Fine-tuned protein language models capture T cell receptor stochasticity

Lewis Cornwall[*] Synteny
Grisha Szep Synteny
James Day Synteny
S R Gokul Krishnan Synteny
David Carter Synteny
Jamie Blundell Synteny
Lilly Wollman Synteny
Neil Dalchau Synteny
Aaron Sim[*] Synteny

Abstract

The combinatorial explosion of T cell receptor (TCRs) sequences enables our immune systems to recognise and respond to an enormous diversity of pathogens. Modelling the highly stochastic TCR generation and selection processes at both sequence and repertoire levels is important for disease detection and advancing therapeutic research. Here we demonstrate that protein language models fine-tuned on TCR sequences are able to capture TCR statistics in hypervariable regions to which mechanistic models are blind, and show that amino acids exhibit strong dependencies on each other within chains but not across chains. Our approach generates representations that improve the prediction of TCR binding specificities.

1 Introduction

T cells are activated when their surface receptors (TCRs) recognise peptides presented on the surface of nearby cells' major histocompatibility complex molecules (pMHCs). To specifically recognise a large diversity of peptide sequences, a TCR consists of two different proteins: an alpha chain and a beta chain[1]. TCRs are generated randomly through the process of V(D)J recombination, which enables huge diversity to be concentrated in the third complementarity-determining region (CDR3) within both the alpha chain and the beta chain [2]. The diversity of TCRs has two related consequences: predicting TCR-pMHC binding specificities is challenging, and generating meaningful TCR representations is non-trivial.

A recent trend in generating protein representations has been towards utilising protein language models (PLMs), and this approach has already demonstrated success in protein folding [3] and predicting disease variant effects [4]. Although the applications of PLMs to antibody sequences have been explored in some detail [5, 6, 7], the potential benefits of fine-tuning on other non-conserved protein sequences such as TCRs remain less well understood. Here we demonstrate that PLMs fine-tuned on TCR sequences reveal correlations between amino acids in hypervariable regions, and that the extracted representations can be useful in downstream applications such as the prediction of TCR-pMHC binding. Throughout this work, we demonstrate these results using a PLM with 150M parameters, ESM-2 [8], fine-tuned on 2.4M single-cell TCR sequences [9].

[*]Corresponding authors: lewis@synteny.ai, aaron@synteny.ai

[1]We restrict our attention to TCRs which consist of an alpha chain and a beta chain, as such TCRs account for 95% of those found in humans [1].

NeurIPS 2023 Generative AI and Biology Workshop.

2 Results

Fine-tuned PLMs recover biological features of TCR sequences that mechanistic models do not. Let $a_1 a_2 \cdots a_L$ be a TCR sequence of length L, with a_i the amino acid at position i. Throughout this work, we will make use of some simplified notation, and take

$$\mathbb{P}(a_i | a_j \cdots a_k)$$

to denote the probability that a model assigns to the true amino acid at position i, given the amino acids at positions $\{j, \ldots, k\}$. With this notation, the perplexity at position i for the TCR is

$$\mathcal{P}_i := \mathbb{P}(a_i | a_1 a_2 \cdots a_{i-1} a_{i+1} \cdots a_L)^{-1}.$$

With the fine-tuned model, we evaluated the perplexity of each amino acid within the hypervariable CDR3 for each TCR in the test set. As shown in Figure 1a, the perplexity is greater in the beta chain, owing to additional combinatorial factors introduced by the D-gene. Note that there is a biophysical lower limit on the perplexity within each chain due to the inherent stochasticity of V(D)J recombination and selection pressures [2]. The distribution of perplexities is bimodal for both chains. Some amino acids are inferred with very low perplexity, whereas the second mode is comparable with the perplexity of the distribution of amino acid frequencies.

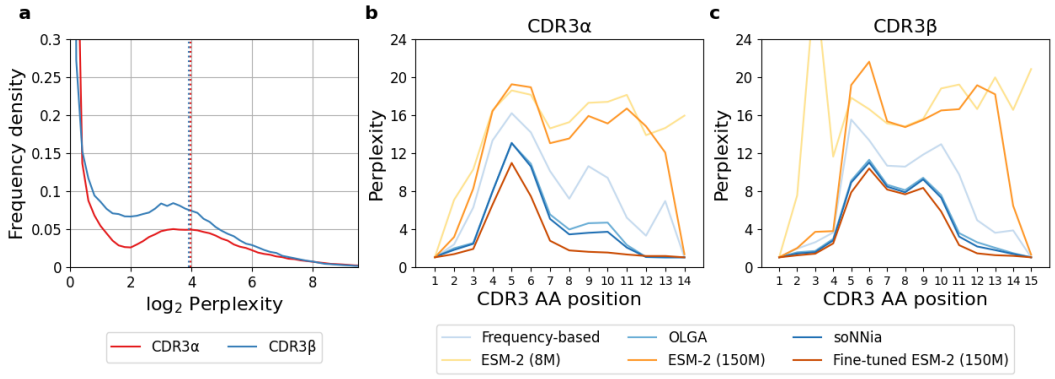

Figure 1: **(a)** The distribution of log perplexity (entropy) for amino acids in the CDR3α and CDR3β. The dotted vertical lines mark the entropy of the distribution of amino acid frequencies for each chain. **(b-c)** The geometric mean perplexity of amino acids by position within the CDR3.

To further explore the nature of the distribution of perplexities in the CDR3, we calculated the perplexity *by position*. The fine-tuned model finds the perplexity to be lower towards the start and end of the CDR3, where the amino acids are highly constrained by the V- and J-genes. Figures 1b and 1c illustrate this for the most common CDR3α and CDR3β lengths of 14 and 15, respectively. The analogous plots for chains of different lengths can be found in Supplementary Figure S1. Within the CDR3β, the perplexity also dips towards the middle of the CDR3, reflecting the fact that the D-gene constrains these amino acids. We include for comparison ESM-2 with just 8M parameters, and note that it does not recognise the relatively conserved amino acids towards to end of the CDR3.

We considered three further methods for comparison: a method based on amino acid frequency; OLGA [10]; and soNNia [11]. The frequency-based method attributes probabilities according to the frequency of each amino acid at each position in the training set. OLGA is a mechanistic model of V(D)J recombination that calculates the probability that a TCR is generated in the absence of any additional selection pressures. soNNia is a deep learning model that extends OLGA by learning such pressures on selection to obtain the probability that a TCR appears in a given repertoire.

Let $p_{\text{gen}}(a_1 \cdots a_L)$ denote the probability of generation as given by OLGA. The perplexity at position i according to OLGA is

$$\begin{aligned} \mathcal{P}_i &= \mathbb{P}(a_i | a_1 a_2 \cdots a_{i-1} a_{i+1} \cdots a_L)^{-1} \\ &= \left[\frac{p_{\text{gen}}(a_1 \cdots a_{i-1} a_i a_{i+1} \cdots a_L)}{\sum_{\tilde{a} \in \mathcal{A}} p_{\text{gen}}(a_1 \cdots a_{i-1} \tilde{a}\, a_{i+1} \cdots a_L))} \right]^{-1}, \end{aligned}$$

where $\mathcal{A}$ is the set of amino acids. An analogous calculation can be performed for soNNia. As shown in Figures 1b and 1c, fine-tuning ESM-2 brings the perplexity by position below that of OLGA and soNNia, demonstrating a strength of our approach over mechanistic models.

Amino acids exhibit strong dependencies within chains, but not across chains. We extended the above analysis of single amino acid statistics to examine the pairwise relationships between amino acids to which fine-tuned PLMs attend. Consider a pair of positions within the TCR, (i, j), where, without loss of generality, $i < j$. Allowing ourselves our earlier abuse of notation, the mutual information for this pair, conditioned on complete knowledge of the rest of the TCR, can be calculated as [12]

$$\mathcal{H}_{ij} \coloneqq \log_2 \frac{\mathbb{P}(a_i a_j | a_1 a_2 \dots a_{i-1} a_{i+1} \dots a_{j-1} a_{j+1} \dots a_L)}{\mathbb{P}(a_i | a_1 a_2 \dots a_{i-1} a_{i+1} \dots a_L)\mathbb{P}(a_j | a_1 a_2 \dots a_{j-1} a_{j+1} \dots a_L)}.$$

We calculated the mean mutual information between amino acids by position before and after fine-tuning the PLM for TCRs in the test set. Before fine-tuning, mutual information is attributable mostly to noise. After fine-tuning, the PLM identifies strong correlations between neighbouring amino acids, and several weaker correlations between more distant amino acids on the same chain. However, despite fine-tuning on both alpha and beta chains together, the PLM *does not* identify any similar relationships between amino acids on complementary chains.

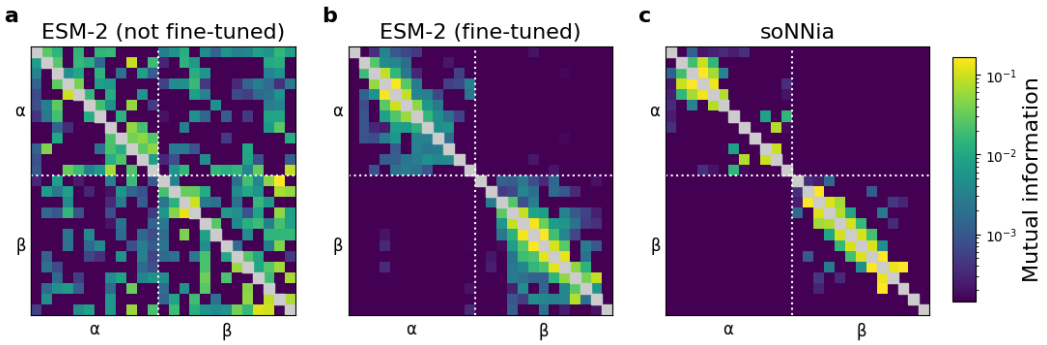

Figure 2: The mutual information between pairs of amino acids for three different models.

We also considered the mutual information for the semi-mechanistic model soNNia. Comparing Figure 2b to 2c, we see that soNNia identifies fewer significant dependencies between distant amino acid pairs, instead relying predominately on short-range dependencies spanning only a handful of amino acid positions. This offers an explanation for the lower perplexity of the fine-tuned PLM compared to soNNia: it is able to uncover weaker correlations between more distant amino acids.

Fine-tuned representations improve TCR-pMHC binding predictions. One strength of our approach of fine-tuning a PLM over mechanistic models is that it provides representations that can be used for downstream tasks, such as TCR-pMHC binding. We designed a novel architecture that takes a TCR and pMHC representation and returns a probability of binding. Crucially, the architecture involves a learnt weighting of amino acid representations in the TCR and pMHC, so that not all amino acids are equally emphasised in the prediction.

Weight initialisation	Fine-tuning on TCRs	AUROC	AUPRC
Random	No	0.738	0.165
ESM-2	No	0.742	0.182
Random	Yes	0.757	0.189
ESM-2	Yes	**0.770**	**0.218**

Table 1: The performance of our binding specificity model. We compare the results for TCR representations generated with and without fine-tuning, and initialised with and without the pre-trained ESM-2 weights.

As shown in Table 1, the binding specificity model performs best on the test set when using representations generated by fine-tuned PLMs. Moreover, we find that using the pre-trained ESM-2 weight initialisation, rather than fine-tuning from a random initial condition, is beneficial. We infer the representations trained from the ESM-2 initial conditions retain latent information about general proteins from ESM-2's training data [13] that can be harnessed for predicting binding specificities.

We conclude that generating meaningful TCR representations, such as those generated through fine-tuned PLMs, is likely to be a crucial element in solving the TCR-pMHC binding problem. Here we have fine-tuned a PLM with 150M parameters, but PLMs have been trained with as many as 100B parameters [14], and further work may explore the relationship of our results to scale.

3 Data and Methods

TCR sequence data. Single-cell TCR sequence data were sourced from six healthy repertoires [9], with gene assignments determined by AbStar [15] and complete TCR sequences constructed using Stitchr [16]. The six repertoires were pooled to create a universal donor repertoire from which we created three different datasets: for inference of the alpha chain, of the beta chain, and of both chains. Within each dataset, two TCRs were deemed duplicate and removed if and only if all chains under consideration for that dataset were identical (for example, two TCRs with the same alpha chain but different beta chain are considered duplicate in the alpha chain dataset, but not in the beta chain or both chains dataset). Each dataset was split into a train, validation, and test set in the ratio 80:10:10.

TCR-pMHC binding data. Binding data were sourced from three public databases: VDJdb [17], IEDB [18], and McPAS-TCR [19]. Non-human TCRs and HLA types other than HLA-A*02:01 were filtered out. Two TCRs were identified as duplicate if they shared the same CDR3α and CDR3β sequences. If both CDR3s were identical, the TCRs were merged. If exactly one chain was identical, the TCRs were placed in the same train/validation/test split to circumvent data leakage [20]. Train, validation, and test sets were generated in the ratio 80:10:10. For each of the datasets, negative binding pairs were generated by exchanging the TCR in a positive binding pair with a random TCR, sampled according to the distribution of TCRs in the dataset [21]. In the validation set, we restricted negative sampling to TCRs within the same batch. The ratio of negative to positive binding pairs was set to 9:1 in the training and validation sets and 50:1 in the test set.

Masked amino acid prediction task. ESM-2 consists of a transformer-like architecture [22] with a RoBERTa head [23]. To fine-tune ESM-2, we followed the ESM-2 procedure of masking out 15% of amino acids using BERT-style replacement [24]. The model was fined-tuned on TCR sequence data for six epochs using Adam [25] with the same hyperparameters as ESM-2. Rather than training the full model, LoRA with rank four was used for each of the transformer layers [26]. To train soNNia, we used a learning rate of 0.001, which we determined to be the optimal learning rate from the set $\{0.01, 0.003, 0.001, 0.0003\}$, and trained until the negative log-likelihood loss was minimised on the validation set.

TCR-pMHC binding prediction task. Let $\mathbf{a}_i$ and $\mathbf{b}_j$ be the representations of the amino acid at position i along the CDR3α and position j along the CDR3β, respectively. For our novel binding architecture, we first calculated a representation of the TCR in terms of the representations of each amino acid in the TCR through

$$\mathbf{t} \coloneqq V\left(\text{GELU}\left[\sum_i (\mathbf{w}_\alpha^\top \mathbf{a}_i)\mathbf{a}_i\right] \oplus \text{GELU}\left[\sum_j (\mathbf{w}_\beta^\top \mathbf{b}_j)\mathbf{b}_j\right]\right),$$

where $\mathbf{w}_\alpha \in \mathbb{R}^n$, $\mathbf{w}_\beta \in \mathbb{R}^n$, and $V \in \mathbb{R}^{n\times m}$ are learnt tensors, and n and m are the dimensions of the amino acid and TCR representations, respectively. Through $\mathbf{w}_\alpha$ and $\mathbf{w}_\beta$, a weighted sum of amino acid representations is learnt, so that amino acids are able to make different contributions in accordance with their relevance for binding.

We sought to minimise the cross entropy loss between the predicted labels, $\hat{y}(p,t)$, and actual labels, $y(p,t)$, across all (peptide, TCR) pairs (p,t) in the dataset, where

$$\hat{y}(p,t) \coloneqq \frac{2}{1+\exp||\mathbf{p}-\mathbf{t}||^2},$$

and $\mathbf{p} \in \mathbb{R}^n$ is a learnt embedding of the pMHC p. We trained the binding head for sixteen epochs using Adam [25] with learning rate 0.001, $\beta_1 = 0.9$, and $\beta_2 = 0.999$.

References

[1] Kenneth Murphy and Casey Weaver. *Janeway's Immunobiology*. Garland Science, 2016.

[2] Frederick W Alt et al. "VDJ recombination". In: *Immunology Today* 13.8 (1992), pp. 306–314.

[3] Ratul Chowdhury et al. "Single-sequence protein structure prediction using a language model and deep learning". In: *Nature Biotechnology* 40.11 (2022), pp. 1617–1623.

[4] Nadav Brandes et al. "Genome-wide prediction of disease variant effects with a deep protein language model". In: *Nature Genetics* 55.9 (2023), pp. 1–11.

[5] Tobias H Olsen, Iain H Moal, and Charlotte M Deane. "AbLang: an antibody language model for completing antibody sequences". In: *Bioinformatics Advances* 2.1 (2022).

[6] Brian L Hie et al. "Efficient evolution of human antibodies from general protein language models". In: *Nature Biotechnology* (2023).

[7] Hongtai Jing et al. "Accurate Prediction of Antibody Function and Structure Using Bio-Inspired Antibody Language Model". In: *arXiv 2308.16713* (2023).

[8] Zeming Lin et al. "Evolutionary-scale prediction of atomic-level protein structure with a language model". In: *Science* 379.6637 (2023), pp. 1123–1130.

[9] Matthew J Spindler et al. "Massively parallel interrogation and mining of natively paired human TCR$\alpha\beta$ repertoires". In: *Nature Biotechnology* 38.5 (2020), pp. 609–619.

[10] Zachary Sethna et al. "OLGA: fast computation of generation probabilities of B-and T-cell receptor amino acid sequences and motifs". In: *Bioinformatics* 35.17 (2019), pp. 2974–2981.

[11] Giulio Isacchini et al. "Deep generative selection models of T and B cell receptor repertoires with soNNia". In: *Proceedings of the National Academy of Sciences* 118.14 (2021).

[12] Aaron D Wyner. "A definition of conditional mutual information for arbitrary ensembles". In: *Information and Control* 38.1 (1978), pp. 51–59.

[13] Baris E Suzek et al. "UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches". In: *Bioinformatics* 31.6 (2015), pp. 926–932.

[14] Bo Chen et al. "xTrimoPGLM: Unified 100b-Scale Pre-Trained Transformer for Deciphering the Language of Protein". In: *bioRxiv 2023.07.05.547496* (2023).

[15] Bryan Briney and Dennis R Burton. "Massively scalable genetic analysis of antibody repertoires". In: *bioRxiv 447813* (2018).

[16] James M Heather et al. "Stitchr: stitching coding TCR nucleotide sequences from V/J/CDR3 information". In: *Nucleic Acids Research* 50.12 (2022), pp. 68–68.

[17] Mikhail Shugay et al. "VDJdb: a curated database of T-cell receptor sequences with known antigen specificity". In: *Nucleic Acids Research* 46.D1 (2018), pp. 419–427.

[18] Randi Vita et al. "The Immune Epitope Database (IEDB): 2018 update". In: *Nucleic Acids Research* 47.D1 (2019), pp. 339–343.

[19] Nili Tickotsky et al. "McPAS-TCR: a manually curated catalogue of pathology-associated T cell receptor sequences". In: *Bioinformatics* 33.18 (2017), pp. 2924–2929.

[20] Barthelemy Meynard-Piganeau et al. "TULIP-a Transformer based Unsupervised Language model for Interacting Peptides and T-cell receptors that generalizes to unseen epitopes". In: *bioRxiv 2023.07.19.549669* (2023).

[21] Bjorn PY Kwee et al. "STAPLER: Efficient learning of TCR-peptide specificity prediction from full-length TCR-peptide data". In: *bioRxiv 2023.04.25.538237* (2023).

[22] Ashish Vaswani et al. "Attention is all you need". In: *Advances in Neural Information Processing Systems* 30 (2017).

[23] Yinhan Liu et al. "RoBERTa: A Robustly Optimized BERT Pretraining Approach". In: *arXiv 1907.11692* (2019).

[24] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *arXiv 1810.04805* (2018).

[25] Diederik P Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *arXiv 1412.6980* (2014).

[26] Edward J Hu et al. “LoRA: Low-Rank Adaptation of Large Language Models”. In: *arXiv 2106.09685* (2021).