Conditional Generation of Antigen Specific T-cell Receptor Sequences

Dhuvarakesh Karthikeyan [*,†]
Curriculum of Bioinformatics and Computational Biology
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 Microbiology and Immunology
University of North Carolina, Chapel Hill
benjamin_vincent@med.unc.edu

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

Abstract

Training and evaluating large language models (LLMs) for use in the design of antigen specific T-cell receptor (TCR) sequences is challenging due to the complex many-to-many mapping between TCRs and their targets, a struggle exacerbated by a severe lack of ground truth data. Traditional NLP metrics can be artificially poor indicators of model performance since labels are concentrated on a few examples, and functional in-vitro assessment of generated TCRs is time-consuming and costly. Here, we introduce TCR-BART and TCR-T5, adapted from the prominent BART and T5 models, to explore the use of these LLMs for conditional TCR sequence generation given a specific target epitope. To fairly evaluate such models with limited labeled examples, we propose novel evaluation metrics tailored to the sparsely sampled many-to-many nature of TCR-epitope data and investigate the interplay between accuracy and diversity of generated TCR sequences.

1 Introduction

T-cells are specialized immune cells responsible for clearing infections, suppressing cancer, and preventing autoimmunity through the specific recognition of peptide-MHC (pMHC) complexes by cognate T-cell receptors (TCRs). Paradoxically, individual TCRs can recognize on the order of 10^6 unique peptides, with pMHCs recognized by a similar number of unique TCRs [1, 2]. However, this many-to-many mapping is sparsely sampled with many experimentally validated TCRs studied in the context of a few pMHCs [3]. Computational models capable of designing antigen specific TCRs, while accounting for epitope cross-reactivity, have the potential to not only drastically accelerate the development of targeted cellular therapies both for cytotoxic [4–7] and tolerogenic uses [8], but also provide foundational insights into the broader mechanisms governing immunogenicity.

Current approaches in modeling antigen specificity of TCRs have predominantly relied on framing TCR-pMHC cross reactivity as a binary classification task [9–21], with limited utility in TCR design [22]. Prior work exploring generative models of TCRs leveraged auto-encoders to generate realistic de-novo TCRs that recapitulated repertoire level phenomena when aggregated [23, 24]. However,

[*]Personalized Immunotherapy Research Lab
[†]Computational Medicine Program, UNC School of Medicine
[‡]Vector Institute for Artificial Intelligence
[§]Lineberger Comprehensive Cancer Center

NeurIPS 2023 Generative AI and Biology Workshop.

their use in generating epitope-specific TCR sequences is constrained to cases of known epitopes having associated paired cognate TCR data.

This work explores the formulation of the TCR design problem as a sequence-to-sequence (seq2seq) task for the conditional generation of antigen-specific TCRs (Figure 1). Encoder:Decoder models have proven highly successful in various seq2seq tasks, including those with complex mappings between source and target sequences, such as machine translation, question answering, and text summarization. The transformer architecture introduced in 2017 [25] currently sets the state-of-the-art across these tasks [26–28]. Building upon the demonstrated robustness of this framework, we investigate this model class's capacity to learn a meaningful mapping between pMHC sequences and their potential cognate TCR sequences, while addressing the limitations in current seq2seq metrics. In this paper, we demonstrate the transformer's robustness to the challenges of this task and measure performance in a holistic manner, focusing on accuracy, generalization, and diversity in sequence outputs. Our results provide a first step towards characterizing the performance of generative models in the high multiplicity and low data regime of generating high-fidelity TCRs for target antigens.

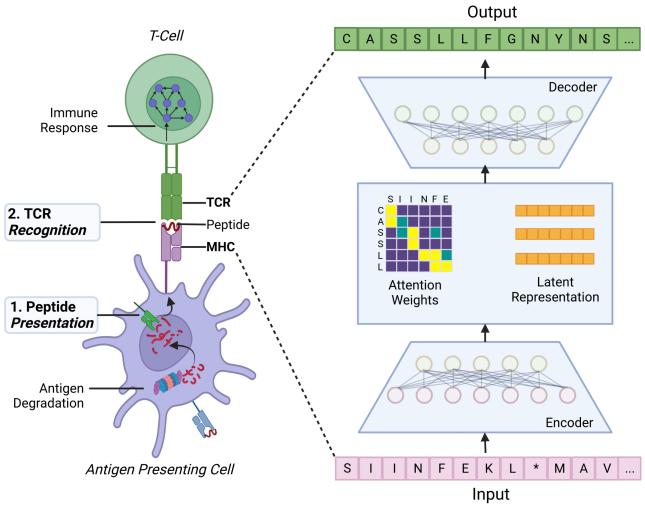


Figure 1: Casting the TCR:pMHC problem as a sequence-to-sequence task.

2 Methodology

2.1 Problem Formulation

We present the TCR:pMHC interaction as a seq2seq given pairs of interacting "source" pMHCs and "target" TCR amino acid sequences. Currently, we represent the TCR by the specific loop that makes the most contact with the epitope known as the complementarity determining region 3 (CDR3b). The pMHC is represented as the concatenation of the peptide sequence and MHC pseudo-sequence defined in [29]. For tokenization, we use a character-level single amino acid tokenization scheme with special tokens to denote sequence start/end and padding. Furthermore, we introduced a separator token to distinguish between the concatenated peptide and MHC sequences: <SOS>PEPTIDE<SEP>PSEUDO<EOS> $\rightarrow$ <SOS>CDR3B<EOS>.

2.2 Data

Our source of ground truth sequences consisted of experimentally validated immunogenic TCR:pMHC pairs, sourced from publicly available datasets: McPAS [30], VDJdb [31], and IEDB [32]. We pre-processed these datasets and collated them, removing duplicates and inferring missing fields where applicable to maximize data retention and completeness. Specifically, CD4/CD8 type was imputed based on MHC class and HLA subgroup was imputed using the canonical subgroup '*01' where unavailable. All HLA nomenclature was normalized using the *mhcgnomes* python package. After, we applied some post-processing steps to filter for only human TCRs, CD8 T-cells,

and entries that included the minimal paired information: (epitope, MHC allele, CDR3b, Vb, Jb).The resulting dataset of over 100,000 paired sequences was partitioned into distinct training (N=1005 unique pMHCs) and test data (N=245 unique pMHCs). This partitioning was designed to balance MHC allele fractions across the two sets and to ensure that validation pMHCs were not exposed during the training phase. In addition to the paired experimental data, we used unlabeled pMHC (N≈400K)[32] and CDR3b sequences (N≈13M) from TCRdb [33] for pre-training.

2.3 Models

For this exploration, we adapted the BART [34] and T5 [28] models, chosen for their demonstrated performance on robust benchmarks spanning various seq2seq tasks in the broad NLP setting [35, 36]. Both TCR-BART and TCR-T5 are specific implementations of denoising autoencoders coupled with autoregressive decoders, whose architectures do not diverge from their inspirations (See Table S1. Additional details may be found in the original papers). TCR-BART was pre-trained via masked amino-acid modeling on a corpus containing both pMHCs and TCRs (30/70 split) whereas TCR-T5 was pre-trained on masked span reconstruction, similar to the original T5 paper. Interestingly, we found that a larger finetuning learning rate and batch sizes yielded better performance, deviating from the relative hyper-parameters used in the original BART and T5 papers (See Table S2). Both models use the same character-level tokenization scheme with minimal adaptations. To assess their performance within the broader encoder:decoder landscape, we established baselines with a bidirectional-RNN, bi-RNN with a cross attention mechanism, and 1-D CNN with cross attention and positional encoding. TCR-BART architecture used 6 encoder and decoder layers and $d_{model} = 768$, totaling around 120 million parameters. In order to maintain equal scales for a fair comparison, we adjusted the embedding dimensions and layer counts in the baseline models to achieve 120M parameters $\pm$ 10M. Here, TCR-T5 (120M) did not converge in the same number of iterations as the other models and thus a smaller TCR-T5 (66M) with the same d_{model} and number of encoder and decoder layers as TCR-BART was used. All models were trained using the cross entropy loss.

2.4 Evaluation

Given the shortage of experimental data and the high cost of performing additional in-vitro validation, evaluating generations of TCR sequences poses a unique challenge, especially in the case of *de novo* sequences. A recent study used an existing binary predictor of TCR and pMHC reactivity to measure their generative model's capacity to produce realistic and antigen-specific TCRs [37]. While this approach has the ability of evaluating *de novo* TCRs, given that these discriminator models generalize poorly on out of distribution TCRs [38], we opt for metrics with a less variable error profile at the cost of potentially underrepresenting performance. However, traditional recall-based metrics prove inadequate and their results, in some instances, can be misleading. When evaluating generated TCRs against a set of a few reference sequences (n < 10), recall based metrics test recapitulation of an arbitrary sample of validated examples among the vast space of all cognate TCRs. Additionally, when evaluating against a set of many TCRs (n > 1000), a model that adequately learns a particular TCR motif [2] may be penalized for lack of diversity. To address this, we employ a combination of tailored metrics to evaluate both generation accuracy and diversity:

- **Char-BLEU**: Calculated using the standard BLEU-4 [39], the character-level BLEU calculates the weighted n-gram precision against the k closest reference sequences to abate unintended penalization of accurate predictions under a large reference set. We set $k = 20$.
- **Precision, Recall, and F1@K**: Borrowed from information retrieval, these metrics gauge precision, recall, and F1 by exact sequence recovery after sampling K times, without rank, given uncertain calibration of model scores and lack of a robust relevance function.
- **Mean Edit Distance**: For each model prediction, the closest match is found and the Levenshtein edit distance is computed and averaged for all predictions.
- **Perplexity**: We report perplexity as a standard measure of language model performance, using the cross entropy loss calculated over the validation corpus.
- **Biological Likelihood**: As an orthogonal measure of model performance, independent of antigen-specificity or labeled data, we compute generation probability of predictions using OLGA, a domain specific generative model that infers TCR sequence likelihood [40].

3 Results

In this section we evaluate the models' generations against known reactive TCRs. We first benchmark across the various architectures, controlling for number of parameters and training steps. All values represent averages across different held out pMHCs. In addition, we curated a "target-rich dataset" comprising the top 10 held-out validation pMHCs with the highest number of cognate TCRs, to observe the effects of evaluating performance with adequately sampled target sequences.

Our results demonstrate the clear effect that evaluating on the target-rich dataset holds with increased performance reported on all metrics (Table 1). However, the relative performance of various models under certain metrics and the discordance between them were less monolithic. While all of the models produced plausible generations that bear semblance to the ground truth sequences (Figure S1), in terms of exact sequence recovery, we found that the best performing model was the TCR-BART model. Between the full validation set and the target-rich set, the TCR-T5 model demonstrated a significant increase in Char-Bleu. Interestingly, the model achieving the lowest perplexity was the vanilla bi-RNN. We suspect that given the high target multiplicity, the use of perplexity as a performance indicator is limited since model convergence towards a sequence "mode" may appear indistinguishable from a poorly trained model altogether. Notably, the TCR-BART model with no pre-training generated more CDR3b sequences unseen during training than the pre-trained version.

Table 1: Generation Accuracy by Model via Greedy Decoding

			Evaluation Metrics			
	Model (120M)	Pre-training?	BLEU	P@1	D_{Edit}	PPL
Full Validation Set (N_{pMHC}=245)	BiRNN	-	.377	.016	6.58	**4.49**
	BiRNN+Attn	-	.407	.057	6.63	5.44
	Conv1D+Attn	-	.410	.049	6.56	6.82
	TCR-BART	-	**.452**	.122	**6.09**	8.73
	TCR-BART	+	**.452**	**.139**	6.16	8.22
	TCR-T5 (66M)	-	.382	.025	6.71	—
	TCR-T5 (66M)	+	.392	.029	6.72	—
Target-Rich Set (N_{pMHC}=10)	BiRNN	-	.786	0.0	3.1	**4.40**
	BiRNN+Attn	-	.803	0.1	3.1	5.31
	Conv1D+Attn	-	.731	0.0	3.6	6.30
	TCR-BART	-	.802	**0.3**	2.6	7.42
	TCR-BART	+	.801	0.2	**2.4**	6.98
	TCR-T5 (66M)	-	.815	0.2	2.5	—
	TCR-T5 (66M)	+	**.826**	0.2	2.7	—

Given the importance of decoding strategies in constructing plausible target sequences [41], we benchmark various decoding methods, with parameters optimized using a grid search method, unless specified in their respective papers (Figure S2). We evaluated precision, recall, and F1@K with K=1000 on multinomial, top-k, top-p, beam, diverse-beam [42], typical [43], and contrastive [44] decoding. For this analysis, we retrained the model with the highest Precision@1, TCR-BART (with pre-training), and finetuned it on three different data splits, each excluding one of the three pMHCs with the highest number of cognate TCRs. We find that across the three different epitopes, beam search decoding outperformed the other methods by sizeable margin, including diverse beam-search (Table 2). We further show that beam search sequence probabilities calculated using token probabilities correlate highly with OLGA generation probabilities, supporting its use as a potential scoring function (Figure S3). However, beam search appeared brittle when recapitulating distributions of target TCR generation probabilities, highlighting a paradox that is currently being investigated (Figure S4).

Table 2: Measuring Generation Diversity by Decoding Method

			Evaluation Metrics		
	Method	Stochastic?	P@1000	R@1000	F1@1000
KLGGALQAK (N_{CDR3b} =12660)	Multinomial	+	.015	.007	.010
	Top-K	+	.005	.005	.005
	Top-P	+	.008	.008	.008
	Beam	-	**.034**	**.034**	**.034**
	Diverse Beam	-	0.0	0.0	0.0
	Typical	+	.018	.008	.011
	Contrastive	+	.007	.007	.007
YVLDHLIVV (N_{CDR3b} =8290)	Multinomial	+	.005	.004	.004
	Top-K	+	.004	.003	.003
	Top-P	+	.003	.003	.003
	Beam	-	**.009**	**.009**	**.009**
	Diverse Beam	-	.003	.001	.002
	Typical	+	.005	.005	.005
	Contrastive	+	0.0	0.0	0.0
GLCTLVAML (N_{CDR3b} =7339)	Multinomial	+	.004	.003	.003
	Top-K	+	.008	.006	.007
	Top-P	+	.007	.006	.006
	Beam	-	**.016**	**.016**	**.016**
	Diverse Beam	-	0.0	0.0	0.0
	Typical	+	.006	.004	.005
	Contrastive	+	.009	.007	.008

4 Discussion

In this preliminary work, we demonstrate the viability of using sequence-to-sequence transformer models like TCR-BART and TCR-T5 to generate antigen-specific TCR sequences conditional on input peptide-MHC, focusing on the need to define meaningful evaluation metrics that capture an unbiased snapshot of model performance. We introduce evaluation metrics tailored to the task that measure accuracy, generalization, and diversity to account for the many-to-many mapping and data sparsity inherent to this problem. However, these metrics are heavily impacted by the availability of experimentally validated data. Even in the cases of epitopes with the largest number of validated TCRs, roughly 1% of the theoretical space is captured, and for the majority of epitopes there is roughly an order of magnitude less. In this setting, evaluating model performance by exact sequence reconstruction against a sparsely sampled label space gives low performance, matching our expectation. However, without *in-silico* methods that can generalize TCR-pMHC binary predictions well or dramatic cost reduction for *in-vitro* methods, we use the aforementioned metrics to characterize model performance in a more holistic manner, acknowledging that many model generations may be true binders *in-vitro*. Our results provide an initial benchmark characterizing the performance trade-offs between these metrics. We note that like many of the binary classification models trained on this data, strong model performance is tied to similarity with training sequences, making out-of-distribution generalization an active work in progress. The capacity to generate large, high-fidelity repertoires of antigen-specific TCRs has immense biological and therapeutic potential. For emerging cellular therapies, it enables rapid, targeted generation of potentially active TCRs that can be tested for cross-reactivity against self-epitopes before being administered to patients. This contrasts current approaches of TCR discovery which require in-vitro identification of reactive T-cell clones from starting biological sample material or the use of discriminatory models against a database of receptor:epitope pairs. As more data becomes available, both model training as well as the metrics used for evaluation stand to improve drastically. Future directions for this work, including implementing diverse pre-training techniques, training and testing on the full TCR and MHC sequences, and utilizing principled data augmentation techniques to address long tail examples all stand to further unlock the promise of this approach across basic immunology and translational medicine.

A Supplementary Information

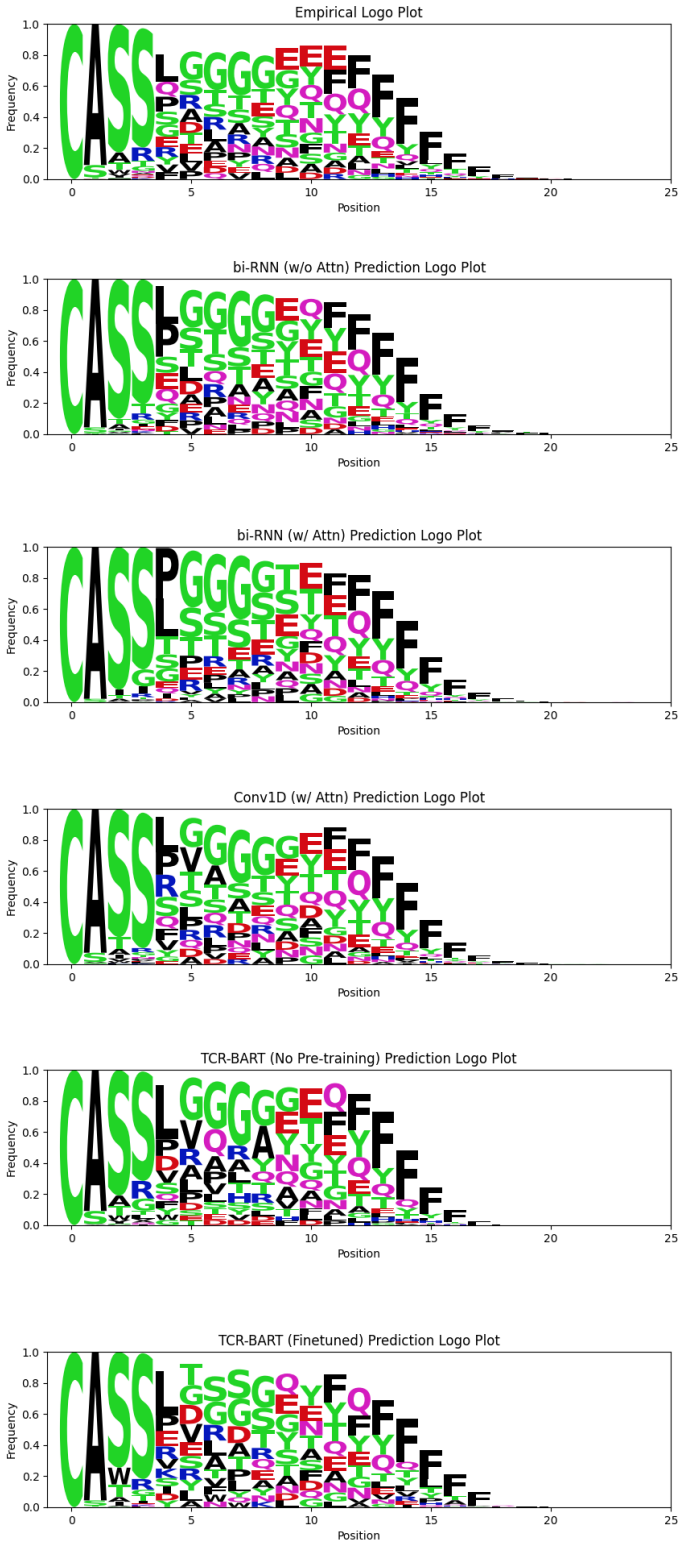
A.1 Hyperparameters

	TCR-BART	TCR-T5
Parameters	120M	66M
d_{model}	768	768
Vocab Size	28	128
Encoder Layers	6	6
Decoder Layers	6	6
Positional Encoding	512	512
Cross Attention	✓	✓

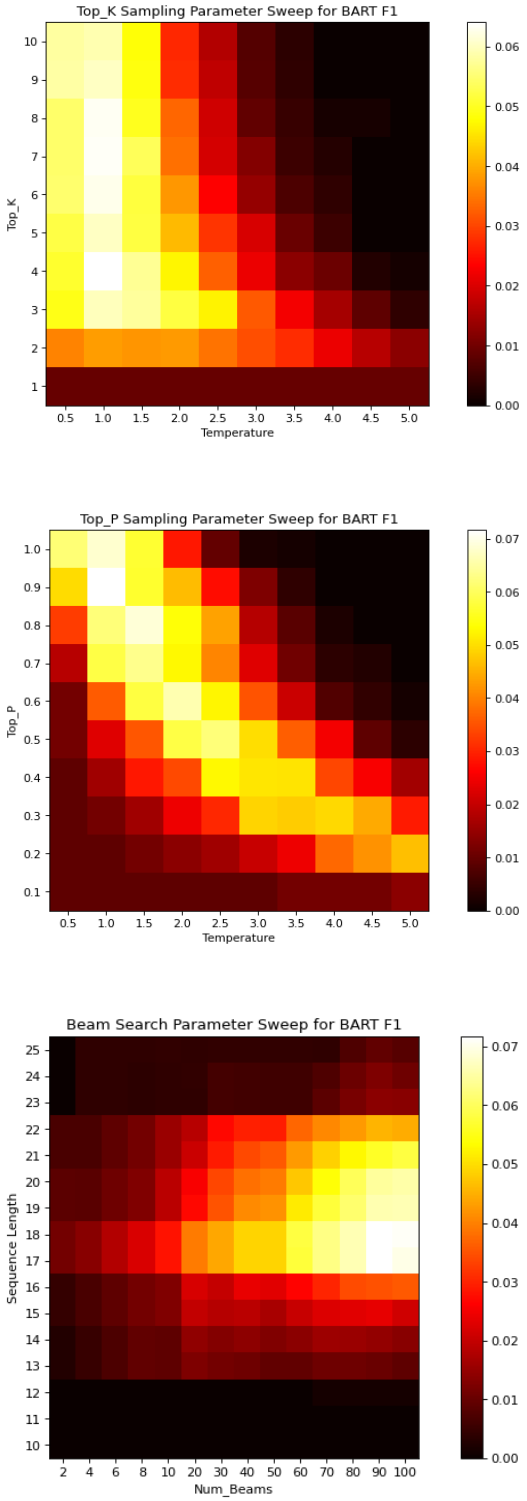
Supplementary Table 1: Model Architecture Hyperparameters

		TCR-BART	TCR-T5
Pre-Training (MLM Objective)	Epochs	1	1
	Batch Size	512	512
	Learning Rate	5e-05	1e-04
	Weight Decay	.01	.01
	Optimizer	AdamW	AdamW
	p_{MLM}	0.15	0.15
Fine-Tuning (Cross Entropy Loss)	Epochs	100	100
	Batch Size	512	512
	Learning Rate	5e-05	1e-04
	Weight Decay	0.0	0.0
	Optimizer	AdamW	AdamW
Direct Training (Cross Entropy Loss)	Epochs	100	100
	Batch Size	512	512
	Learning Rate	5e-05	1e-04
	Weight Decay	0.0	0.0
	Optimizer	AdamW	AdamW

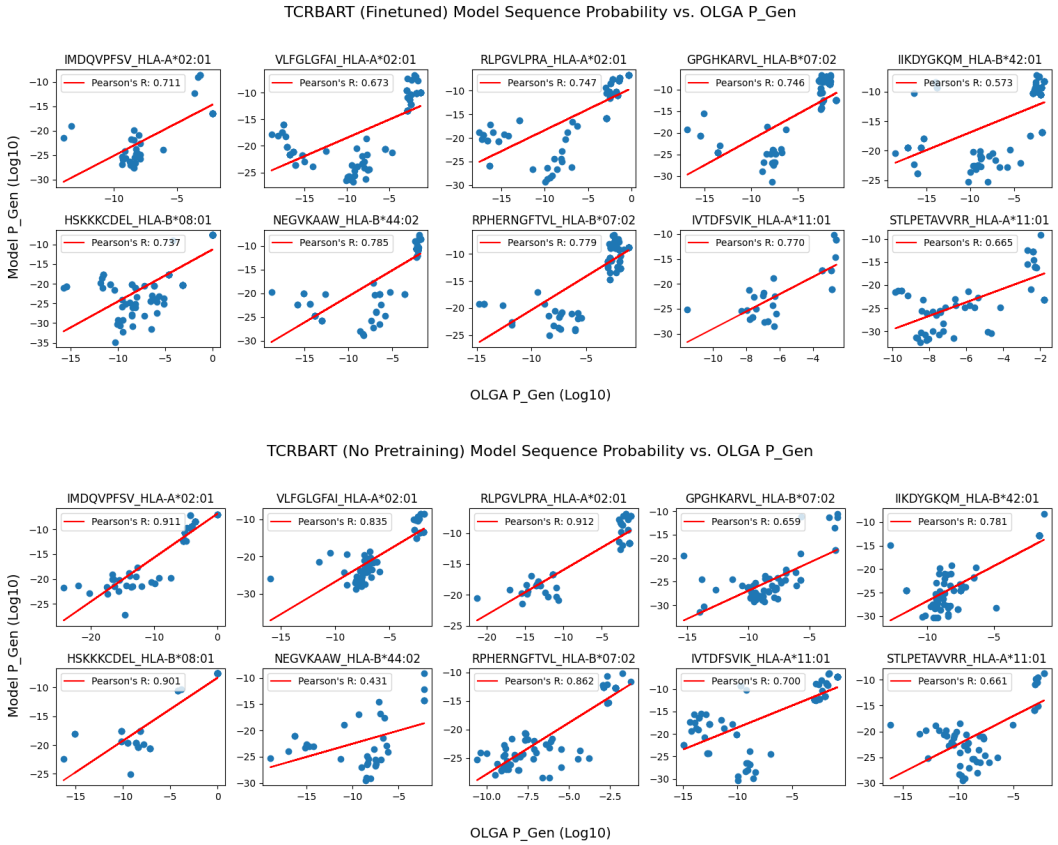
Supplementary Table 2: Model Training Parameters



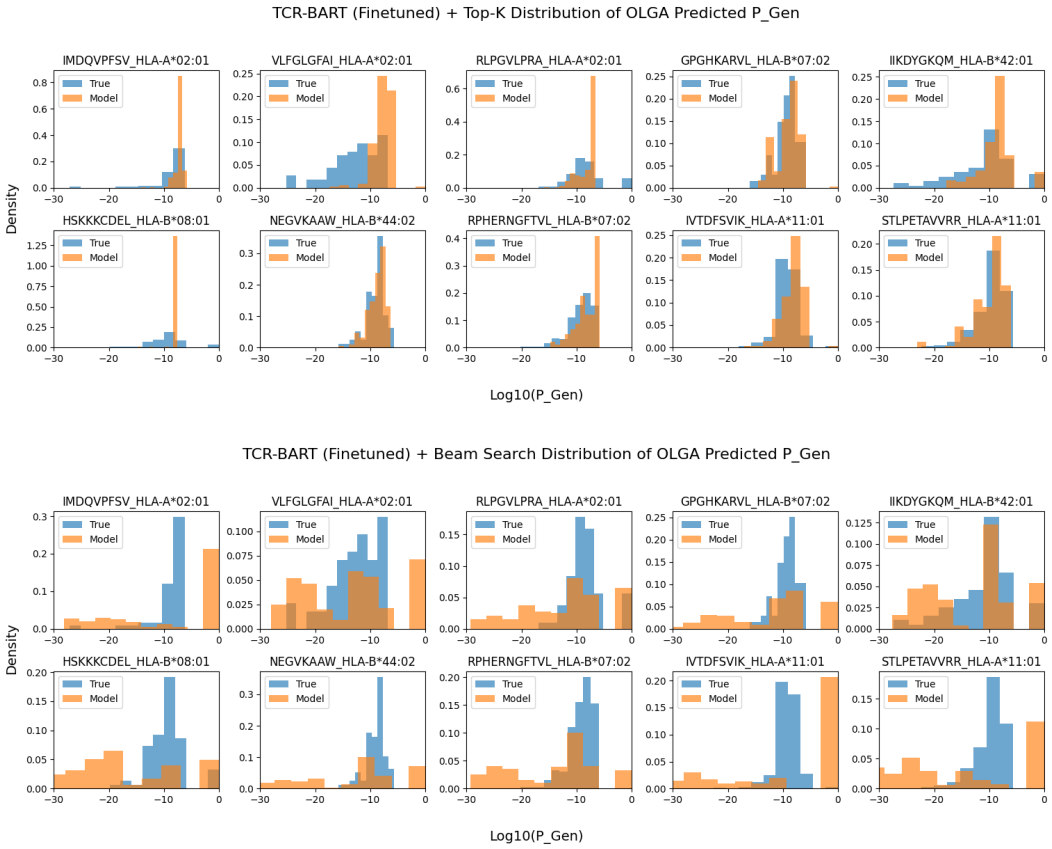
Supplementary Figure 1: CDR3b sequence logo plots for the Hepatitis B virus epitope (STLPETAVVRR) on HLA-A*11:01, held out from training data. Figure shows the empirical logo plot given known ground truth sequences compared to logo plots generated by sampling at a depth of K=1000 using top-k sampling (k=8). Special tokens were removed for clarity.



Supplementary Figure 2: Heatmap of bivariate parameter sweeps for top-k,top-p, and beam search decoding methods colored by F1@100 score. Evaluated on target rich set using the TCR-BART model without pre-training to inform parameters in Table 2 with minimal data leakage in a low-data setting.



Supplementary Figure 3: Scatter-plot of model predicted sequence log-probabilities vs. OLGA generation log-probabilities (Pgen) for the target-rich pMHCs. Red line represents best fit line. Pearson Correlation Coefficient is reported as Pearson's R.



Supplementary Figure 4: Density plot of OLGA predicted generation probabilities for real and conditionally sampled TCRs from target-rich pMHCs using top-k and beam search decoding. Zero probability real sequences are the result of OLGA error whereas sampled sequences with zero probability may indicate error or poor CDR3b sequences (sequences that do not start with a C or end with an F). These sequences were set to 1 for transformation by log scale.

Acknowledgments and Disclosure of Funding

This work was supported largely by the National Science Foundation Graduate Research Fellowship. The authors have no competing interests to declare.

References

[1] 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.

[2] 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.

[3] 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*, pages 1–11, 2023.

[4] Xingyu Cao, Guangna Liu, Jianping Zhang, Yanli Zhao, Hua Chen, Hongli Zheng, Wei Rui, Lemei Jia, Xueqiang Zhao, Xin Lin, and Peihua Lu. 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.

[5] 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.

[6] Susan P. Foy, Kyle Jacoby, Daniela A. Bota, Theresa Hunter, Zheng Pan, Eric Stawiski, Yan Ma, William Lu, Songming Peng, Clifford L. Wang, Benjamin Yuen, Olivier Dalmas, Katharine Heeringa, Barbara Sennino, Andy Conroy, Michael T. Bethune, Ines Mende, William White, Monica Kukreja, Swetha Gunturu, Emily Humphrey, Adeel Hussaini, Duo An, Adam J. Litterman, Boi Bryant Quach, Alphonsus H. C. Ng, Yue Lu, Chad Smith, Katie M. Campbell, Daniel Anaya, Lindsey Skrdlant, Eva Yi-Hsuan Huang, Ventura Mendoza, Jyoti Mathur, Luke Dengler, Bhamini Purandare, Robert Moot, Michael C. Yi, Roel Funke, Alison Sibley, Todd Stallings-Schmitt, David Y. Oh, Bartosz Chmielowski, Mehrdad Abedi, Yuan Yuan, Jeffrey A. Sosman, Sylvia M. Lee, Adam J. Schoenfeld, David Baltimore, James R. Heath, Alex Franzusoff, Antoni Ribas, Arati V. Rao, and Stefanie J. Mandl. Non-viral precision t cell receptor replacement for personalized cell therapy. *Nature*, 615(7953):687–696, Mar 2023. ISSN 1476-4687. doi: 10.1038/s41586-022-05531-1. URL https://doi.org/10.1038/s41586-022-05531-1.

[7] Andrew Poole, Vijaykumar Karuppiah, Annabelle Hartt, Jaafar N. Haidar, Sylvie Moureau, Tomasz Dobrzycki, Conor Hayes, Christopher Rowley, Jorge Dias, Stephen Harper, Keir Barnbrook, Miriam Hock, Charlotte Coles, Wei Yang, Milos Aleksic, Aimee Bence Lin, Ross Robinson, Joe D. Dukes, Nathaniel Liddy, Marc Van der Kamp, Gregory D. Plowman, Annelise Vuidepot, David K. Cole, Andrew D. Whale, and Chandramouli Chillakuri. 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.

[8] Christoph T. Ellebrecht, Vijay G. Bhoj, Arben Nace, Eun Jung Choi, Xuming Mao, Michael Jeffrey Cho, Giovanni Di Zenzo, Antonio Lanzavecchia, John T. Seykora, George Cotsarelis, Michael C. Milone, and Aimee S. Payne. 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.

[9] Jacob Glanville, Huang Huang, Allison Nau, Olivia Hatton, Lisa Wagar, Florian Rubelt, Xuhuai Ji, Arnold Han, Sheri Krams, Christina Pettus, Nikhil Haas, Cecilia Lindestam Arlehamn, Alessandro Sette, Scott Boyd, Thomas Scriba, Olivia Martinez, and Mark Davis. Identifying specificity groups in the t cell receptor repertoire. *Nature*, 547, 06 2017. doi: 10.1038/nature22976.

[10] Pradyot K Dash, Andrew J. Fiore-Gartland, Tomer Hertz, George C. Wang, Shalini Sharma, Aisha Souquette, Jeremy Chase Crawford, E Bridie Clemens, Thi H.O. Nguyen, Katherine Kedzierska, Nicole L. La Gruta, Philip Bradley, and Paul G. Thomas. Quantifiable predictive features define epitope-specific t cell receptor repertoires. *Nature*, 547(7661):89–93, July 2017. ISSN 0028-0836. doi: 10.1038/nature22383.

[11] Vanessa Isabell Jurtz, Leon Eyrich Jessen, Amalie Kai Bentzen, Martin Closter Jespersen, Swapnil Mahajan, Randi Vita, Kamilla Kjærgaard Jensen, Paolo Marcatili, Sine Reker Hadrup, Bjoern Peters, and Morten Nielsen. Nettcr: sequence-based prediction of tcr binding to peptide-mhc complexes using convolutional neural networks. *bioRxiv*, 2018. doi: 10.1101/433706. URL https://www.biorxiv.org/content/early/2018/10/02/433706.

[12] Pieter Moris, Joey De Pauw, Anna Postovskaya, Sofie Gielis, Nicolas De Neuter, Wout Bittremieux, Benson Ogunjimi, Kris Laukens, and Pieter Meysman. Current challenges for epitope-agnostic tcr interaction prediction and a new perspective derived from image classification. *bioRxiv*, 2020. doi: 10.1101/2019.12.18.880146. URL https://www.biorxiv.org/content/early/2020/09/11/2019.12.18.880146.

[13] Esteban Lanzarotti, Paolo Marcatili, and Morten Nielsen. Identification of the cognate peptide-mhc target of t cell receptors using molecular modeling and force field scoring. *Molecular Immunology*, 94:91–97, 2018. ISSN 0161-5890. doi: https://doi.org/10.1016/j.molimm.2017.12.019. URL https://www.sciencedirect.com/science/article/pii/S0161589017306211.

[14] Xingcheng Lin, Jason T. George, Nicholas P. Schafer, Kevin Ng Chau, Michael E. Birnbaum, Cecilia Clementi, José N. Onuchic, and Herbert Levine. Rapid assessment of t-cell receptor specificity of the immune repertoire. *bioRxiv*, 2021. doi: 10.1101/2020.04.06.028415. URL https://www.biorxiv.org/content/early/2021/04/22/2020.04.06.028415.

[15] Masato Ogishi and Hiroshi Yotsuyanagi. Quantitative prediction of the landscape of t cell epitope immunogenicity in sequence space. *Frontiers in Immunology*, 10, 2019. ISSN 1664-3224. doi: 10.3389/fimmu.2019.00827. URL https://www.frontiersin.org/articles/10.3389/fimmu.2019.00827.

[16] Martina Milighetti, John Shawe-Taylor, and Benny Chain. Predicting t cell receptor antigen specificity from structural features derived from homology models of receptor-peptide-major histocompatibility complexes. *Frontiers in Physiology*, 12, 2021. ISSN 1664-042X. doi: 10.3389/fphys.2021.730908. URL https://www.frontiersin.org/articles/10.3389/fphys.2021.730908.

[17] Yao Tong, Jiayin Wang, Tian Zheng, Xuanping Zhang, Xiao Xiao, Xiaoyan Zhu, Xin Lai, and Xiang Liu. Sete: Sequence-based ensemble learning approach for tcr epitope binding prediction. *Computational Biology and Chemistry*, 87:107281, 2020. ISSN 1476-9271. doi: https://doi.org/10.1016/j.compbiolchem.2020.107281. URL https://www.sciencedirect.com/science/article/pii/S1476927120303194.

[18] 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.

[19] Anna Weber, Jannis Born, and María Rodriguez Martínez. TITAN: T-cell receptor specificity prediction with bimodal attention networks. *Bioinformatics*, 37(Supplement):i237–i244, 07 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab294. URL https://doi.org/10.1093/bioinformatics/btab294.

[20] 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.

[21] Bjørn P. Y. Kwee, Marius Messemaker, Eric Marcus, Giacomo Oliveira, Wouter Scheper, Catherine J. Wu, Jonas Teuwen, and Ton N. Schumacher. Stapler: Efficient learning of tcr-peptide specificity prediction from full-length tcr-peptide data. *bioRxiv*, 2023. doi: 10.1101/2023.04.25.538237. URL https://www.biorxiv.org/content/early/2023/04/28/2023.04.25.538237.

[22] 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.

[23] 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.

[24] 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.

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

[26] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation, 2020.

[27] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text, 2016.

[36] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding, 2019.

[37] Ziqi Chen, Martin Renqiang Min, Hongyu Guo, Chao Cheng, Trevor Clancy, and Xia Ning. T-cell receptor optimization with reinforcement learning and mutation policies for precesion immunotherapy, 2023.

[38] Filippo Grazioli, Anja Mösch, Pierre Machart, Kai Li, Israa Alqassem, Timothy J O'Donnell, and Martin Renqiang Min. On tcr binding predictors failing to generalize to unseen peptides. *Frontiers in Immunology*, 13:1014256, 2022.

[39] 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, page 311–318, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://doi.org/10.3115/1073083.1073135.

[40] 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.

[41] Alexandra DeLucia, Aaron Mueller, Xiang Lisa Li, and João Sedoc. Decoding methods for neural narrative generation, 2021.

[42] Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R. Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. Diverse beam search: Decoding diverse solutions from neural sequence models, 2018.

[43] Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. Locally typical sampling, 2023.

[44] Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. A contrastive framework for neural text generation, 2022.