TCR-BERT: learning the grammar of T-cell receptors for flexible antigen-binding analyses

Kevin E. Wu [1 2 3] **Kathryn E. Yost** [2] **Bence Daniel** [2 4] **Julia A. Belk** [1 4] **Yu Xia** [5] **Takeshi Egawa** [5]
Ansuman Satpathy [4] **Howard Y. Chang** [2 6] **James Zou** [1 3]

Abstract

The T-cell receptor (TCR) allows T-cells to recognize and respond to antigens presented by infected and diseased cells. However, due to TCRs' staggering diversity and complex binding dynamics, it is challenging to predict which antigens a given TCR may bind. Here, we present TCR-BERT, which applies language modeling techniques to learn a general, versatile representation of TCR sequences, enabling numerous downstream applications.

1. Introduction

Mature T cells continuously monitor their surroundings for signs of diseased cells and help activate other immune defenses upon recognition. This recognition is mediated by the T-cell receptor (TCR), which binds to antigens – short peptide chains presented on the external surface of cells. Healthy cells present antigens to identify themselves as benign, whereas infected cells present antigens that signal that they are diseased. In a viral infection, a cell might present viral peptides as antigens, and in cancer, mutated protein fragments called neoantigens are presented (Strønen et al., 2016). Autoreactivity, or the aberrant recognition of self-antigens as invaders can cause autoimmune disorders like type 1 diabetes (Pugliese, 2017). Across these settings, understanding TCRs' binding and recognition of antigen sequences is key to understanding the underlying mechanisms of disease and developing effective treatments.

[1]Department of Computer Science, Stanford University, Stanford, CA, USA [2]Center for Personal Dynamic Regulomes, Stanford University, Stanford, CA, USA [3]Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA, USA [4]Department of Pathology, Stanford University School of Medicine, Stanford, CA, USA [5]Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, MO, USA [6]Howard Hughes Medical Institute, Stanford University, Stanford, CA, USA. Correspondence to: Kevin Wu <wukevin@stanford.edu>, James Zou <jamesz@stanford.edu>.

The 2022 ICML Workshop on Computational Biology.

Despite the importance and therapeutic potential of T cells, it is challenging to predict TCR-antigen recognition behavior. The TCR itself is a dimeric protein with two hypervariable chains – typically α and β chains, encoded by the TRA and TRB genes. These TRA and TRB sequences are specified through recombination of the variable (V), diversity (D), and junction (J) gene segments, as well as through random insertions and deletions. This stochastic process generates a staggering diversity of TCR sequences – often estimated to be on the order of (hundreds of) millions for a healthy human individual (Qi et al., 2014). This diversity crucially lends the immune system its ability to recognize a vast array of antigens, but also makes precisely understanding and predicting TCR-antigen specificity difficult. This challenge is compounded by cross reactivity – a single TCR often recognizes multiple antigens, and conversely, an antigen may be recognized by multiple TCRs (Petrova et al., 2012). Furthermore, only a small fraction of TCR sequences are "labelled" with known antigen binding specificities, and TRA-TRB sequence pairings are frequently unavailable.

With these challenges in mind, we developed TCR-BERT, a large language model trained specifically on TCR amino acids with TCR-specific pre-training tasks. Unlike previous approaches to modeling TCR-antigen binding, which have focused on training supervised models using relatively limited datasets of TCRs with known specificities, TCR-BERT leverages the wealth of unlabeled TCR sequences for pre-training to achieve state-of-the-art performance on a wide variety of downstream tasks and applications. These range from predicting the antigen binding preferences of a TCR sequence to grouping TCRs with similar binding properties, and extends to computationally generating synthetic TCRs with engineered specificities.

2. Results

TCR-BERT takes a TCR sequence (e.g., CASRPDGRETQYF) and tokenizes it into individual residues as input to a BERT transformer architecture (Bertoletti & Tan, 2020) which is pre-trained on two objectives sequentially. First, we pre-train TCR-BERT to predict masked amino acids (MAA); this is analogous to the canonical

masked language modeling objective in BERT and aims to learn the grammar of what constitutes a valid TCR. This is done using a dataset of ($n = 88403$) predominantly human TRA and TRB amino acid sequences (without flags distinguishing them) from the VDJdb (Bagaev et al., 2020) and PIRD (Zhang et al., 2020) databases. After MAA pre-training, we leverage the fact that some TCRs are "labelled" with known binding affinities to train TCR-BERT to predict antigen specificity given a TCR sequence. This takes the form of a multi-class classification problem with 44 classes – each corresponding to a different antigen – using TRB sequences from the PIRD database (insufficient labelled TRA sequences). This biologically-motivated antigen classification pre-training task particularly improves model performance (Appendix A).

2.1. Predicting antigen specificity

After pre-training, we can use TCR-BERT to generate embedding vectors for TCR sequences. These embeddings can be paired with a support vector machine (SVM) trained to predict whether TCRs bind to a specific antigen. This approach is simple and produces strong results when available data is in the order of tens to hundreds of examples.

For a subset ($n = 26$) of antigens used in our second classification pre-training step that have at least 20 known binding TCR sequences, we repeat classification pre-training excluding that antigen and its associated TRBs. We then use the resulting model to embed and classify the held-out antigen's binding TRBs against a background of naturally-occurring human TCR sequences, which are assumed to be non-binding and are sampled at 5 negatives to each binding sequence. The average area under the precision-recall curve (AUPRC) across these antigens is 0.91, compared to an AUPRC of 0.17 for a random classifier. We compare this performance to that of a convolutional neural network – a common architecture for classifying biological sequences (Zou et al., 2019) – trained separately for each antigen and find that TCR-BERT's embeddings with an SVM provides better performance for all but one antigen (Figure 1).

We repeat this "antigen cross-validation" process with several other classifiers to contextualize TCR-BERT's performance. To evaluate the advantage of having a TCR-specific transformer model, we used the same SVM classifier, but instead used different protein transformers to embed the TRB sequences. Compared to general protein transformers TAPE (Rao et al., 2019) and ESM (Rives et al., 2021), TCR-BERT provides superior performance classifying every evaluated antigen, despite the fact that ESM and TAPE have much larger training sets. We additionally evaluate TCR-BERT against a previously published supervised model, SETE (Tong et al., 2020), and again found that TCR-BERT provides improved performance in all antigens (Appendix B).

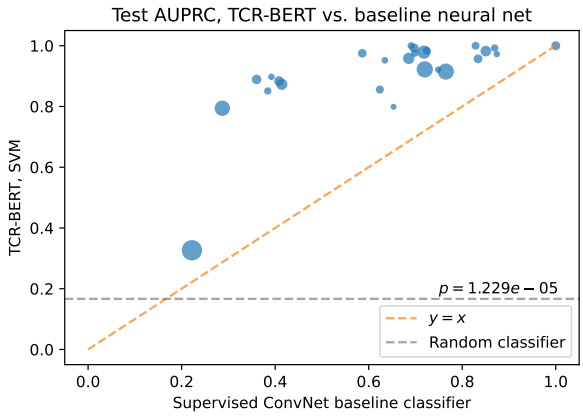


Figure 1. Comparison of TCR-BERT embeddings with an SVM classifier (y axis) against a supervised convolutional network (x axis). Each dot indicates test AUPRC classifying TCRs that bind or do not bind to a given antigen. Size of each dot corresponds to the number of training points (21-231, single outlier of 4115); p-value from two-sided Wilcoxon test.

We additionally sought to evaluate how TCR-BERT would perform under patient-based data splits. As each individual's immune system independently generates TCR sequences, cross-patient generalization is more indicative of how antigen binding classifiers might be used in a clinical setting. We focus on ($n = 214$) human TRB sequences binding to the NP177 influenza A viral antigen, which was not seen during pre-training (Glanville et al., 2017). We use one patient's data ($n = 176$ binding TRBs) for training, and 4 patients' data ($n = 38$) for evaluation. Negative sequences are sampled from naturally-occurring human TRBs at a ratio of 5:1. Among all evaluated approaches (ESM with SVM, TAPE with SVM, SETE, ConvNet), TCR-BERT provides the best test AUPRC of 0.40 (Appendix B).

Finally, we explored how TCR-BERT could be used to build classifiers in a more data-rich scenario that enables fine-tuning the TCR-BERT transformer itself. We use a dataset of ($n = 17702$, 13% positive binding) murine TRA-TRB pairs with measured binding to the murine GP33 antigen (Daniel et al., 2021). We create two copies of the TCR-BERT architecture, each with weights initialized from the MAA pre-training step. These two arms embed the TRA and TRB sequences, respectively; these embeddings are concatenated and passed through a fully-connected classification head, and this entire network is fine-tuned. We compare this approach to a similar "two-armed" convolutional network, with two convolutional networks each responsible for embedding the TRA or TRB. We additionally benchmark against embedding the TRA and TRB using ESM or TAPE, concatenating the embeddings, and training an SVM. We also evaluate DeepTCR (Sidhom et al., 2021),

a prior method for embedding TRA-TRB pairs, that we train on murine TCRs (not including GP33-binders). We also include logistic regression on a *k*-mer featurization of the TRA and TRB sequences. Across all these methods, TCR-BERT provides the best performance (Figure 2).

Overall, our results demonstrate that TCR-BERT provides a strong, versatile foundation for building classifiers predicting antigen specificity. This can be achieved via using TCR-BERT to generate sequence embeddings, which is typically optimal for smaller datasets, or by using larger datasets to fine-tune TCR-BERT. We also show that TCR-BERT can be flexibly used to predict binding for TRB sequences alone, or for TRA/TRB sequence pairs.

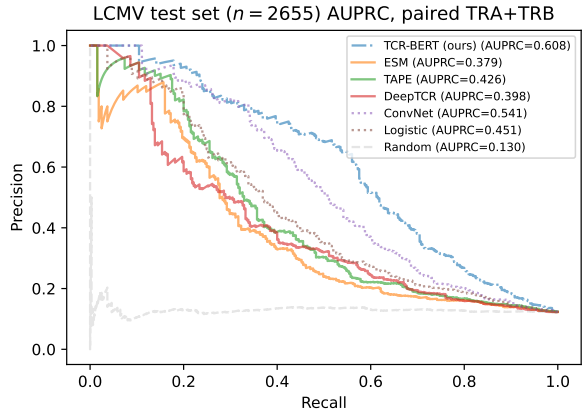


Figure 2. Binding/non-binding classification performance on murine TRA-TRB pairs with measured binding to GP33. Across all evaluated methods, TCR-BERT exhibits the best AUPRC.

2.2. Unsupervised exploration of TCRs

In many cases, researchers may not know *a priori* which specific antigen(s) to predict TCR affinity for, necessitating more exploratory analyses of TCR sequence data. One common task is identifying groups of TCRs likely to share antigen binding properties, which can help identify motifs and provides a more succinct representation of the TCR repertoire than analyzing each unique TCR. TCR-BERT facilitates such analyses via clustering its TCR sequence embeddings via an algorithm like Leiden (Traag et al., 2019).

We evaluate this using a subset ($n = 2443$) of aforementioned murine GP33 dataset containing TRA/TRB pairs. When clustering these sequences, we ideally want most input sequences to belong to a cluster (i.e., the clustering meaningfully captures similarity between sequences), and each cluster should be homogeneous (i.e., GP33-binding sequences should not be clustered with non-binders). These two ideas are quantified by two canonical metrics in TCR clustering (Glanville et al., 2017): percent clustered (proportion of sequences in a cluster with at least 3 constituent members) and percent correctly clustered (clustering accuracy when assigning each cluster's "label" to be the dominant binding state). We benchmark using Leiden to cluster TCR-BERT's TRB sequence embeddings[1] against two popular methods for grouping TCRs: GLIPH (Glanville et al., 2017) and TCRDist3 (Mayer-Blackwell et al., 2021). Both these prior methods use sequence heuristics to compare and group TCRs. GLIPH can only process TRBs, whereas TCRDist3 can process TRBs or TRA/TRB pairs. For all methods, we evaluate various clustering resolutions to evaluate the range of percent clustered and percent correctly clustered (Figure 3). We observe that TCR-BERT provides a smooth tradeoff between these metrics, whereas both GLIPH and TCRDist3 struggle to cluster a meaningful proportion of sequences with above-random accuracy.

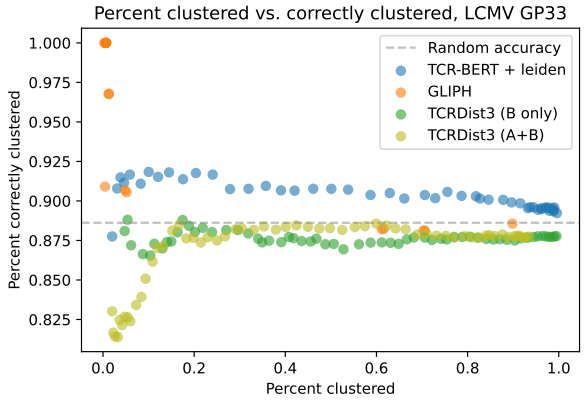


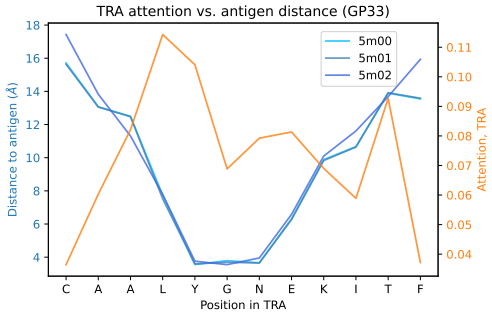
Figure 3. Clustering performance of TCR-BERT and heuristic-based methods GLIPH and TCRDist3 on the GP33 dataset. The x and y axes capture the tradeoff between informativeness and correctness of the clustering, respectively.

2.3. Attentions focus on biologically relevant residues

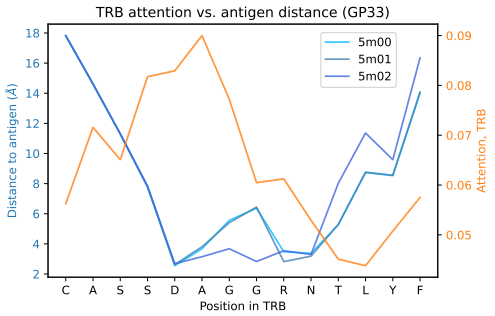
Not only does TCR-BERT excel for a range of common downstream TCR analyses, it does so by recognizing biologically reasonable signals. To demonstrate this, we visualize TCR-BERT's attentions when predicting GP33 binding. Recall that this model was fine-tuned with two "arms" each embedding the TRA or TRB. We examine ($n = 157$) test set sequences of uniform length (12 and 14 residues for TRA and TRB, respectively) and average each TRA/TRB arm's per-residue attentions across these examples. We compare these attentions to distances from each TCR residue to the closest antigen residue, using three experimentally profiled structures profiling a similar GP33 system as references (PDB IDs 5m00, 5m01, 5m02). Attention is anti-correlated

[1]Embedding both TRA and TRB sequences does not yield different performance for TCR-BERT compared to TRB alone.

with distance to the antigen for both the TRA and TRB chains (Figure 4). This matches biological intuition that physical contact between the TCR and the antigen is a primary mechanism for recognition.



(a) TRA attentions vs. distance to antigen



(b) TRB attentions vs. distance to antigen

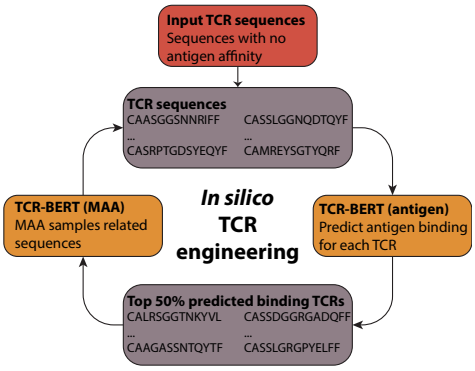
Figure 4. TCR-BERT attentions (orange), averaged across GP33 test sequences of uniform length, plotted against distances from TCR to antigen from three similar empirical GP33 structures (hues of blue). TCR-BERT assigns the greatest attention to residues closest to the antigen – residues most likely to contact the antigen.

2.4. Designing novel TCR sequences using TCR-BERT

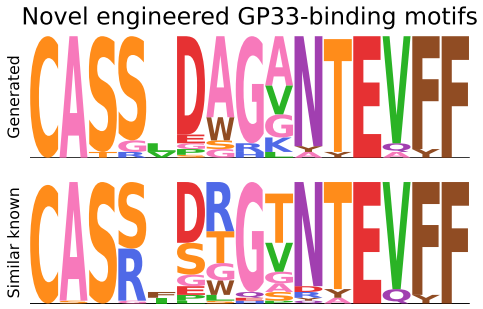
TCR-BERT also enables novel computational approaches to experimental and clinical challenges involving TCRs. Among these, one exciting domain is TCR engineering, which seeks to redirect T cell specificity by introducing synthetic TCR sequences into T-cells. In principle, this should boost T cells' ability to recognize specific antigens and pathogens, enhancing the immune system's response to specific infections (Rapoport et al., 2015).

We generate novel TCR sequences targeting the murine GP33 antigen using the version of TCR-BERT fine-tuned to predict GP33 binding as an oracle. We start with 100 TRA-TRB pairs with no measured binding drawn from the test set. For each sequence in the top half of predicted binding affinities, we mask random residues and use TCR-BERT's masked amino acid predictions to generate new, similar sequences that adhere to the "grammar" of natural TCR sequences. We repeat this cycle, starting from the top half of these generated sequences with strongest predicted affinities, until we converge to a set of sequences with strong binding (Figure 5(a), Appendix C).

To computationally validate these generated sequences, we use BLAST (Altschul et al., 1990) to match the generated TRBs against all known murine TRB sequences (insufficient data to similarly compare TRA sequences). We find that our generated sequences bear significant similarity to a set of TRBs found to bind GP33 in a separate experiment not included in training (Figure 5(b)). This strongly suggests that we are able to generate biologically reasonable sequences with desirable binding patterns.



(a) TCR engineering procedure



(b) Generated vs. known GP33 binders

Figure 5. TCR engineering using TCR-BERT as a generative model and oracle. We use this process to generate synthetic TCRs highly similar to previously-characterized GP33 binders, starting from sequences with no GP33 binding activity.

3. Discussion

TCR-BERT is a large language model trained to embed T-cell receptor sequences. Compared to prior approaches, TCR-BERT leverages unlabelled data to achieve state-of-the-art performance across a wide variety of tasks, and serves as a computational platform for future technologies.

4. Software and Data

All code is available from https://github.com/wukevin/tcr-bert. Pre-trained models are available; see instructions on GitHub.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.

Bagaev, D., Vroomans, R., Samir, J., Stervbo, U., Rius, C., Dolton, G., Greenshields-Watson, A., Attaf, M., Egorov, E., Zvyagin, I., Babel, N., Cole, D., Godkin, A., Sewell, A., Kesmir, C., Chudakov, D., Luciani, F., and Shugay, M. Vdjdb in 2019: database extension, new analysis infrastructure and a t-cell receptor motif compendium. *Nucleic Acids Res*, 48(D1):D1057–D1062, 2020.

Bertoletti, A. and Tan, A. Challenges of car- and tcr-t cell-based therapy for chronic infections. *J Exp Med*, 217(5): e20191663, 2020.

Daniel, B., Yost, K. E., Sandor, K., Xia, Y., Qi, Y., Hiam-Galvez, K. J., Meier, S. L., Belk, J. A., Giles, J. R., Wherry, E. J., Chang, H. Y., Egawa, T., and Satpathy, A. T. Divergent clonal differentiation trajectories of t cell exhaustion. *bioRxiv*, 2021. doi: 10.1101/2021.12.16.472900. URL https://www.biorxiv.org/content/early/2021/12/17/2021.12.16.472900.

Glanville, J., Huang, H., Nau, A., Hatton, O., Wagar, L., Rubelt, F., Ji, X., Han, A., Krams, S., Pettus, C., Haas, N., Arlehamn, C., Sette, A., Boyd, S., Scriba, T., Martinez, O., and Davis, M. Identifying specificity groups in the t cell receptor repertoire. *Nature*, 547(7661):94–98, 2017.

Mayer-Blackwell, K., Schattgen, S., Cohen-Lavi, L., Crawford, J., Souquette, A., Gaevert, J., Hertz, T., Thomas, P., Bradley, P., and Fiore-Gartland, A. Tcr meta-clonotypes for biomarker discovery with tcrdist3 enabled identification of public, hla-restricted clusters of sars-cov-2 tcrs. *Elife*, 10:e68605, 2021.

Petrova, G., Ferrante, A., and Gorski, J. Cross-reactivity of t cells and its role in the immune system. *Crit Rev Immunol*, 32(4):349–372, 2012.

Pugliese, A. Autoreactive t cells in type 1 diabetes. *J Clin Invest*, 127(8):2881–2891, 2017.

Qi, Q., Liu, Y., Cheng, Y., Glanville, J., Zhang, D., Lee, J., Olshen, R., Weyand, C., Boyd, S., and Goronzy, J. Diversity and clonal selection in the human t-cell repertoire. *Proc Natl Acad Sci U S A*, 111(36):13139–13144, 2014.

Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, X., Canny, J., Abbeel, P., and Song, Y. Evaluating protein transfer learning with tape. *Adv Neural Inf Process Syst*, 32:9689–9701, 2019.

Rapoport, A., Stadtmauer, E., Binder-Scholl, G., Goloubeva, O., Vogl, D., Lacey, S., Badros, A., Garfall, A., Weiss, B., Finklestein, J., Kulikovskaya, I., Sinha, S., Kronsberg, S., Gupta, M., Bond, S., Melchiori, L., Brewer, J., Bennett, A., Gerry, A., Pumphrey, N., Williams, D., Tayton-Martin, H., Ribeiro, L., Holdich, T., Yanovich, S., Hardy, N., Yared, J., Kerr, N., Philip, S., Westphal, S., Siegel, D., Levine, B., Jakobsen, B., Kalos, M., and June, C. Ny-eso-1-specific tcr-engineered t cells mediate sustained antigen-specific antitumor effects in myeloma. *Nat Med*, 21(8):914–921, 2015.

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C., Ma, J., and Fergus, R. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci U S A*, 118(15):e2016239118, 2021.

Sidhom, J., Larman, H., Pardoll, D., and Baras, A. Deeptcr is a deep learning framework for revealing sequence concepts within t-cell repertoires. *Nat Commun*, 12(1):1605, 2021.

Strønen, E., Toebes, M., Kelderman, S., van Buuren, M., Yang, W., van Rooij, N., Donia, M., Böschen, M., Lund-Johansen, F., Olweus, J., and Schumacher, T. Targeting of cancer neoantigens with donor-derived t cell receptor repertoires. *Science*, 352(6291):1337–1341, 2016.

Tong, Y., Wang, J., Zheng, T., Zhang, X., Xiao, X., Zhu, X., Lai, X., and Liu, X. Sete: Sequence-based ensemble learning approach for tcr epitope binding prediction. *Comput Biol Chem*, 87:107281, 2020.

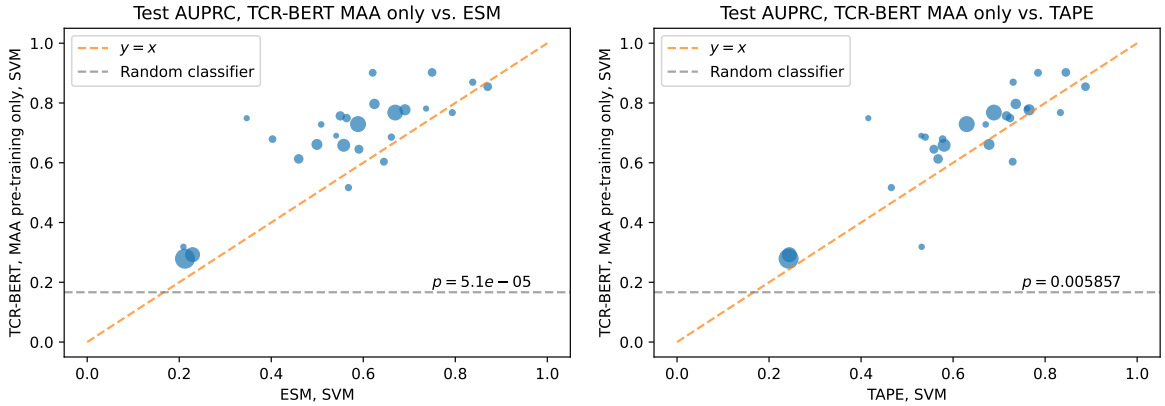
Traag, V., Waltman, L., and van Eck, N. From louvain to leiden: guaranteeing well-connected communities. *Sci Rep*, 9(1):5233, 2019.

Zhang, W., Wang, L., Liu, K., Wei, X., Yang, K., Du, W., Wang, S., Guo, N., Ma, C., Luo, L., Wu, J., Lin, L., Yang, F., Gao, F., Wang, X., Li, T., Zhang, R., Saksena, N., Yang, H., Wang, J., Fang, L., Hou, Y., Xu, X., and Liu, X. Pird: Pan immune repertoire database. *Bioinformatics*, 36(3):897–903, 2020.

Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., and Telenti, A. A primer on deep learning in genomics. *Nature Genetics*, 51(1):12–18, 2019.

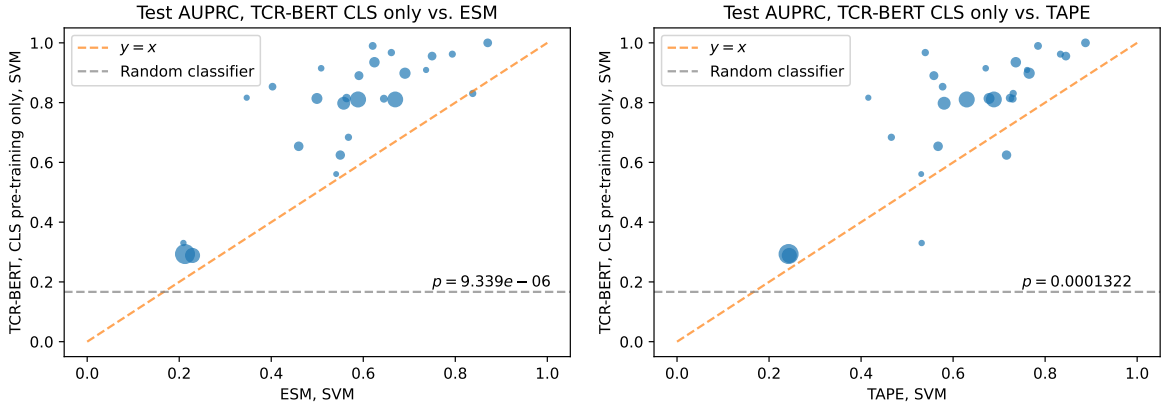
A. Ablation of pre-training objectives reveals importance of biologically relevant tasks

We demonstrate that TCR-BERT provides superior performance compared to general-purpose protein transformer models, specifically ESM (Rives et al., 2021) and TAPE (Rao et al., 2019). Since these models are architecturally similar to TCR-BERT, they allow us to understand how our two pre-training tasks and our selection of data for pre-training contribute to TCR-BERT's performance advantage. We re-trained TCR-BERT using either only masked amino acid prediction, or using only antigen classification pre-training, and evaluated the resulting models against ESM and TAPE. In all cases, evaluation is done by taking the model's fixed embedding and training an SVM classifier, followed by running the previously described antigen cross-validation procedure (Figure 1).



(a) TCR-BERT with MAA only vs. ESM (b) TCR-BERT with MAA only vs. TAPE

Figure 6. We evaluate TCR-BERT trained only with MAA against general protein transformer models. We observe a slight but significant performance advantage compared to either ESM or TAPE (p-value calculated using a two-sided Wilcoxon test).



(a) TCR-BERT with classification pre-training only vs. ESM (b) TCR-BERT with classification pre-training only vs. TAPE

Figure 7. We evaluate TCR-BERT trained with only antigen classification against general protein transformer models (p-value from two-sided Wilcoxon test). We observe a larger performance gain than in Figure 6, which trains TCR-BERT with only masked amino acids.

We first evaluate TCR-BERT with only masked amino acid (MAA) prediction pre-training (Figure 6). We find that while TCR-BERT's performance with only MAA still exceeds that of either ESM or TAPE, their performance is much more comparable. Since ESM and TAPE use a similar architecture and the same MAA pre-training objective, the primary difference is the dataset of sequences used for MAA pre-training. TCR-BERT uses relatively few sequences, all of which are TCRs ($n = 88403$) whereas ESM and TAPE are trained on much larger sets of general protein sequences (250 million sequences for ESM). TCR-BERT's performance advantage here suggests that using a pre-training *dataset* that is closer to

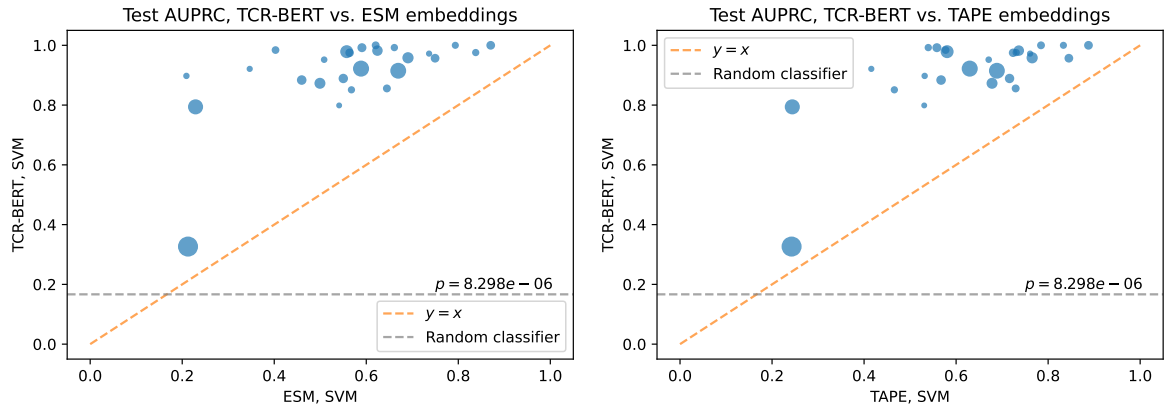
the target application can yield a better embedding, even if this entails having much fewer training examples.

We then evaluate TCR-BERT with only the classification pre-training task (Figure 7) and find a much larger performance gap compared to ESM and TAPE. Here, the model architectures are comparable, so this comparison focuses on the difference between ESM and TAPE's general MAA objective, compared to TCR-BERT's TCR-antigen classification objective alone. TCR-BERT's TCR-specific classification pre-training task appears to yields a large performance advantage. This suggests that having pre-training *tasks* tailored to target applications can yield a substantial performance advantage.

Overall, these results suggest that intentionally designing pre-training datasets *and* (particularly) tasks with downstream applications in mind can yield performance improvements even if doing so entails having much fewer training examples – as is the case here for TCR-BERT.

B. Comprehensive benchmarking of TCR-BERT's classification performance

We compare TCR-BERT's TCR-antigen binding classification performance to that of several methods using antigen cross-validation. To study the advantage of TCR-BERT over using general purpose protein transformers, we evaluate TAPE and ESM using the same approach of generating an embedding and training an SVM on that embedding. TCR-BERT outperforms both TAPE and ESM, which suggests that TCR-BERT's performance advantages are not simply architectural and that having a TCR-specific pre-training is beneficial (Figure 8, see also Appendix A).



(a) TCR-BERT vs. ESM

(b) TCR-BERT vs. TAPE

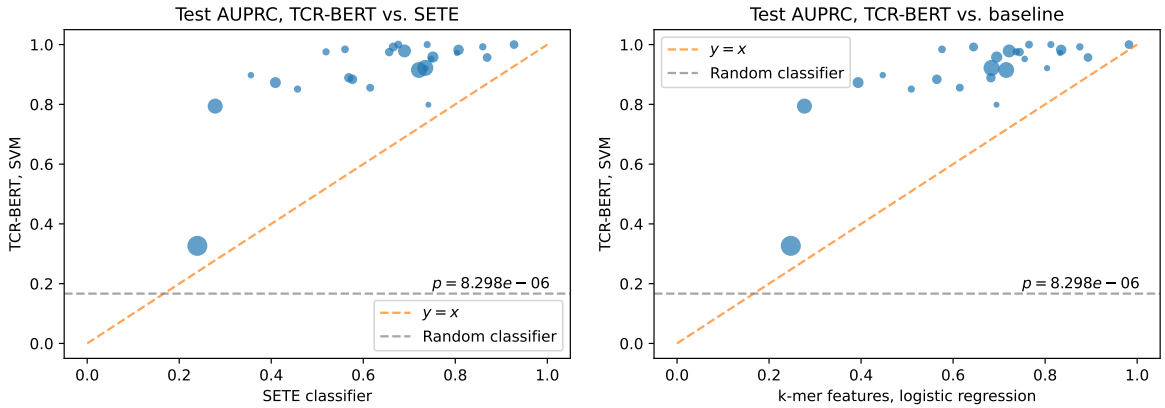
Figure 8. TCR-BERT embeddings compared to embeddings from general protein transformers.

We also compared TCR-BERT against supervised approaches that do not leverage pre-training or transfer learning. In addition to the ConvNet comparison (Figure 1), we evaluate SETE, a method that applies a tree-based classifier on k-mer featurization of TCR sequences (Tong et al., 2020), and a simple logistic regression on k-mer featurization. We find that TCR-BERT outperforms these methods as well.

Since each individual's immune system generates a unique set of TCRs, we evaluated how these classifiers generalize across patient-based splits using a set of TCRs binding NP177 (Glanville et al., 2017). We randomly sampled background human TCRs at a ratio of 5:1. Using one individual's data to train and evaluating on four patients' data, we observe that protein large language models (TCR-BERT, ESM and TAPE) that leverage pre-training tend to produce stronger generalizability than supervised models (SETE, ConvNet), with TCR-BERT again being the most performant (Figure 10).

C. TCR engineering iterations

Our TCR engineering process uses TCR-BERT as both an oracle and a generative model. With each iteration, we sample the top 50% of predicted binders, and use masked amino acid prediction to generate new similar sequences. We stop iterating when all generated sequences have at least a predicted binding of 0.95. Our methodology results in a steady increase in predicted binding of generated sequences without simply repeating sequences seen during training (Figure 11). This process can be repeated with different initial sets of negative sequences to produce different final engineered sequences.



(a) TCR-BERT compared to SETE (b) TCR-BERT compared to logistic regression on k-mers

Figure 9. TCR-BERT performance compared to supervised baselines. Both methods are built on k-mer featurizations.

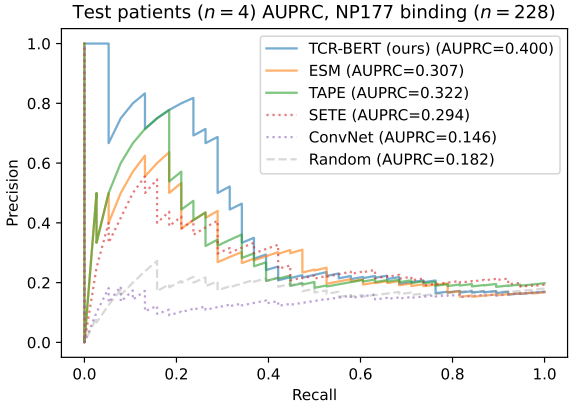
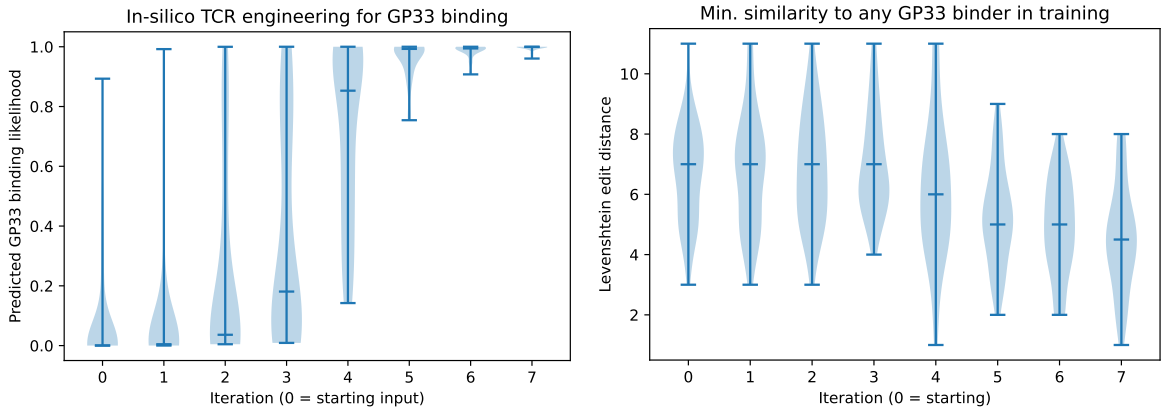


Figure 10. Performance of TCR-BERT and comparison methods when trained on one patient's TCRs and evaluated on 4 different patients' TCRs. Solid lines indicate large language models trained on protein sequences; dotted lines indicate supervised models.



(a) Predicted GP33 binding over iterations (b) Min. edit distance to any positive training ex. per iteration

Figure 11. Over each iteration of TCR engineering, we increase the predicted likelihood of binding (a) without simply regurgitating training sequences (b).