Evaluating zero-shot conditional generation of TCR sequences

Divya Nori [*1] Bridget Li [*1] Rac Mukkamala [*1] Ananth Shyamal [*1]

Abstract

The efficient design of a diverse array of novel T cell receptors (TCRs) would transform the field of immunology, providing a vastly expanded search space from which new immunotherapies can be designed. However, data on TCRs is limited and often biased. Recent advances in diffusion models have proven to be effective in designing biological sequences that are diverse and specific to a given context. We investigate whether a pre-trained sequence diffusion model, EvoDiff, can be used to zero-shot generate TCR CDR3 sequences, conditioned on variable and constant chains. The EvoDiff model generates CDR3s that do not recapitulate the native TCR distribution in humans but have high structural fidelity and diversity, showing potential for synthetic TCR generation.

1. Introduction

T cell receptors (TCRs) are key mediators of adaptive immunity: they recognize short antigen peptides presented on major histocompatibility complex (MHC) surface proteins, leading to T cell activation and downstream effector T cell functions (Shah et al., 2021). Improving *de novo* design of TCRs would enable the development of more effective and robust immunotherapies for various diseases, such as cancer. For example, enhanced TCR design could be used to produce customized TCR-based chimeric antigen receptor T-cells (CAR T cells) with precise tropism towards a desired disease-specific epitope, thereby allowing initiation and coordination of a highly specific immune response (Poorebrahim et al., 2021).

TCRs comprise an extracellular α and a β chain (Figure 1), each of which contains a variable and constant domain. The three complementarity-determining regions (CDRs), located in the variable domains, are where the TCR binds the antigen and, consequently, are the source of the majority of the sequence variation among TCRs (Sun et al., 2021). As a result, efforts to design TCRs have focused on designing CDRs, and in particular, CDR3 - the most variable CDR.

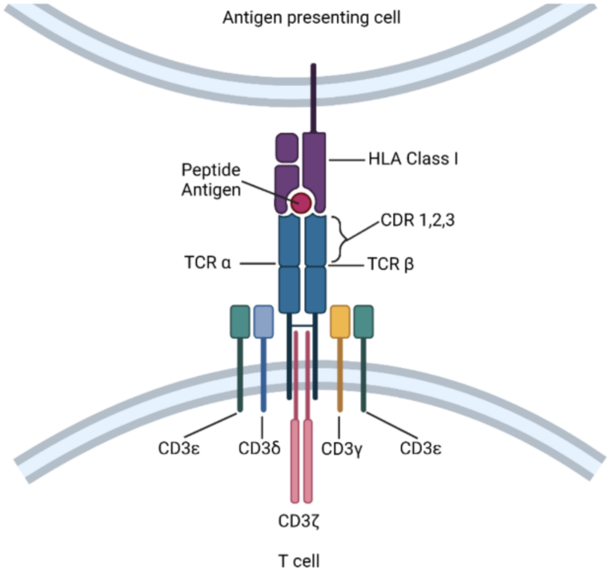

Figure 1. TCRs are key mediators of adaptive immunity.

Machine learning-based models have been increasingly used for the design of biomacromolecules, including TCRs, due to the high experimental costs and durations associated with the synthesis and analysis of these molecules. While the field of *in silico*, *de novo* antibody design has experienced many significant developments over the past few years, *de novo* TCR design is still relatively unexplored despite the structural similarities between TCRs and antibodies. One reason for this is TCR sequencing data is rather limited and biased. Most TCR sequence databases consist of only β chains, and paired α-β chain data is rare. Additionally, databases contain TCR repertoires from limited samples (e.g. 24 people for TCRdb (Chen et al., 2021)) which would heavily bias a generative model trained on this data. Thus, we evaluate whether general protein design models could be leveraged in zero-shot fashion for TCR generation.

A diverse repertoire of TCRs is generated *in vivo* via a mechanism called V(D)J recombination. TCR variable domain sequences comprise of V (variable), D (diversity), and J

[1] Massachusetts Institute of Technology, MA, USA. **AUTHOR-ERR: Missing \icmlcorrespondingauthor.**

The 2022 ICML Workshop on Computational Biology. Baltimore, Maryland, USA, 2022. Copyright 2022 by the author(s).

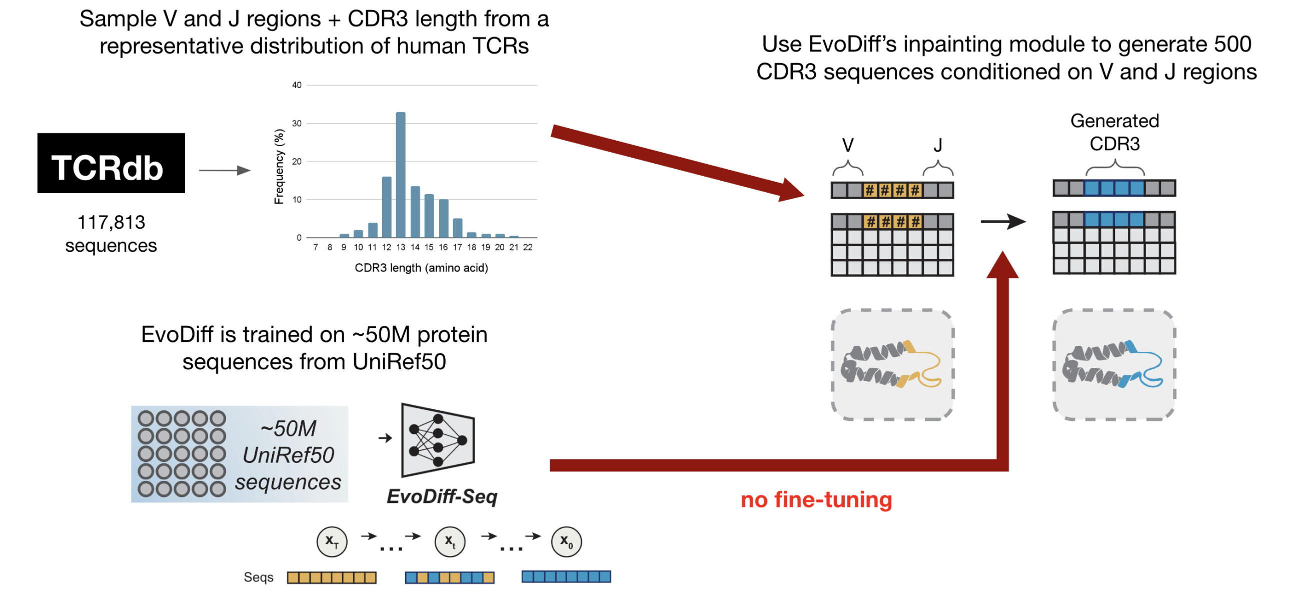
Sample V and J regions + CDR3 length from a representative distribution of human TCRs
TCRdb
117,813 sequences
Frequency (%)
40
30
20
10
0
7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
CDR3 length (amino acid)
Use EvoDiff's inpainting module to generate 500 CDR3 sequences conditioned on V and J regions
V
J
Generated CDR3
EvoDiff is trained on ~50M protein sequences from UniRef50
~50M UniRef50 sequences
EvoDiff-Seq
no fine-tuning
Seqs

Figure 2. We construct a joint distribution of V and J sequences and a distribution over CDR lengths from true human TCRs. To generate a CDR3 sequence using EvoDiff, we sample a V/J sequence pair and CDR3 length to condition the generation. EvoDiff is applied without fine-tuning to generate 500 CDR3 sequences.

(joining) regions, with the CDR1 and CDR2 sequences in the V region and the CDR3 sequence spanning part of the V region and all of the D and J regions. Recently, diffusion models have been used to generate highly diverse and designable proteins, both for *de novo* design and to re-engineer particular domains (Watson et al., 2023). We hypothesize that diffusion models can effectively capture the vast sequence space of TCRs resulting from V(D)J recombination.

We investigate the ability of sequence diffusion models, conditioned on TCR variable and constant chains, to design new CDR3 regions in a zero-shot manner. In particular, we analyze the performance of EvoDiff (Alamdari et al., 2023), a recent discrete diffusion model for controllable protein generation, in designing realistic CDR3 sequences. By modeling the stochastic process of V(D)J recombination and leveraging EvoDiff in a zero-shot manner, enabling a less biased exploration of a wide range of potential TCR sequences, we hope to generate diverse TCR designs.

2. Related Work

2.1. Diffusion Models for Protein Design

Many recent works have developed diffusion models for general protein design tasks, including motif scaffolding, binder design, and shape-conditioned design (Watson et al., 2023; Wu et al., 2024). Many of these models rely on structure; however, the amount of structural data on TCRs is very limited. Hence, we focus on EvoDiff, which enables controllable protein generation in sequence space while maintaining structural plausibility in its predictions. Critically, EvoDiff excels at designing proteins with disordered regions. Since TCR CDR regions are somewhat disordered, we hypothesize that a sequence diffusion model is well-suited for this task.

2.2. Models for CDR Design

While some deep learning-based methods have been developed for antibody CDR design (Jin et al., 2021), there have been relatively fewer methods for TCR CDR design. The most common approach used to date has involved training VAE models on TCR sequencing data to model the native TCR repertoire distribution (Davidsen et al., 2019; Sidhom et al., 2021). These supervised variational autoencoder (VAE) models have shown a strong ability to recapitulate the native TCR sequence distribution and are capable of learning low-dimensional embedding representations of TCR/CDR motifs. However, because these models are trained to recapitulate the TCR repertoire distribution obtained from a limited patient sample set, they may not be effective sampling tools for *de novo* enhanced TCR design; the native distributions they are fitted to can be biased due to prior disease exposure and/or may not be representative of the full theoretical diversity of TCR sequences. Additionally, as mentioned earlier, there is very limited availability of TCRα chain sequencing data, which is perhaps why both of the TCR VAE examples cited earlier limited their analysis to just the TCR β chain.

3. Methods

3.1. Sampling from EvoDiff

As shown in Figure 2, we begin by creating a joint distribution of V and J sequence co-frequency, as well as a distribution of CDR3 lengths, from which we can sample condition sequences. To construct these distributions, we use the TCRdb database (Chen et al., 2021), which contains $117,813$ unique TCR sequences, annotated by V, D, and J regions as well as clone fraction. We sample 500 condition sequence pairs and use EvoDiff to inpaint a CDR3 sequence for each.

3.2. Supervised Learning Baseline

To baseline against a supervised method, we train a model with a simple long-short LSTM encoder layer and an autoregressive LSTM decoder layer to predict CDR3 sequence given the V region sequence. The model was trained on $2,910$ true V-region, CDR3 pairs from VDJdb (Shugay et al., 2018) and supervised with a standard cross entropy loss until convergence. We sample 500 sequences following the same V-region and CDR length distribution as EvoDiff sampling.

3.3. Structural Fidelity

One of the most important properties that a TCR must have is structural fidelity. Particularly, the CDR3 loop must assume specific three-dimensional structure that allows for effective antigen recognition and binding. We evaluate whether the EvoDiff-generated CDR3s assume realistic structures despite being sampled in a zero-shot manner. We fold 100 randomly selected EvoDiff-generated sequences using ESMFold (Lin et al., 2023), folding the generated CDR3 scaffolded into the sampled V/J regions and β constant region. We report the average CDR3 pLDDT across samples. As a positive control, we sample 100 true TCRs from the TCRdb-derived distribution. We also compare against the LSTM baseline.

3.4. Sequence Diversity

Next, we evaluate the diversity amongst the generated set of sequences, comparing EvoDiff and the LSTM baseline. We measure similarity between each pair of generated sequences by summing the BLOSUM-62 scores for corresponding amino acid pairs across the entire length of the aligned sequences and dividing by the length of the alignment. We report the average pairwise similarity across generated samples, so lower scores indicate higher diversity.

3.5. Sequence Fidelity

Finally, we evaluate how realistic the generated CDR sequences are with respect to a distribution of native human TCR β chains. Although we do not expect EvoDiff-generated sequences to score highly on this metric, given that we employ the model in a zero-shot manner, we seek to evaluate how strongly the provided V/J context steers the model towards human-like CDR3 sequences.

To evaluate sequence fidelity, we employ the Optimized Likelihood estimate of immunoGlobin Amino-acid sequences (OLGA) algorithm (Sethna et al., 2019). The OLGA algorithm calculates the probability of a particular CDR3 sequence under a stochastic model of V(D)J recombination. Particularly, the probability of a generation event E is determined by the probability of observing the given V, D, and J gene templates along with a given number of nucleotide deletions d_V, d_J. These probabilities are computed under distributions defined by true human β-chain TCR variable regions, as shown in the equation below. The probability of a particular CDR3 sequence is given by the sum over the probabilities of all possible generation events that could have resulted in the observed sequence.

$$P(E) = P_V(V)P_{DJ}(D,J)P_{delV}(d_V|L)P_{delJ}(d_J|J)$$

$$P(a_1...a_L) = \sum_{E \to \sigma \sim a} P(E)$$

4. Results

We evaluate the EvoDiff-generated sequences along three axes: (i) structural fidelity, (ii) sequence diversity, and (iii) sequence fidelity which are described in the subsequent sections.

4.1. Structural Fidelity

Table 1 shows the pLDDT of structures obtained by folding EvoDiff-generated sequences, LSTM-generated sequences, and true TCRs. Given that the pLDDT of EvoDiff-generated CDR3s is within the error bars of both LSTM-generated and true TCR sequences, it indicates that the structures of EvoDiff-generated CDR3s are realistic compared to true TCRs. EvoDiff is also competitive with the supervised LSTM baseline, even though it is a zero-shot method.

Table 1. Comparison of Models on Structural Fidelity

Model	**Avg. CDR3 pLDDT**
EvoDiff	57.54 ± 4.98
LSTM	61.25 ± 5.41
True TCRs	61.55 ± 5.20

We conduct further analysis on the pLDDT values as a function of sequence position, as shown for a sampled protein in Figure 3. We observe that while the structural confidence goes down in the generated CDR region, there are regions of the constant domain that have lower confidence. We conclude that this level of structural uncertainty is inherent to disordered regions like CDR loops, matching the trends observed in true TCR structures.

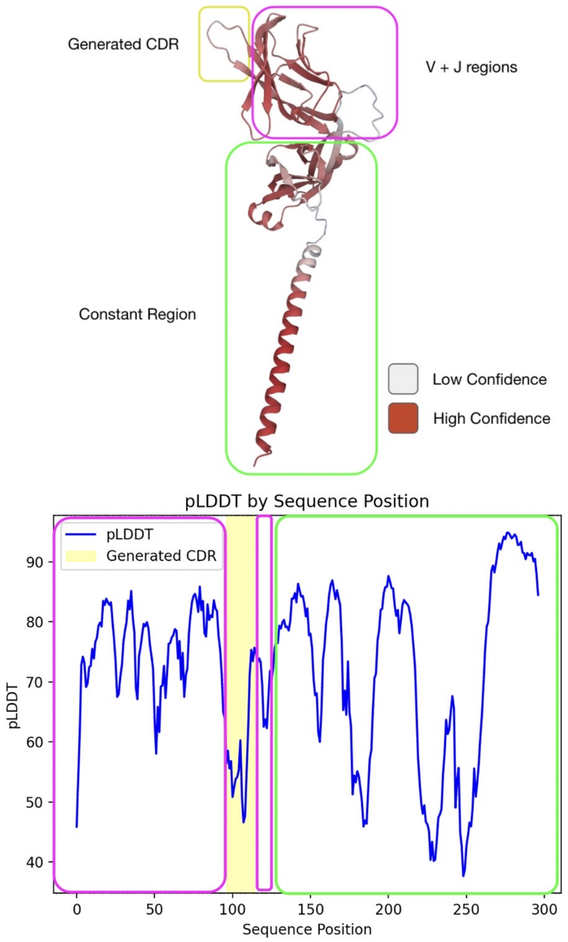

Figure 3. plDDT as a function of sequence position. There is some structural uncertainty in the generated CDR loop and segments of the constant region.

4.2. Sequence Diversity

Table 2 compares the average pairwise similarity scores of EvoDiff and the LSTM. Lower similarity scores imply higher diversity among the generated sequences, meaning that the sequences produced by EvoDiff are more varied compared to those from the LSTM model. This increased diversity in the sequences generated by EvoDiff is promising as it mirrors the natural variability found in the immune system, allowing for a broader exploration of potential TCR behaviors and interactions.

Table 2. Comparison of Models on Average Pairwise Similarity Score

Model	**Average Pairwise Similarity**
EvoDiff	37.18
LSTM	39.91

4.3. Sequence Fidelity

As shown in Table 3, we report the average OLGA probability across generated sequences. As expected, true TCRs have the highest probability. We find that LSTM-generated sequences are more probable than EvoDiff-generated sequences. This shows that zero-shot generalization from general protein design to TCR generation is difficult. Although the LSTM only receives context from the V region, EvoDiff is not able to generate sequences of higher fidelity, highlighting a limitation of zero-shot generation.

Table 3. Comparison of Models on Sequence Fidelity

Model	**Average OLGA Log Probability**
EvoDiff	-27.97
LSTM	-11.75
True TCRs	-7.28

5. Conclusions

While EvoDiff finds it difficult to zero-shot generalize to the distribution of true TCR sequences, we show that EvoDiff can generate TCR CDR3s with high structural fidelity and diversity. By enabling the model to explore the inherently vast sequence space of TCRs, we show that this method can effectively sample synthetic TCRs in low-data settings.

The ability to maintain a high degree of structural fidelity while generating a wide variety of TCR sequences is essential. This ensures that the synthetic TCRs produced by EvoDiff are not only diverse but also maintain the necessary structural features that could be crucial for them to maintain their *in vivo* effector function. By enabling exploration across a vast spectrum of TCR sequences, EvoDiff facilitates a comprehensive understanding and discovery of potentially effective TCRs, even when empirical data is scarce. This opens up new avenues for investigating TCR behaviors and developing novel therapeutic approaches in immunology.

Code Availability

The code and data for the experiments described in this paper is available at `https://github.com/divnori/tcr_gen`.

References

Alamdari, S., Thakkar, N., van den Berg, R., Lu, A. X., Fusi, N., Amini, A. P., and Yang, K. K. Protein generation with evolutionary diffusion: sequence is all you need. *bioRxiv*, 2023.

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*, 2021.

Davidsen, K., Olson, B. J., DeWitt, William S, I., Feng, J., Harkins, E., Bradley, P., and Matsen, Frederick A, I. Deep generative models for t cell receptor protein sequences. *eLife*, 8, sep 2019. doi: 10.7554/eLife.46935.

Jin, W., Wohlwend, J., Barzilay, R., and Jaakkola, T. Iterative refinement graph neural network for antibody sequence-structure co-design. *arXiv preprint arXiv:2110.04624*, 2021.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637): 1123–1130, 2023.

Poorebrahim, M., Mohammadkhani, N., Mahmoudi, R., Gholizadeh, M., Fakhr, E., and Cid-Arregui, A. Tcr-like cars and tcr-cars targeting neoepitopes: an emerging potential. *Nature, Cancer Gene Therapy*, 2021.

Sethna, Z., Elhanati, Y., Jr., C. G. C., Walczak, A. M., and Mora, T. Olga: fast computation of generation probabilities of b- and t-cell receptor amino acid sequences and motifs. *Bioinformatics*, 2019.

Shah, K., Al-Haidari, A., Sun, J., and Kazi, J. U. T cell receptor (tcr) signaling in health and disease. *Nature, Signal Transduction and Targeted Therapy*, 2021.

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., Dyk, E. V., Dash, P., Attaf, M., Rius, C., Ladell, K., McLaren, J. E., Matthews, K. K., Clemens, E. B., Douek, D. C., Luciani, F., van Baarle, D., Kedzierska, K., Kesmir, C., Thomas, P. G., Price, D. A., Sewell, A. K., and Chudakov, D. M. Vdjdb: a curated database of t-cell receptor sequences with known antigen specificity. *Nucleic Acids Research*, 2018.

Sidhom, J.-W., Larman, H. B., Pardoll, D. M., and Baras, A. S. Deeptcr is a deep learning framework for revealing sequence concepts within t-cell repertoires. *Nature Communications*, 12(1), Mar 2021. doi: 10.1038/s41467-021-21879-w.

Sun, Y., Li, F., Sonnemann, H., Jackson, K. R., Talukder, A. H., Katailiha, A. S., and Lizee, G. Evolution of cd8+ t cell receptor (tcr) engineered therapies for the treatment of cancer. *cell*, 2021.

Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., Ahern, W., Borst, A. J., Ragotte, R. J., Milles, L. F., Wicky, B. I. M., Hanikel, N., Pellock, S. J., Courbet, A., Sheffler, W., Wang, J., Venkatesh, P., Sappington, I., Torres, S. V., Lauko, A., Bortoli, V. D., Mathieu, E., Ovchinnikov, S., Barzilay, R., Jaakkola, T. S., DiMaio, F., Baek, M., and Baker, D. De novo design of protein structure and function with rfdiffusion. *Nature*, 2023.

Wu, K. E., Yang, K. K., van den Berg, R., Alamdari, S., Zou, J. Y., Lu, A. X., and Amini, A. P. Protein structure generation via folding diffusion. *Nature Communications*, 15(1):1059, 2024.