A*11:01 | TCRBART-0 | 0.731 | 0.01 | 0.01 | 0.01 | 81.4 |
| | TCRBART-0 (B) | 0.783 | 0.00 | 0.00 | 0.00 | 77.2 |
| | TCRBART-0 (M) | 0.783 | 0.00 | 0.00 | 0.00 | 81.6 |
| | TCRT5-0 | 0.831 | 0.00 | 0.00 | 0.00 | 83.0 |
| | TCRT5-0 (B) | 0.339 | 0.01 | 0.01 | 0.01 | 81.9 |
| | TCRT5-0 (M) | 0.417 | 0.01 | 0.01 | 0.01 | 82.1 |
| KLGGALQAK-A*03:01 | TCRBART-0 | 0.976 | 0.18 | 0.18 | 0.18 | 92.9 |
| | TCRBART-0 (B) | 0.931 | 0.09 | 0.09 | 0.09 | 90.6 |
| | TCRBART-0 (M) | 1.000 | 0.15 | 0.15 | 0.15 | 92.9 |

Table 4: (continued)

| | | Evaluation Metrics | | | | |
|---|---|---|---|---|---|---|
| | Model | Char-BLEU | P@100 | R@100 | F1@100 | % Recovery |
| | TCRT5-0 | 0.931 | 0.17 | 0.17 | 0.17 | 93.4 |
| | TCRT5-0 (B) | 0.384 | 0.25 | 0.25 | 0.25 | 93.7 |
| | TCRT5-0 (M) | 0.384 | 0.27 | 0.27 | 0.27 | 94.0 |
| LLLDRLNQL-A*02:01 | TCRBART-0 | 0.821 | 0.01 | 0.01 | 0.01 | 79.3 |
| | TCRBART-0 (B) | 0.836 | 0.00 | 0.00 | 0.00 | 76.2 |
| | TCRBART-0 (M) | 0.821 | 0.00 | 0.00 | 0.00 | 78.5 |
| | TCRT5-0 | 0.821 | 0.00 | 0.00 | 0.00 | 79.0 |
| | TCRT5-0 (B) | 0.379 | 0.01 | 0.01 | 0.01 | 78.2 |
| | TCRT5-0 (M) | 0.379 | 0.02 | 0.02 | 0.02 | 79.5 |
| LLWNGPMAV-A*02:01 | TCRBART-0 | 0.976 | 0.03 | 0.03 | 0.03 | 86.5 |
| | TCRBART-0 (B) | 0.946 | 0.09 | 0.09 | 0.09 | 87.7 |
| | TCRBART-0 (M) | 0.976 | 0.08 | 0.08 | 0.08 | 87.3 |
| | TCRT5-0 | 0.976 | 0.08 | 0.08 | 0.08 | 87.8 |
| | TCRT5-0 (B) | 0.341 | 0.12 | 0.12 | 0.12 | 86.9 |
| | TCRT5-0 (M) | 0.341 | 0.06 | 0.06 | 0.06 | 86.9 |
| LPRRSGAAGA-B*07:02 | TCRBART-0 | 1.000 | 0.01 | 0.01 | 0.01 | 86.5 |
| | TCRBART-0 (B) | 1.000 | 0.1 | 0.1 | 0.1 | 89.5 |
| | TCRBART-0 (M) | 1.000 | 0.1 | 0.1 | 0.1 | 88.8 |
| | TCRT5-0 | 1.000 | 0.11 | 0.11 | 0.11 | 88.9 |
| | TCRT5-0 (B) | 0.339 | 0.1 | 0.1 | 0.1 | 89.4 |
| | TCRT5-0 (M) | 0.339 | 0.15 | 0.15 | 0.15 | 91.0 |
| LVVDFSQFSR-A*11:01 | TCRBART-0 | 0.895 | 0.01 | 0.01 | 0.01 | 84.6 |
| | TCRBART-0 (B) | 0.955 | 0.00 | 0.00 | 0.00 | 84.5 |
| | TCRBART-0 (M) | 0.895 | 0.01 | 0.01 | 0.01 | 85.5 |
| | TCRT5-0 | 0.831 | 0.02 | 0.02 | 0.02 | 85.6 |
| | TCRT5-0 (B) | 0.301 | 0.04 | 0.04 | 0.04 | 84.0 |
| | TCRT5-0 (M) | 0.301 | 0.01 | 0.01 | 0.01 | 85.3 |
| NLVPMVATV-A*02:01 | TCRBART-0 | 1.000 | 0.09 | 0.09 | 0.09 | 90.1 |
| | TCRBART-0 (B) | 1.000 | 0.11 | 0.11 | 0.11 | 90.1 |
| | TCRBART-0 (M) | 0.974 | 0.11 | 0.11 | 0.11 | 89.9 |
| | TCRT5-0 | 0.976 | 0.11 | 0.11 | 0.11 | 90.2 |
| | TCRT5-0 (B) | 0.301 | 0.12 | 0.12 | 0.12 | 90.1 |
| | TCRT5-0 (M) | 0.301 | 0.12 | 0.12 | 0.12 | 89.9 |
| RAKFKQLL-B*08:01 | TCRBART-0 | 0.912 | 0.01 | 0.01 | 0.01 | 79.1 |
| | TCRBART-0 (B) | 0.955 | 0.00 | 0.00 | 0.00 | 82.2 |
| | TCRBART-0 (M) | 0.912 | 0.06 | 0.06 | 0.06 | 83.8 |
| | TCRT5-0 | 0.851 | 0.00 | 0.00 | 0.00 | 81.3 |
| | TCRT5-0 (B) | 0.339 | 0.04 | 0.04 | 0.04 | 84.1 |
| | TCRT5-0 (M) | 0.646 | 0.02 | 0.02 | 0.02 | 84.7 |
| SPRWYFYYL-B*07:02 | TCRBART-0 | 0.886 | 0.05 | 0.05 | 0.05 | 86.5 |
| | TCRBART-0 (B) | 0.938 | 0.04 | 0.04 | 0.04 | 87.4 |
| | TCRBART-0 (M) | 0.886 | 0.06 | 0.06 | 0.06 | 86.6 |
| | TCRT5-0 | 0.976 | 0.11 | 0.11 | 0.11 | 87.7 |
| | TCRT5-0 (B) | 0.353 | 0.11 | 0.11 | 0.11 | 86.9 |
| | TCRT5-0 (M) | 0.619 | 0.1 | 0.1 | 0.1 | 87.9 |

Table 4: (continued)

| | | Evaluation Metrics | | | | |
|---|---|---|---|---|---|---|
| | Model | Char-BLEU | P@100 | R@100 | F1@100 | % Recovery |
| STLPETAVVRR-A*11:01 | TCRBART-0 | 0.831 | 0.00 | 0.00 | 0.00 | 80.7 |
| | TCRBART-0 (B) | 0.904 | 0.01 | 0.01 | 0.01 | 84.6 |
| | TCRBART-0 (M) | 0.912 | 0.00 | 0.00 | 0.00 | 83.2 |
| | TCRT5-0 | 0.912 | 0.00 | 0.00 | 0.00 | 82.8 |
| | TCRT5-0 (B) | 0.330 | 0.00 | 0.00 | 0.00 | 84.5 |
| | TCRT5-0 (M) | 0.330 | 0.00 | 0.00 | 0.00 | 82.6 |
| TPRVTGGGAM-B*07:02 | TCRBART-0 | 1.000 | 0.06 | 0.06 | 0.06 | 88.3 |
| | TCRBART-0 (B) | 0.946 | 0.08 | 0.08 | 0.08 | 88.6 |
| | TCRBART-0 (M) | 1.000 | 0.13 | 0.13 | 0.13 | 90.0 |
| | TCRT5-0 | 1.000 | 0.12 | 0.12 | 0.12 | 88.8 |
| | TCRT5-0 (B) | 0.339 | 0.11 | 0.11 | 0.11 | 89.5 |
| | TCRT5-0 (M) | 0.339 | 0.14 | 0.14 | 0.14 | 90.6 |
| TTDPSFLGRY-A*01:01 | TCRBART-0 | 0.727 | 0.00 | 0.00 | 0.00 | 78.1 |
| | TCRBART-0 (B) | 1.000 | 0.00 | 0.00 | 0.00 | 77.0 |
| | TCRBART-0 (M) | 0.727 | 0.00 | 0.00 | 0.00 | 81.7 |
| | TCRT5-0 | 0.727 | 0.01 | 0.01 | 0.01 | 80.6 |
| | TCRT5-0 (B) | 0.375 | 0.00 | 0.00 | 0.00 | 81.9 |
| | TCRT5-0 (M) | 0.375 | 0.00 | 0.00 | 0.00 | 81.7 |
| YLQPRTFLL-A*02:01 | TCRBART-0 | 0.919 | 0.4 | 0.4 | 0.4 | 93.2 |
| | TCRBART-0 (B) | 0.912 | 0.46 | 0.46 | 0.46 | 93.3 |
| | TCRBART-0 (M) | 0.919 | 0.01 | 0.01 | 0.01 | 79.0 |
| | TCRT5-0 | 0.831 | 0.00 | 0.00 | 0.00 | 78.5 |
| | TCRT5-0 (B) | 0.285 | 0.00 | 0.00 | 0.00 | 80.9 |
| | TCRT5-0 (M) | 0.285 | 0.02 | 0.02 | 0.02 | 79.8 |
| YVLDHLIVV-A*02:01 | TCRBART-0 | 0.931 | 0.04 | 0.04 | 0.04 | 86.7 |
| | TCRBART-0 (B) | 0.912 | 0.04 | 0.04 | 0.04 | 85.7 |
| | TCRBART-0 (M) | 0.931 | 0.04 | 0.04 | 0.04 | 87.3 |
| | TCRT5-0 | 0.874 | 0.02 | 0.02 | 0.02 | 87.2 |
| | TCRT5-0 (B) | 0.361 | 0.05 | 0.05 | 0.05 | 86.9 |
| | TCRT5-0 (M) | 0.361 | 0.06 | 0.06 | 0.06 | 88.4 |