ImmSET: Sequence-Based Predictor of TCR-pMHC Specificity at Scale

Marco Garcia Noceda   MGARCIANOCEDA@ADAPTIVEBIOTECH.COM
Matthew T. Noakes   MNOAKES@ADAPTIVEBIOTECH.COM
Andrew FigPope   AFIGPOPE@ADAPTIVEBIOTECH.COM
Daniel E. Mattox   DMATTOX@ADAPTIVEBIOTECH.COM
Bryan Howie   BHOWIE@ADAPTIVEBIOTECH.COM
Harlan Robins   HROBINS@ADAPTIVEBIOTECH.COM
Adaptive Biotechnologies, Seattle, WA, USA

Abstract

T cells are a critical component of the adaptive immune system, playing a role in infectious disease, autoimmunity, and cancer. T cell function is mediated by the T cell receptor (TCR) protein, a highly diverse receptor targeting specific peptides presented by the major histocompatibility complex (pMHCs). Predicting the specificity of TCRs for their cognate pMHCs is central to understanding adaptive immunity and enabling personalized therapies. However, accurate prediction of this protein-protein interaction remains challenging due to the extreme diversity of both TCRs and pMHCs. Here, we present ImmSET (Immune Synapse Encoding Transformer), a novel sequence-based architecture designed to model interactions among sets of variable-length biological sequences. We train this model across a range of dataset sizes and compositions and study the resulting models' generalization to pMHC targets. We describe a failure mode in prior sequence-based approaches that inflates previously reported performance on this task and show that ImmSET remains robust under stricter evaluation. In systematically testing the scaling behavior of ImmSET with training data, we show that performance scales consistently with data volume across multiple data types and compares favorably with the pre-trained protein language model ESM2 fine-tuned on the same datasets. Finally, we demonstrate that ImmSET can outperform AlphaFold2 and AlphaFold3-based pipelines on TCR-pMHC specificity prediction when provided sufficient training data. This work establishes ImmSET as a scalable modeling paradigm for multi-sequence interaction problems, demonstrated in the TCR-pMHC setting but generalizable to other biological domains where high-throughput sequence-driven reasoning complements structure prediction and experimental mapping.

Keywords: T cell receptor (TCR), Peptide-MHC (pMHC), Immunoinformatics, Protein-protein interaction, Sequence representation learning, Scaling Laws

1. Introduction

Many clinically important prediction problems hinge on how multiple biological segments interact rather than properties of any single component in isolation. In immunology, one of the most challenging examples is predicting T cell receptor (TCR) recognition of peptide-major histocompatibility complex (pMHC) ligands, the interaction that determines whether a T cell initiates an immune response (Gray et al., 2025).

Each T cell carries a unique receptor built from two protein chains, α and β (TCRα and TCRβ), that together contribute six complementarity-determining regions (CDRs). These hypervariable loops scan peptides displayed by human leukocyte antigen (HLA) proteins on other cells, acting as a sentinel for infection, cancer, and other abnormalities. In cancer immunotherapy, vaccine development, and autoimmune disease research, a critical barrier is the ability to anticipate responses to previously unstudied peptides (Hudson et al., 2023). For example, the broad category of cancer neoantigens comprises tumor-specific peptides arising from subtle alterations of self-proteins and therefore falling outside any pre-computable library of targets (Xie et al., 2023). The scale of this problem is staggering: the theoretical diversity of the TCR repertoire exceeds 10^{15} unique sequences (Valkiers et al., 2022), while deep sequencing has revealed that each person expresses on the order of 10^8 unique TCRs (Sun et al., 2022). The peptide space is likewise astronomical. Considering 8-10 amino acid peptides, even if only 0.01% of these are processed and presented in a biologically relevant context, this still leaves on the order of 10^9 potential peptide targets (Sewell, 2012).

Despite decades of research, there are no simple rules governing TCR–pMHC recognition (Gray et al., 2025). Recent work shows that even weaker-binding TCR-pMHC interactions can sometimes be more resilient under force than stronger ones, underscoring that binding strength alone does not predict activation (Pettmann et al., 2023). Functional responses may depend on multiple transient binding events and distinct conformations—different three-dimensional arrangements of the same molecules—rather than a single stable geometry (Siller-Farfán and Dushek, 2018; Buckle and Borg, 2018). Structure-based tools like AlphaFold (Jumper et al., 2021; Abramson et al., 2024) have been transformative for protein science, but they are optimized to predict a single structure per complex (Masrati et al., 2021). Capturing the full dynamic landscape underlying TCR–pMHC recognition would require modeling multiple conformations or running molecular dynamics (MD) simulations, which is computationally intensive even for a single interaction (Hollingsworth and Dror, 2018). Empirically, existing AlphaFold-based predictors demonstrate some *ab initio* skill at predicting TCR-pMHC interaction but fall short of the high classification skill requisite for many practical applications (Bradley, 2023; Yin et al., 2023). Generating additional structure data through crystallography or cryo-EM is impractically expensive as a data-driven solution to further optimizing these models (Wang et al., 2021). In contrast, sequence-labeling assays determining the sequence but not the structure of interacting TCR-pMHC complexes are scalable and cost-effective (Klinger et al., 2015). A sequence-based model capable of improving through training on high throughput labeled sequence data would provide a powerful complement to structure prediction, especially for tasks like repertoire-level screening or safety evaluation, where millions of candidate interactions must be scored rapidly (June et al., 2018).

In this work, we introduce ImmSET (Immune Synapse Encoding Transformer), a transformer-based architecture designed for modeling interactions among sets of variable-length biological sequences with known or inferable roles while maintaining robustness for inference with partial inputs. Instead of first encoding each segment independently and only combining them at a later stage, ImmSET encodes sequences jointly from the outset, allowing the model to capture cross-segment dependencies throughout. This design avoids imposing a fixed contact graph or assuming a single "correct" structure, making it particularly well-suited for systems where function arises from dynamic and context-dependent interactions. To reduce biological variability while demonstrating a generalizable modeling paradigm, we focus on TCR-pMHC prediction in the exclusive context of HLA-A*02:01 – one of the most prevalent class I HLA alleles globally – and train on large, high-quality datasets of TCR-pMHC interactions. Despite being trained exclusively on A*02:01, ImmSET generalizes to other HLA contexts, capturing shared principles of TCR recognition. Our evaluation reveals a failure mode in prior sequence-based approaches that inflates reported metrics. We show that ImmSET remains robust under stricter evaluation, generalizing to peptides far outside the training distribution. We demonstrate predictable scaling with additional training data, with ImmSET consistently outperforming the pre-trained protein language model ESM2 (Lin et al., 2023) fine-tuned on the same datasets. Finally, ImmSET achieves over four orders of magnitude faster inference than structure-based pipelines while establishing state-of-the-art performance: it outperforms confidence-based AlphaFold2 and AlphaFold3 pipelines specialized for TCR-pMHC modeling (Bradley, 2023; Yin et al., 2023) on HLA-A*02:01 prediction tasks.

2. ImmSET Architecture

ImmSET is a transformer-based encoder and training framework designed to predict T cell activation directly from the TCR and pMHC amino acid sequences while scaling effectively with larger datasets. A central design goal is to encourage models to learn cross-chain dependencies – critical for pMHC recognition – while avoiding rigid inductive biases that may perform well in small data regimes but bottleneck scaling as datasets grow. Given that our understanding of TCR-pMHC biology remains incomplete and continues to evolve (Gray et al., 2025), ImmSET emphasizes a flexible, data-driven design that can adapt and improve as more labeled data becomes available. To achieve this goal, ImmSET introduces several novel design elements which contribute in total toward its performance on the TCR-pMHC task (Appendix H). A schematic overview of ImmSET is shown in Figure 1.

2.1. Auxiliary objectives for interaction learning

To encourage inter-sequence reasoning without imposing a fixed interaction graph, ImmSET incorporates two auxiliary objectives in addition to the primary classification loss. These auxiliary losses are applied only to positive (activation) examples to avoid introducing misleading correlations from negatives. Additional implementation details and training settings for these objectives are provided in Appendix A.

Masked Language Modeling (MLM): Individual amino acids within each chain are masked and predicted from their full multi-chain context. The MLM loss is computed separately for each chain, allowing chain-specific weighting so gradients from smaller chains (e.g., peptide) are not diluted by those from larger chains (e.g., HLA). This ensures that all chains contribute proportionally to their biological relevance.

Complete Segment Masking (CSM): Entire segments (e.g., the peptide) are masked and reconstructed from the remaining inputs, explicitly pressuring the model to learn cross-segment dependencies. This task is absent in existing pretrained protein language models, which focus solely on token-level recovery and do not enforce inter-segment integration. CSM is particularly important in cases where short, hypervariable regions interact with much longer, more conserved sequences. Without explicit pressure

to model these interactions, models can default to relying on features from a single dominant or stable segment, overlooking critical context provided by smaller, variable segments.

The overall objective function combines classification, MLM, and CSM losses as a weighted sum:

$$\mathcal{L} = \lambda_{\text{cls}}\,\mathcal{L}_{\text{cls}} + \lambda_{\text{mlm}}\,\mathcal{L}_{\text{mlm}} + \lambda_{\text{csm}}\,\mathcal{L}_{\text{csm}} \qquad (1)$$

2.2. Encoder, Order shuffling, and Integrated ensembling

ImmSET uses a single shared transformer encoder with segment-specific special tokens that delimit biologically meaningful regions. During training we also randomize the order of segments. These two choices serve complementary goals:

1. Prevent positional shortcuts: Order randomization forces the encoder to rely on token identity rather than fixed input slot.

2. Provides robustness to partial inputs during inference: In many datasets and applications, key inputs are missing (e.g., TCRα is often absent). The combination of special segment-specific tokens with shuffled input order during training makes the encoder naturally tolerant to absent sequences.

3. Enable clean implementation of auxiliary losses: Segment specific delimiter tokens allow for straightforward implementation of the per segment MLM losses and full segment masking for the CSM loss.

4. Enable future integration of partial datasets: The design provides a framework for incorporating auxiliary losses on incomplete inputs during training, a potential area of future research.

Predictions are produced from three independent CLS tokens (CLS1, CLS2, CLS3), each with its own learned embeddings and feed-forward classification head. The final prediction is obtained by averaging the three outputs, forming a lightweight, integrated ensemble that stabilizes training and reduces variance while sharing a single encoder.

2.3. Focused input for efficiency

To focus capacity on biologically relevant features, ImmSET encodes only the six complementarity-determining regions (CDRs) instead of full TCRα
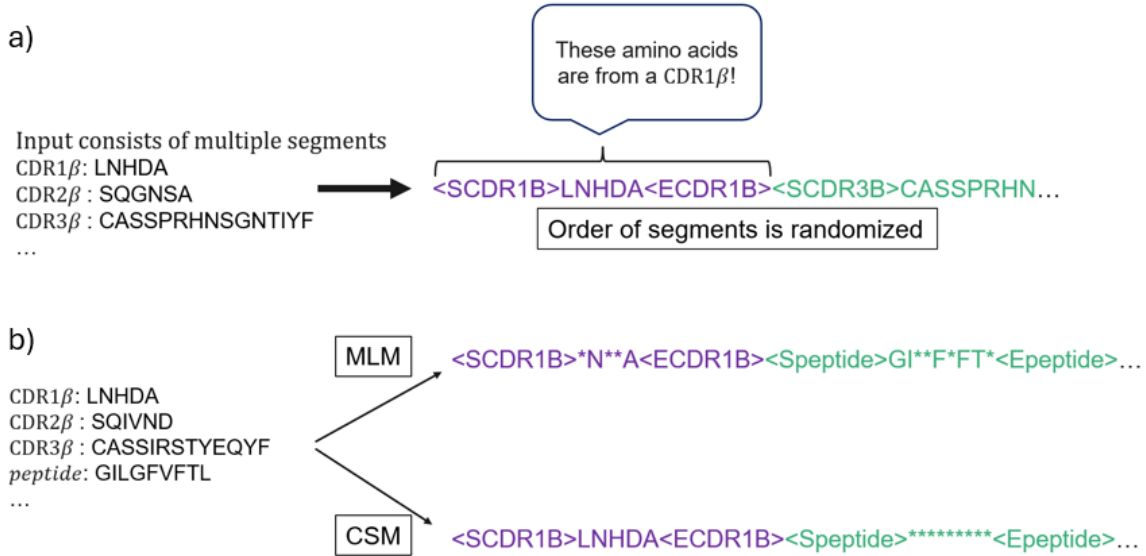
Figure 1: ImmSET architecture and training overview. (a) Segment-specific start and end tokens are used to differentiate the multiple input segments. (b) The MLM objective (above) masks random characters, while the CSM objective (below) masks entire segments.

and TCRβ chains. This decision leverages the flexibility of special tokens to define clear segment boundaries while reducing sequence length and compute.

As this study trains exclusively on A*02:01, we omit HLA representation reducing sequence length while retaining context critical for modeling.

3. Results

3.1. Failure modes for sequence-based models

Prior work has noted pitfalls in TCR–pMHC modeling such as label leakage or improper dataset partitioning which can inflate reported performance (Appendix B). In addition to those previously reported, we identify an additional failure mode that can also obscure true generalization:

Shortcut learning via TCR motif memorization. Models can appear artificially accurate by memorizing TCR–label associations rather than learning true interaction features. This arises because similar peptides often elicit overlapping TCR patterns (Gouttefangeas et al., 2023; Birnbaum et al., 2014). Even when performance is stratified by peptide edit distance (e.g., 1-Levenshtein splits), strong

signal on near-neighbor peptides that decays with distance can create a misleading impression of generalization. Low peptide diversity in the training data and architectures that encode TCR and peptide separately or emphasize local motif detection (e.g., CNNs) exacerbate this issue.
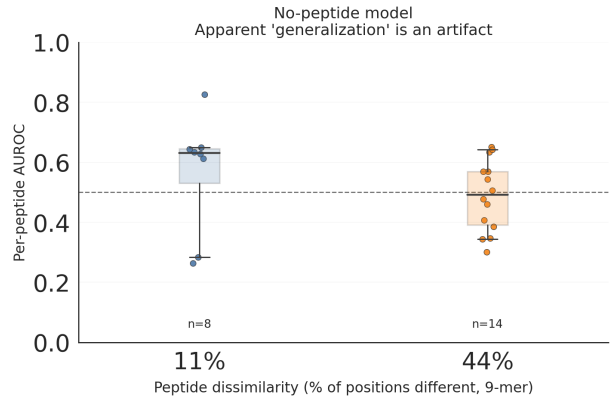

Figure 2: Performance of a model encoding only the TCR sequence and no peptide features.

To demonstrate this artifact, we trained a trans-former model on a small TCR-pMHC dataset with the peptide input entirely removed (details in Appendix C). Because peptide identity is the critical determinant of recognition, such a model should by construction have no predictive signal. Yet, as shown in Figure 2, the model still appears to "generalize" to unseen peptides at distance-1 neighbors ($\approx 11\%$ dissimilarity), while the signal disappears for distance-4 peptides ($\approx 44\%$ dissimilarity).

To mitigate this failure mode and avoid those previously reported, all subsequent evaluations are restricted to test peptides at least four Levenshtein distance away from any training or validation peptide to evaluate on truly unseen/dissimilar peptides, and draw negative labels only from within the limited set of the holdout pMHCs to prevent label leakage from train to test.

### 3.2. ImmSET performance scales predictably with training data

We sought to understand the relationship between training data and ImmSET performance. Identifying consistent patterns of how ImmSET improves with different quantities and compositions of training data would provide a valuable road map toward future data generation efforts. Such patterns would allow us to predict how much data is required for ImmSET to reach a given level of performance and inform what type of data is most impactful for driving better performance.

There are two main strategies for generating more data: identifying additional TCRs specific for a fixed set of peptides, or identifying TCRs to novel peptides. While generating additional TCRs on a fixed set of peptides is typically faster and cheaper, determining whether novel peptides are more impactful for model performance would motivate the time and cost required to generate that type of data.

To test the relative impact of each of these modes of data generation, we trained ImmSET using a single classification head across a full grid of training sets varying from 20 to 500 unique peptides and from 20 to 180 unique TCRs per peptide, all in the context of A*02:01. We repeated this scaling grid across 3 unique datasets comprising different random sub-samples of the full set of available peptides, termed "dataset 0", "dataset 1" and "dataset 2" in figure legends throughout this work (Appendix D).
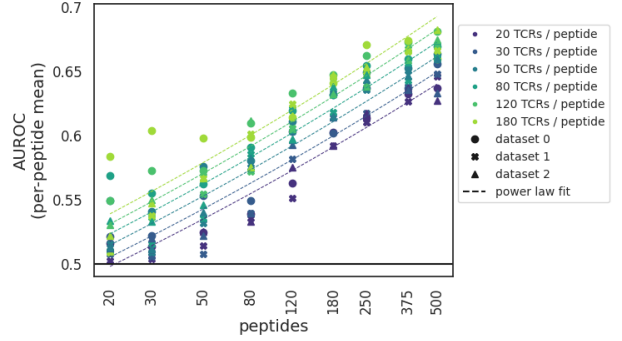
(a) Scaling with peptides
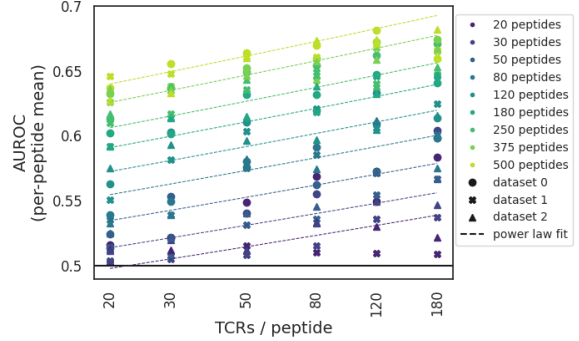
(b) Scaling with TCRs per peptide

Figure 3: ImmSET scaling behavior as a function of training data

After training separate ImmSET instances across all datasets, we then evaluated each of these trained models on a fixed holdout set of A*02:01-restricted peptides such that all holdout peptides are a minimum of 4 edits away from any training or validation example (Appendix D).

We observed that performance improves as a function of both the number of unique peptides and the number of TCRs per peptide (Figure 3). Both trends are significant (p < 1e-6 for both peptides; p < 3e-5 for TCRs, Jonckheere-Terpstra test) but models improve more quickly as a function of the number of peptides than they do as function of the number of TCRs per peptide. In addition to AUROC, these trends with training data hold across multiple performance metrics (Appendix F).

We fit a power law to the performance of all models, treating performance as a function of both the number of peptides (P) and the number of TCRs per peptide (T):

$$AUROC \propto P^{\alpha}T^{\beta} \qquad (2)$$

We find a best fit of $\alpha = 0.078$ (95% CI = (0.075, 0.082)) and $\beta = 0.036$ (95% CI = (0.031, 0.040)) (Appendix G). This fit reinforces the visual trend that while both peptides and TCRs per peptide benefits ImmSET performance, increasing peptides is more impactful.

These results suggest that ImmSET performance is a predictable function of the provided training data, quantified by the number of distinct peptides and the number of TCRs per peptide. Additionally, these results suggest that future data generation should focus on the collection of TCRs against new peptides as this data type provides a stronger lever toward increasing model performance.

3.3. Comparison with a pretrained sequence model (ESM2)

The ImmSET results presented thus far could be attributable to one or both of two fundamental advantages: its uniquely adapted architecture for this problem class, or its access to large training datasets. Given that the training data used in this work is from a proprietary dataset much larger than public data repositories of TCR-pMHC interactions, we compared ImmSET's performance as a function of training data to that of the well-established pretrained protein language model ESM2 (Lin et al., 2023) fine-tuned on the same proprietary data.
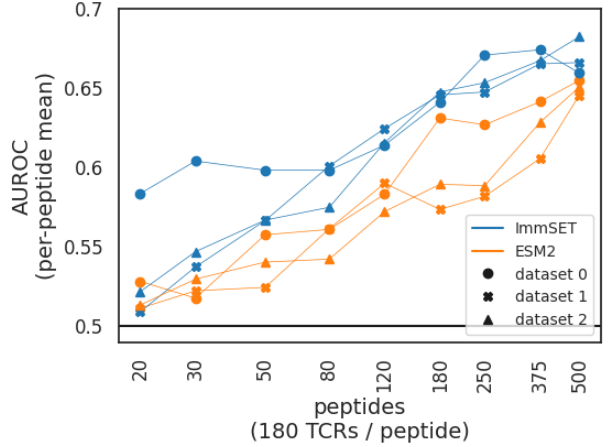
(a) Varying peptides
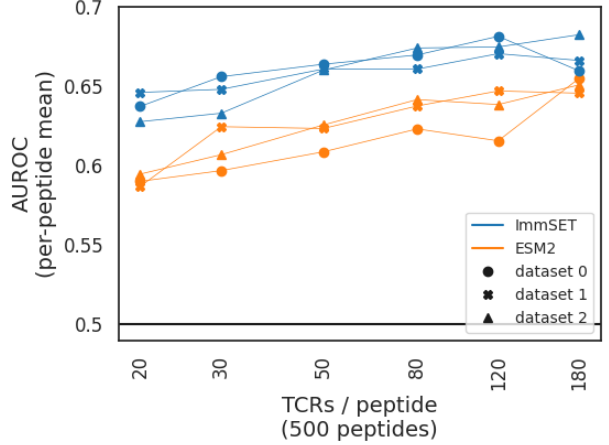
(b) Varying TCRs per peptide

Figure 4: Comparison of ImmSET vs ESM2 fine-tuned on the same datasets.

To compare model performance and scaling behavior, we compared the ImmSET models derived from the previous scaling analysis against similarly-sized ESM2 models (esm2_t6_8M_UR50D) fine-tuned across these same training data. Specifically, we fine-tuned ESM2 using all datasets with 180 TCRs per peptide (and any number of peptides) and all datasets with 500 peptides (and any number of TCRs per peptide). ImmSET training and architecture was as previously described. For ESM2, we initialized the model with the pre-trained model weights, added a sequence-level classification head, and trained the entire model on TCR-pMHC examples to learn a classification of activating and non-activating complexes (Appendix E).

We observed that both models learned an overall signal when trained across these datasets and both exhibit predictable scaling behavior as a function of the number of distinct peptides and the number of TCRs per peptide (Figure 4). However, ImmSET consistently achieves better overall performance with the same training data compared to ESM2 ($p < 5\mathrm{e}$-13, Wilcoxon test), while also retaining 4-fold faster inference time (Appendix M). This advantage holds across multiple performance metrics (Appendix F). As both models train on the same data, we attribute ImmSET's better overall performance to its architecture's adaptation for multi-segment interaction modeling. In addition to better overall performance, ImmSET achieves faster inference time (1.1ms per input ImmSET; 5.0 ms per input ESM2, measured on a single Nvidia T4 GPU) due to its parsimonious tokenization of the interacting chains in lieu of the full-length protein sequences used by ESM2.

The scaling of both models as a function of peptides and TCRs per peptide is comparable (Appendix G). Thus, ImmSET is not unique in its ability to achieve predictable performance on this task with training data but does achieve better performance with fixed training data than state-of-the-art pretrained protein language models like ESM2. Notably, these results were obtained with a single classification head from a single ImmSET instance, rather than our top ensemble configuration with three heads (see Appendix H for ablation study).

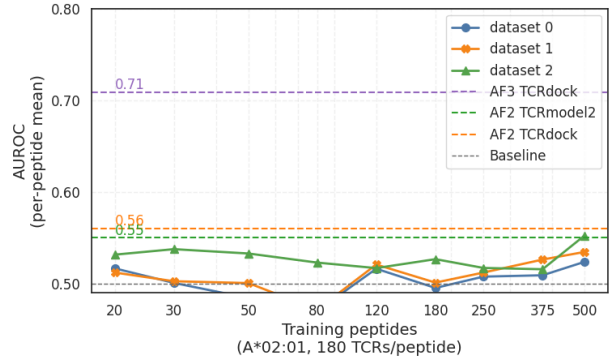### 3.4. ImmSET outperforms AlphaFold-based pipelines on A*02:01

We evaluated an ensemble variant of ImmSET (see Appendix I) against AlphaFold-derived pipelines
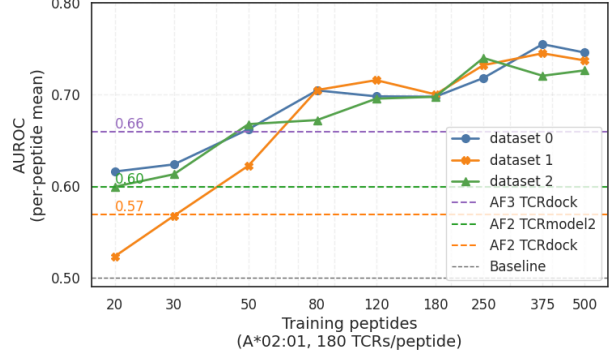
(a) A*02:01 evaluation, full input
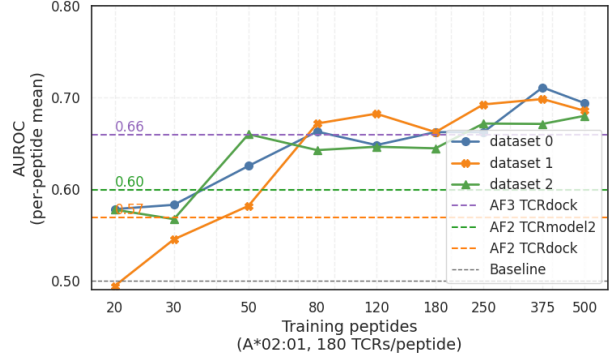
(b) A*02:01 evaluation, TCRα omitted

(c) B*40:01 evaluation, full input

Figure 5: ImmSET ensembles evaluated against AlphaFold on the IMMREP25 benchmark.

7

on a benchmark dataset previously used to compare TCR-pMHC modeling approaches in the IMMREP2025 competition (Organizers, 2025; Noakes et al., 2025) (Appendix J). As A*02:01-specific training data increases, ImmSET first surpasses AlphaFold2 pipelines and eventually overtakes AlphaFold3 on A*02:01 (Figure 5(a)). Inference of the ensemble model remains at ≈ 10 ms per TCR-pMHC on an NVIDIA T4 GPU - over four orders of magnitude faster than AlphaFold-based pipelines (Appendix M).

We further evaluate ImmSET's robustness to partial inputs by evaluating performance when information from TCRα is missing during inference, as arises often in real-world applications where bulk sequencing is used rather than single cell (Pai and Satpathy, 2021). Using only TCRβ and peptide and omitting TCRα, the top ImmSET models can still outperform AlphaFold-based pipelines on A*02:01 (Figure 5(b)), despite the latter using the full input.

These ImmSET models were trained exclusively on A*02:01 sequences, and their performance does not generalize to the untrained, dissimilar B*40:01 allele in the same IMMREP25 benchark. There, AlphaFold-based pipelines maintain a clear advantage (Figure 5(c)). Detailed understanding of the learning and scaling of ImmSET models trained and evaluated across multiple HLAs will require further research. However, given the importance of cross-HLA generalization to practical application of these models, we systematically test the generalization of A*02:01-trained models on more HLAs below.

3.5. Cross-HLA Generalization

We extended the evaluation of the A*02:01-trained ensemble models to other alleles following the same evaluation criteria as used for A*02:01: peptides at least four edits away from training peptides, a balanced sampling of 50 positives per peptides, and 200 negatives per peptide. Figure 6 shows as A*02:01-specific training data increases, ImmSET performance in other alleles increases as well, with further results in Appendix L. Notably, we find non-random signal to B*40:01, even though no such signal was observed for the B*40:01 subset of IMMREP25. To better understand this behavior, we examined the underlying per-peptide performance and, consistent with our observations in Appendix H, found that the distribution of per-peptide AUROCs is very broad, as summarized in Table 1.

Notably, the lowest per-peptide AUROC observed for B*40:01 was 0.336 and the highest 0.832. Since the IMMREP25 benchmark included only ten peptides, its results are inherently prone to variability, whereas our new evaluation is based on 67 peptides, providing a more stable estimate. Examining per-peptide AUROCs in IMMREP25, we find that for A*02:01 and B*40:01, ImmSET achieved ranges of (0.520-0.867) and (0.367-0.749), respectively, while the AF3-based TCRdock baseline achieved (0.562-0.793) for A*02:01 and (0.481-0.822) for B*40:01. These observations suggest that broad variability in per-peptide AUROC is a general phenomenon reflecting the inherent complexity of TCR-pMHC recognition, rather than a limitation specific to sequence-based models. They also highlight that conclusions drawn from small datasets should be interpreted with caution given their limited peptide diversity and sample size.

While the precise mechanisms driving cross-allelic generalization remain to be determined, a preliminary look into the biological interpretability of ImmSET is presented in Appendix N, where we analyze model responses to single amino acid substitutions as a first step toward mechanistic insight.

Moving forward, grouping alleles by similarity could help guide targeted data collection across representative HLA groups. Because ImmSET already encodes multiple biological segments within a unified architecture, it is naturally capable of incorporating the HLA sequence or pseudo-sequence as an additional input segment. We therefore believe that expanding training to include multi-allelic datasets, particularly spanning distinct HLA families, and incorporating the HLA sequence or pseudo-sequence as an input segment should allow the model to learn transferable, pan-allelic representations without requiring fundamental architectural changes. Understanding the learning and scaling across multiple HLAs will require further research.

4. Discussion

This study demonstrates ImmSET's capabilities in TCR–pMHC recognition, a challenging but critical task in immunology. Beyond this specific application, ImmSET introduces a modeling paradigm that is broadly applicable to other multi-sequence interaction problems—such as CRISPR–Cas targeting (Yao et al., 2025; Chen et al., 2025) or RNA–protein interactions networks (Lee et al., 2024)—where learn-

Table 1: Per peptide AUROC distribution is broad

| HLA	n_peptides	mean	median	percentile_25	percentile_75	min	max
A*01:01	37	0.607	0.619	0.552	0.694	0.348	0.818
A*02:01	134	0.706	0.704	0.639	0.772	0.393	0.959
A*29:02	32	0.644	0.619	0.542	0.761	0.423	0.887
A*68:01	42	0.649	0.636	0.554	0.702	0.446	0.939
B*08:01	18	0.579	0.554	0.493	0.631	0.374	0.908
B*18:01	46	0.624	0.632	0.553	0.700	0.129	0.853
B*40:01	67	0.600	0.601	0.505	0.671	0.336	0.832
C*03:04	18	0.675	0.736	0.547	0.774	0.356	0.948
C*04:01	15	0.567	0.540	0.483	0.680	0.386	0.693
C*07:02	27	0.636	0.627	0.550	0.736	0.312	0.857
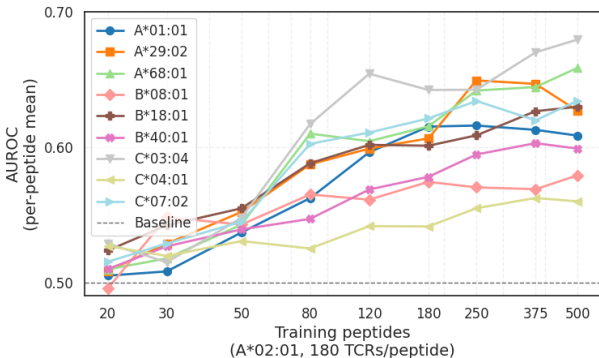
Figure 6: ImmSET cross-HLA generalization.

ing directly from large-scale sequence data is essential to complement structure-based and experimental approaches.

ImmSET exemplifies a function-first strategy: rather than relying on predefined structural assumptions, it leverages a shared transformer encoder with explicit sequence delimiters and auxiliary objectives that encourage cross-chain reasoning. This design enables the model to flexibly incorporate additional data and to capture interaction features even when structural details are uncertain or variable. In the context of immunology, our experiments show that ImmSET can not only match but surpass structure-based modeling while remaining orders of magnitude more efficient, and that its performance continues to improve predictably as training data increases. These scaling properties suggest a favorable trajectory as larger datasets become available.

While all ImmSET models in this study were trained exclusively on HLA-A*02:01 data, we observe measurable signal across multiple untrained class I HLAs. This demonstrates that cross-allelic generalization can emerge naturally from sequence-level learning, even without explicit exposure to multiple HLAs during training. ImmSET's multi-segment architecture is readily extensible to training on multi-HLA data as such datasets come available. We leave the exploration of scaling patterns across multiple HLAs to future work.

The choice of modeling strategy in TCR-pMHC applications should ultimately be guided by the requirements of the task. Structure-based pipelines are and will continue to be useful when training data are limited, data generation cost is high, inference cost is secondary, or moderate accuracy is sufficient for downstream analysis. However, applications that demand higher confidence or finer discrimination require not only much larger training datasets but also models that continue to improve as that data grows. In such settings, the ability to scale becomes decisive. Our results suggest that ImmSET offers a sequence-based path forward under this condition and that it can inspire hybrid approaches that combine the scalability of sequence-driven learning with the interpretability of structure-based pipelines.

Acknowledgments

References

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024.

Dmitriy V. Bagaev, Michael Izraelson, Eduard Ga"izarov, Konstantin Isaev, Alexey N. Davydov, Anastasiya V. Omelchenko, Rita Shaykhutdinova, Kirill Peskov, Pavel Volchkov, Elvira Khamaganova, Dmitriy Vinnikov, Dmitriy S. Shcherbinin, Olga V. Britanova, and Mikhail Shugay. Vdjdb in the pandemic era: a compendium of t cell receptors specific for sars-cov-2 and other viruses. *Nature Methods*, 19(9):1017–1019, 2022. doi: 10.1038/s41592-022-01557-0. URL https://vdjdb.cdr3.net.

Michael E Birnbaum, Juan L Mendoza, Dhruv K Sethi, Shen Dong, Jacob Glanville, Jessica Dobbins, Engin Özkan, Mark M Davis, Kai W Wucherpfennig, and K Christopher Garcia. Deconstructing the peptide-mhc specificity of t cell recognition. *Cell*, 157(5):1073–1087, 2014.

Philip Bradley. Structure-based prediction of t cell receptor: peptide-mhc interactions. *elife*, 12:e82813, 2023.

Ashley M Buckle and Natalie A Borg. Integrating experiment and theory to understand tcr-pmhc dynamics. *Frontiers in immunology*, 9:2898, 2018.

Stephen K. Burley et al. Updated resources for exploring experimentally-determined pdb structures and computed structure models at the rcsb protein data bank. *Nucleic Acids Research*, 53(D1):D564–D574, 2025. doi: 10.1093/nar/gkae1091. URL https://www.rcsb.org.

Jiyun Chen, Xiaofeng Lin, Wenwen Xiang, Ying Chen, Yueming Zhao, Linglong Huang, and Liang Liu. Dna target binding-induced pre-crrna processing in type ii and v crispr-cas systems. *Nucleic Acids Research*, 53(3):gkae1241, 2025.

Martin Culka, Nicolas W Lounsbury, William Thrift, Santrupti Nerli, Andrew Wallace, Gergő Nikolényi, Darya Orlova, Kiran Mukhyala, and Mohammed AlQuraishi. Predicting specificity of tcr-pmhc interactions using machine learning and biophysical models. *bioRxiv*, pages 2025–04, 2025.

Ceder Dens, Kris Laukens, Wout Bittremieux, and Pieter Meysman. The pitfalls of negative data bias for the t-cell epitope specificity challenge. *Nature Machine Intelligence*, 5(10):1060–1062, 2023.

Richard Evans, Michael O'Neill, Alexander Pritzel, Natasha Antropova, Andrew Senior, Tim Green, Augustin Žídek, Russ Bates, Sam Blackwell, Jason Yim, et al. Protein complex prediction with alphafold-multimer. *biorxiv*, pages 2021–10, 2021.

Cécile Gouttefangeas, Reinhild Klein, and Ana Maia. The good and the bad of t cell cross-reactivity: challenges and opportunities for novel therapeutics in autoimmunity and cancer. *Frontiers in Immunology*, 14:1212546, 2023.

George I Gray, P Chukwunalu Chukwuma, Bassant Eldaly, WWJ Gihan Perera, Chad A Brambley, Tatiana J Rosales, and Brian M Baker. The evolving t cell receptor recognition code: The rules are more like guidelines. *Immunological Reviews*, 329(1):e13439, 2025.

Scott A Hollingsworth and Ron O Dror. Molecular dynamics simulation for all. *Neuron*, 99(6):1129–1143, 2018.

Bryan Howie, Anna M Sherwood, Ashley D Berkebile, Jan Berka, Ryan O Emerson, David W Williamson, Ilan Kirsch, Marissa Vignali, Mark J Rieder, Christopher S Carlson, et al. High-throughput pairing of t cell receptor α and β sequences. *Science translational medicine*, 7(301):301ra131–301ra131, 2015.

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*, 23(8):511–521, 2023.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

Carl H June, Roddy S O'Connor, Omkar U Kawalekar, Saba Ghassemi, and Michael C Milone. Car t cell immunotherapy for human cancer. *Science*, 359(6382):1361–1365, 2018.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Mark Klinger, Francois Pepin, Jen Wilkins, Thomas Asbury, Tobias Wittkop, Jianbiao Zheng, Martin Moorhead, and Malek Faham. Multiplex identification of antigen-specific t cell receptors using a combination of immune assays and immune receptor sequencing. *PloS one*, 10(10):e0141561, 2015.

Bjørn PY 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*, pages 2023–04, 2023.

Yu Hsuan Lee, Evan P Hass, Will Campodonico, Yong Kyu Lee, Erika Lasda, Jaynish S Shah, John L Rinn, and Taeyoung Hwang. Massively parallel dissection of rna in rna–protein interactions in vivo. *Nucleic Acids Research*, 52(10):e48–e48, 2024.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379 (6637):1123–1130, 2023.

Gal Masrati, Meytal Landau, Nir Ben-Tal, Andrei Lupas, Mickey Kosloff, and Jan Kosinski. Integrative structural biology in the era of accurate structure prediction. *Journal of molecular biology*, 433 (20):167127, 2021.

Barthelemy Meynard-Piganeau, Christoph Feinauer, Martin Weigt, Aleksandra M Walczak, and Thierry Mora. Tulip: A transformer-based unsupervised language model for interacting peptides and t cell receptors that generalizes to unseen epitopes. *Proceedings of the National Academy of Sciences*, 121 (24):e2316401121, 2024.

Matthew T Noakes, Damon H May, Bryan Howie, and Harlan S. Robins. A Public Benchmark for TCR-pMHC Prediction. 10 2025. doi: 10.6084/m9.figshare.30491582.v1. URL https://figshare.com/articles/dataset/A_Public_Benchmark_for_TCR-pMHC_Prediction/30491582.

Sean Nolan, Marissa Vignali, Mark Klinger, Jennifer N Dines, Ian M Kaplan, Emily Svejnoha, Tracy Craft, Katie Boland, Mitchell W Pesesky, Rachel M Gittelman, et al. A large-scale database of t-cell receptor beta sequences and binding associations from natural and synthetic exposure to sars-cov-2. *Frontiers in Immunology*, 16:1488851, 2025.

Shane Nolan, Marissa Vignali, Mark Klinger, John N. Dines, Ivan M. Kaplan, Thomas Craft, Kevin Boland, Michael Roskey, Brian Martin, Thomas M. Snyder, and Harlan Robins. A large-scale database of t-cell receptor beta (tcr) sequences and binding associations from natural and synthetic exposure to sars-cov-2. *bioRxiv*, 2020. doi: 10.1101/2020.08.24.264630. URL https://www.biorxiv.org/content/10.1101/2020.08.24.264630v1. Preprint.

IMMREP25 Organizers. Immrep25: Tcr specificity prediction challenge. https://kaggle.com/competitions/immrep25, 2025. Kaggle.

Joy A Pai and Ansuman T Satpathy. High-throughput and single-cell t cell receptor sequencing technologies. *Nature methods*, 18(8):881–892, 2021.

Johannes Pettmann, Lama Awada, Bartosz Różycki, Anna Huhn, Sara Faour, Mikhail Kutuzov, Laurent Limozin, Thomas R Weikl, P Anton van der Merwe, Philippe Robert, et al. Mechanical forces impair antigen discrimination by reducing differences in t-cell receptor/peptide–mhc off-rates. *The EMBO Journal*, 42(7):e111841, 2023.

Roshan M Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. Msa transformer. In *International conference on machine learning*, pages 8844–8856. PMLR, 2021.

Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

Andrew K Sewell. Why must t cells be cross-reactive? *Nature Reviews Immunology*, 12(9):669–677, 2012.

Eilon Sharon, Leah V Sibener, Alexis Battle, Hunter B Fraser, K Christopher Garcia, and Jonathan K Pritchard. Genetic variation in mhc proteins is associated with t cell receptor expression biases. *Nature genetics*, 48(9):995–1002, 2016.

Jesús A Siller-Farfán and Omer Dushek. Molecular mechanisms of t cell sensitivity to antigen. *Immunological reviews*, 285(1):194–205, 2018.

Thomas M Snyder, Rachel M Gittelman, Mark Klinger, Damon H May, Edward J Osborne, Ruth Taniguchi, H Jabran Zahid, Ian M Kaplan, Jennifer N Dines, Matthew T Noakes, et al. Magnitude and dynamics of the t-cell response to sars-cov-2 infection at both individual and population levels. *Frontiers in Immunology*, 15:1488860, 2025.

Xiaoping Sun, Thomas Nguyen, Achouak Achour, Annette Ko, Jeffrey Cifello, Chen Ling, Jay Sharma, Toyoko Hiroi, Yongqing Zhang, Chee W Chia, et al. Longitudinal analysis reveals age-related changes in the t cell receptor repertoire of human t cell subsets. *The Journal of clinical investigation*, 132(17), 2022.

S Joshua Swamidass, Chloé-Agathe Azencott, Kenny Daily, and Pierre Baldi. A croc stronger than roc: measuring, visualizing and optimizing early retrieval. *Bioinformatics*, 26(10):1348–1356, 2010.

Naama Tickotsky, Tal Sagiv, Jaime Prilusky, Erez Shifrut, and Nir Friedman. Mcpas-tcr: A manually curated catalogue of pathology-associated t cell receptor sequences. *Frontiers in Immunology*, 8:459, 2017. doi: 10.3389/fimmu.2017.00459. URL https://friedmanlab.weizmann.ac.il/McPAS-TCR/.

Sebastiaan Valkiers, Nicky de Vrij, Sofie Gielis, Sara Verbandt, Benson Ogunjimi, Kris Laukens, and Pieter Meysman. Recent advances in t-cell receptor repertoire analysis: bridging the gap with multimodal single-cell rna sequencing. *ImmunoInformatics*, 5:100009, 2022.

Randi Vita, Nadica Blazeska, Diana Marrama, IEDB Curation Team Members, Steven Duesing, John Bennett, Jason Greenbaum, Monica De Almeida Mendes, Jyoti Mahita, Daniel K. Wheeler, Jonathan R. Cantrell, John A. Overton, Darren A. Natale, Alessandro Sette, and Bjoern Peters. The immune epitope database (iedb): 2024 update. *Nucleic Acids Research*, 53(D1): D436–D443, 2025. doi: 10.1093/nar/gkae1092. URL https://www.iedb.org.

Zuoneng Wang, Qingyang Zhang, and Carsten Mim. Coming of age: Cryo-electron tomography as a versatile tool to generate high-resolution structures at cellular/biological interfaces. *International journal of molecular sciences*, 22(12):6177, 2021.

Na Xie, Guobo Shen, Wei Gao, Zhao Huang, Canhua Huang, and Li Fu. Neoantigens: promising targets for cancer therapy. *Signal transduction and targeted therapy*, 8(1):9, 2023.

Zhihao Yao, Wanglu Li, Kaiyu He, Hongmei Wang, Yan Xu, Qun Wu, Liu Wang, and Yao Nie. Facilitating crrna design by integrating dna interaction features of crispr-cas12a system. *Advanced Science*, page 2501269, 2025.

Rui Yin, Helder V Ribeiro-Filho, Valerie Lin, Ragul Gowthaman, Melyssa Cheung, and Brian G Pierce. Tcrmodel2: high-resolution modeling of t cell receptor recognition using deep learning. *Nucleic Acids Research*, 51(W1):W569–W576, 2023.

Appendix A. ImmSET Training

A.1. Tokenization

Each of the 20 amino acids is assigned a unique token. In addition, every sequence segment is delimited by dedicated start and end tokens. A padding token is included for batch alignment; this token is excluded from both inference and loss computation. Finally, up to three classification (CLS) tokens are appended, each serving as an anchor for downstream prediction. Together, these tokens define the model vocabulary.

A.2. Parameters

We used a transformer encoder with Rotary Positional Embeddings (RoPE). The architectural parameters are listed in Table 2. The classifier is implemented as a feed-forward neural network (FFNN) applied to the output corresponding to the CLS token; when multiple CLS tokens are present, each is paired with its own FFNN classifier. Parameters were initialized using PyTorch's default initialization. The masked language modeling (MLM) and cross-segment modeling (CSM) heads each consist of a single linear layer mapping from `embed_dim` to the vocabulary size defined in the previous section.

A.3. Optimization

The overall objective combines classification, MLM, and CSM losses as a weighted sum:

$$\mathcal{L} = \lambda_{\text{cls}} \, \mathcal{L}_{\text{cls}} + \lambda_{\text{mlm}} \, \mathcal{L}_{\text{mlm}} + \lambda_{\text{csm}} \, \mathcal{L}_{\text{csm}}. \qquad (3)$$

Table 2: Transformer and classifier parameters used in ImmSET.

Parameter	Value
Embedding dim.	420
Attention heads	5
Hidden dim.	420
Encoder layers	10
Dropout	0.01
CLS classifier	128
	128
	64
	32
	1
MLM/CSM head	Linear (`embed_dim` → #tokens)

Table 3: Loss weighting scheme stability.

MLM	CSM	CLS	mean_AUROC
1	1	1	0.616
1	4	4	0.656
1	4	16	0.684
1	4	48	0.680
1	3	55	0.675
1	5	60	0.673
1	4	64	0.674
1	6	72	0.675
1	4	80	0.679
1	4	256	0.668

We set $\lambda_{\mathrm{cls}} = 64$, $\lambda_{\mathrm{csm}} = 4$, and $\lambda_{\mathrm{mlm}} = 1$ for all experiments and the weights stayed robust across the tested range of dataset sizes and compositions. The large weight on the classification loss is not merely to downweight auxiliary losses but to counteract the tendency of the model to overfit the comparatively easy MLM and CSM task. Since no grid search was done to select these values, to further assess the stability of the weighting scheme, we extended training with dataset 0 with 500 peptides and 180 TCRs per peptide, using multiple losses weights configurations and evaluating on holdout, results are summarized in Table 3. The results confirm that the model's performance is generally robust to moderate changes in the auxiliary loss weights only significantly decreasing performance under large changes like using all identical weights. Results also suggest hyperparameter grid search could provide a modest but reproducible improvement in performance. The classification loss used is binary cross entropy.

The MLM loss is cross entropy, computed per segment and summed across segments:

$$\mathcal{L}_{\mathrm{mlm}} = \sum_{\mathrm{segments}} \mathcal{L}_{\mathrm{mlm}}(\mathrm{segment}). \tag{4}$$

A fixed masking probability of 0.3 was used for all amino acids. For a selected position, the replacement strategy was: mask token (0.8), random token substitution (0.1), or unchanged (0.1).

The CSM loss is defined by replacing an entire segment with mask tokens:

$$\mathcal{L}_{\mathrm{csm}} = \sum_{\mathrm{segments}^*} \mathcal{L}_{\mathrm{csm}}(\mathrm{segment}^*), \tag{5}$$

where segments* merges CDR1α with CDR2α, and CDR1β with CDR2β, reflecting their common V-gene origin and lower relevance for peptide correlation compared to full gene–peptide interactions.

Optimization used Adam (Kingma and Ba, 2014) with $\beta = (0.9, 0.999)$ and weight decay 0.01. Learning rate followed a schedule with linear warmup to the peak value, followed by exponential decay clipped at 1% of the initial rate.

A.4. Data Sampling

We used a batch size of 256 with a dataloader size of 12,800, yielding 50 batches per epoch between validation steps. If the number of positives was smaller than 12,800, the closest multiple of 256 was used. Each epoch samples from the full training set with enforced uniform sampling across peptides and 1:1 matching of positives to negatives, since negatives vastly outnumber positives.

A.5. Checkpoints

For holdout evaluation, we selected the checkpoint maximizing a weighted metric combining AUROC and AUCROC (area under the concentrated ROC curve (Swamidass et al., 2010), with $\alpha = 14$ in our implementation). Specifically, we optimized:

$$\mathrm{score} = \mathrm{AUROC} + 5 \times \mathrm{AUCROC}.$$

For ensembles, three checkpoints per training run were retained, corresponding to the top three scores during training.

Appendix B. Pitfalls in TCR–pMHC Sequence Modeling

A wide range of sequence-based models have been proposed for TCR–pMHC prediction, including convolutional networks, transformers, and pre-training strategies. Despite this activity, models only appear to succeed under favorable conditions, and none have demonstrated reliable generalization to unseen peptides.

B.1. Inflated performance from negative data design

Reported accuracy often depends heavily on how negatives are constructed (Dens et al., 2023). Shuffling TCRs and peptides within the same dataset, or using repertoires from healthy individuals as negatives, can introduce leakage signals that models exploit instead of learning recognition rules. STAPLER explicitly demonstrated how such leakage can confound evaluation and proposed best practices to mitigate it (Kwee et al., 2023).

B.2. Benchmark overlap and limited peptide coverage

Several community benchmarks contain overlap between training and test sets at the level of peptides or TCRs. Such designs allow models to rely on memorization within distribution, rather than testing the intended challenge of unseen peptide prediction. The Genentech/Columbia study systematically examined these issues and showed that prior claims of strong generalization were largely due to flawed split design (Culka et al., 2025).

B.3. Sensitivity to negative sampling

Supervised classifiers are highly sensitive to how negatives are chosen. For example, if TCRs are used as negatives in the test set but appear as positives in the training set, models can trivially learn to classify those TCRs as always negative, introducing a strong bias (Kwee et al., 2023).

B.4. Failure to generalize to unseen peptides

The central conclusion across recent work is that no current method achieves robust peptide-level generalization. The Genentech study evaluated models on a blinded, proprietary dataset of cancer neoantigens and found that even methods designed to correct for leakage and bias (e.g., TULIP (Meynard-Piganeau et al., 2024)) performed no better than chance in this zero-shot setting (Culka et al., 2025).

Both STAPLER and TULIP further show that model performance consistently decays toward baseline as peptide distance from the training set increases (Kwee et al., 2023; Meynard-Piganeau et al., 2024). In the main text, we also describe an additional failure mode: shortcut learning through TCR motif memorization. Models that effectively ignore the peptide can appear to generalize to unseen peptides that are close to the training distribution, but their signal collapses with increasing peptide distance. This creates an illusion of generalization and underscores the need for carefully designed training, validation, and holdout datasets.

Appendix C. No-peptide model

To illustrate the apparent generalization effect observed when evaluating on closely-related peptides (as described in the main text), we trained a transformer with the same architecture as ImmSET (Appendix A), but with the peptide sequence excluded from the input. The model received only the six CDR regions in fixed order, separated by SEP tokens, with a CLS token prepended and a classifier applied to the CLS output. Training used a randomly chosen dataset comprising 20 peptides with 50 TCRs per peptide. Data were split into training, validation, and holdout sets, with strict enforcement of non-overlap across peptides and TCRs—including negatives—to prevent leakage as noted in Appendix B.

Appendix D. Dataset details

D.1. Data generation

The labeled data used in both training and evaluation across this work are the amino acid sequences of TCR-pMHC combinations that do (positive labels) and do not (negative labels) result in T cell activation. The data was generated through a combination of immunological assays previously described

in other work. Briefly, naïve CD8+ T cells are incubated against a pool of peptides to drive T cell clonal expansion. A portion of this expanded T cell population is input into a pairSEQ experiment to determine the TCRα–TCRβ pairings as described in Howie et al. (2015). The remainder of the expanded T cell sample is input into the MIRA assay (Multiplex Identification of T cell Receptor-Antigen specificity), returning mappings of TCRβ sequences to their pMHC specificity as described in Klinger et al. (2015); Nolan et al. (2025) (experimental methods) and Snyder et al. (2025) (statistical methods). Joining the TCRα–TCRβ pairings (from pairSEQ) with TCRβ-pMHC mappings (from MIRA) on the TCRβ sequence then provides complete TCR-pMHC sequence information.

In addition to the positive TCR-pMHC combinations generated by these assays (TCR-pMHC combinations leading to T cell activation), the experiments also provide empirical negative labels (TCR-pMHC combinations confirmed not to lead to T cell activation). These negative labels are derived from the multiplexed layout of the MIRA assay (Figure 7). TCRβ clonotypes are assessed for activation against all peptides in the experiment, and as they are only reported as a positive for a given peptide when their activation pattern corresponds exactly to the unique occupancy pattern of that peptide. Consequently, any TCRβ reported as a positive to one peptide must necessarily not activate in response to any other peptides in the experiment.

D.2. Datasets for scaling experiments and ESM2 comparison

Training datasets   We constructed a grid of nested training datasets to study the general scaling behavior of ImmSET and to compare ImmSET to ESM2 across a range of training data compositions.

To assess variation of model performance and scaling across different random subsets of the data, we created 3 replicate training grids referred to as "dataset 0", "dataset 1", and "dataset 2" across this study. Each grid is constructed as follows:

Starting from a large proprietary database of HLA-A*02:01 restricted peptides, we randomly sample 500 unique peptides from this database such that all selected peptides had a minimum of 180 labeled TCRs. For each of those 500 peptides, we then randomly sampled 180 positive TCRs to each of them, giving us an overall dataset of 500 peptides by 180 TCRs each.

From that starting point, we iteratively downsampled the peptides from 500 to 375, then from 375 to 250, and so on until we had nested sets of peptides ranging in count from 500 down to 20. In total, we sampled 9 unique peptide counts, roughly spaced by a factor of 1.5x relative to their neighbor: 20, 30, 50, 80, 120, 180, 250, 375, and 500. At each peptide downsampling, we retain the full initial set of 180 TCRs per peptide.

Likewise, we conduct a similar nested downsampling per peptide on its full set of 180 TCRs. From that starting set of 180 TCRs, we selected 120 TCRs, then 80 from the downsampled set of 120, and so on until we reached a minimum size of 20 TCRs for the peptide. In total, we sample 6 unique counts of TCRs for each peptide: 20, 30, 50, 80, 120, and 180, again roughly spaced by a factor of 1.5x relative to their neighbors.

We then cross each variably sized peptide subset with each of the variably sized TCR-per-peptide subsets to create complete grid of training sets varying in both the number of peptides and the number of TCRs per peptide (Figure 8). This procedure yields a grid of training sets such that each individual set S_i (characterized by a number of peptides P_i and of TCRs per peptide T_i) is a strict subset of all other sets S_j provided that $P_j \geq P_i$ and $T_j \geq T_i$.

With these 3 replicate grids of 54 training sets each, we ultimately trained ImmSET across 162 total unique TCR by peptide training compositions. For ESM2, we trained across all 3 replicate datasets, but only using those sets that included either 500 peptides (any number of TCRs) or 180 TCRs per peptide (any number of peptides), for a total of 42 training compositions.

Validation dataset   All training runs – both for ImmSET and ESM2 – used the same set of validation data to select the final model checkpoint for evaluation. The validation set constructed by selecting 115 unique HLA-A*02:01-restricted peptides such that each validation peptide was 4 or more levenshtein distance away from any training peptide in any training dataset. Each selected validation peptide had a minimum of 50 unique positive labeled TCRs. From these data, we randomly subsampled 50 positive TCRs per peptide. Finally, we subsampled the negative label data to achieve a ratio of 4 negative examples for each positive example.
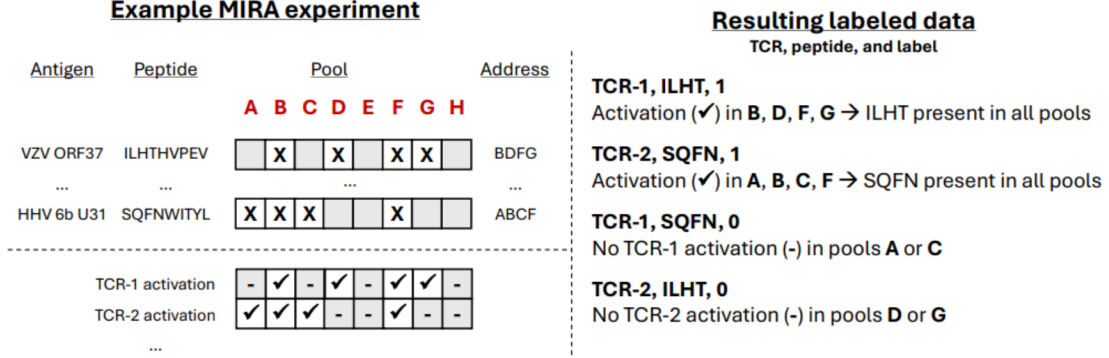
Figure 7: Derivation of positive and negative labels from MIRA experiments. Each peptide in the the experiment is present in a unique combination of the lettered pools (upper left). The activation patter of each TCR tested in the experiment is read out as activated (check mark) or not activated (dash) in each of the letted pools (lower left). Matching these activation patterns to the peptide pool assignments provides both positive and negative evidence of TCR-pMHC specificity (right).

Holdout dataset In both the ImmSET scaling analysis and the ESM2 comparison analysis, all models were evaluated on a consistent holdout dataset. The holdout set was constructed from 134 unique HLA-A*02:01-restricted peptides such that all holdout peptides were 4 or more Levenshtein away from all peptides across all training datasets and the shared validation dataset. As with the validation data, we used only peptides with a minimum of 50 positive labeled TCRs and subsampled 50 positive TCRs per peptide. Matching the composition of the validation set, we also subsample the negative label data to achieve the same 4:1 ratio of negative to positive labels.

D.3. Dataset Composition

The training and evaluation dataset used in this study are proprietary but can be summarized in terms of their composition and diversity. Table 4 provides an overview of the five A*02:01 datasets employed: three independent training datasets, one validation dataset, and one holdout dataset. For each dataset, we report the total number of unique peptides, the distribution of peptide lengths, and the number of peptide clusters. Peptide clusters are defined as the connected components of a graph with nodes corresponding to peptides and edges between any pair of peptides with an edit distance below
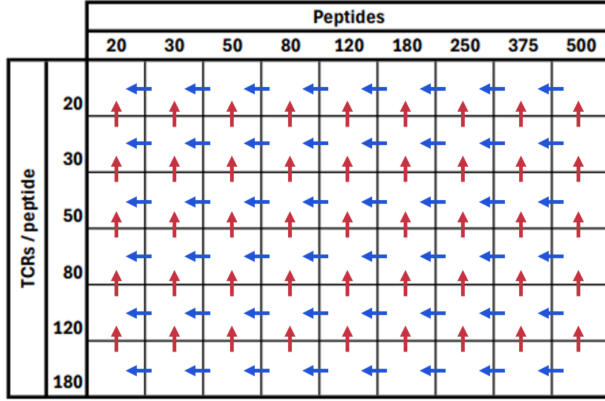

Figure 8: Format of a single training grid. The blue arrows show steps down in the number of peptides and red arrows show steps down in the number of TCRs. Each individual set is a strict subset of all parents – both those to the right in the grid (more peptides) and below in the grid (more TCRs).

16

a given threshold. Here, we define graphs using a threshold of 1 edit, 2 edits, and 3 edits.

Table 5 describes the composition of the same 5 datasets from the perspective of their constituent TCRs. We report the number of positive TCRs per peptide, the total number of positive and negative records, the ratio of negative to positive examples, and the number of unique TCR sequences in each dataset. We also provide the number of unique TCRα V gene / TCRβ V gene combinations in each dataset as a broad descriptor of the CDR diversity present.

Finally, we report in Table 6 the median and 10th-to-90th percentile range of CDR lengths for each of the 6 CDRs.

Appendix E. ESM2 fine-tuning experiments

We fine-tuned ESM2 (Lin et al., 2023) to predict binary classification of TCR-pMHC complexes resulting in (positive label) and not resulting in (negative label) T cell activation as follows.

We initialize a sequence-level classification model to predict a binary label across the entire input sequence using the HuggingFace `EsmForSequenceClassification` class, taking the pretrained weights of the 8 million parameter ESM2 model (esm2_t6_8M_UR50D) as a starting point.

The TCR-pMHC sequences are input to the model during training as follows. At each training step, we take the full-length amino acid sequence of the variable domains of the TCRα and TCRβ chains, as well as the peptide (MHC sequence is omitted as we train only in the single context of A*02:01). We randomly order these 3 chains at each training step and link them using poly-G sequences (length 20). All inputs longer than 300 tokens are randomly cropped to a maximum input length of 300 tokens, with shorter inputs padded out to the same length. Across the training, validation, and holdout data used in this study, all inputs are below the 300 token crop size limit – the crop size limit is only retained in case of future prediction on rare but biologically possible extremely long CDR3 sequences in TCRα and TCRβ.

During training, we update all weights of the ESM2 classification model except for those of the contact head as we are not training or using the model for direct structure prediction.

We train the model to minimize the binary cross-entropy loss of its classification prediction against the binary labels using the Adam optimizer (Kingma and

Table 4: Dataset composition: peptide lengths and edit-distance clusters.

Dataset	Peptides	# 9-mers	# 10-mers	# 11-mers	# clusters (1 edit)	# clusters (2 edits)	# clusters (3 edits)
Train-0	500	453	41	6	417	399	390
Train-1	500	452	43	5	421	406	399
Train-2	500	453	41	6	413	398	392
Validation	115	110	5	0	115	115	115
Holdout	134	125	7	2	103	89	87

Table 5: Dataset composition: TCR sampling and record statistics.

Dataset	TCRs / peptide	Positive records	Negative records	Neg:Pos Ratio	Unique TCRs	TCRαV/TCRβV combinations
Train-0	180	90000	5349353	59	89935	1583
Train-1	180	90000	5375072	60	89935	1578
Train-2	180	90000	5201633	58	89914	1572
Validation	50	5750	23000	4	5750	1071
Holdout	50	6700	26800	4	6692	1104

Ba, 2014), with a batch size of 128 and learning rate of $1E-4$. Each batch samples an equal number of positive and negative labeled examples. We evaluate the model against the validation set every 500 training steps. We select the final model checkpoint for holdout evaluation using the same criterion as for the ImmSET training runs: choosing the checkpoint maximizing the validation set score of

$$score = AUROC + 5 \times AUCROC.$$

As for the ImmSET criterion, AUCROC is calculated using a transformation parameter $\alpha = 14$.

Appendix F. Additional performance metrics

In addition to AUROC, we evaluated the performance of both ImmSET and ESM2 across the many training runs by their average precision (AP), sensitivity at 90% specificity (SENS_AT_90), and area under the concentrated ROC (AUCROC) (Swamidass et al., 2010). In the AUCROC calculations, we used a transformation parameter $\alpha = 14$.

These additional evaluation metrics emphasize model sensitivity at high specificity. This is a critical regime for many applications of TCR-pMHC specificity prediction models, where the model must confidently identify a small number of positive combinations out of a large set of negative possibilities.

The ImmSET scaling results under these additional performance metrics (Figure 9) show the same trend observed in the AUROC analysis: performance improves monotonically with both the number of peptides and the number of TCRs per peptide, though the growth is more rapid as a function of the number of peptides.

Likewise, the comparison of ESM2 to ImmSET across these additional metrics shows the same trend as the AUROC analysis. ImmSET maintains a consistent performance gap over the fine-tuned ESM2 model across all metrics. Both models improve monotonically with increasing data and show a stronger dependence on peptides than TCRs per peptide under all 3 additional metrics.

Table 6: CDR length statistics per dataset (median and range).

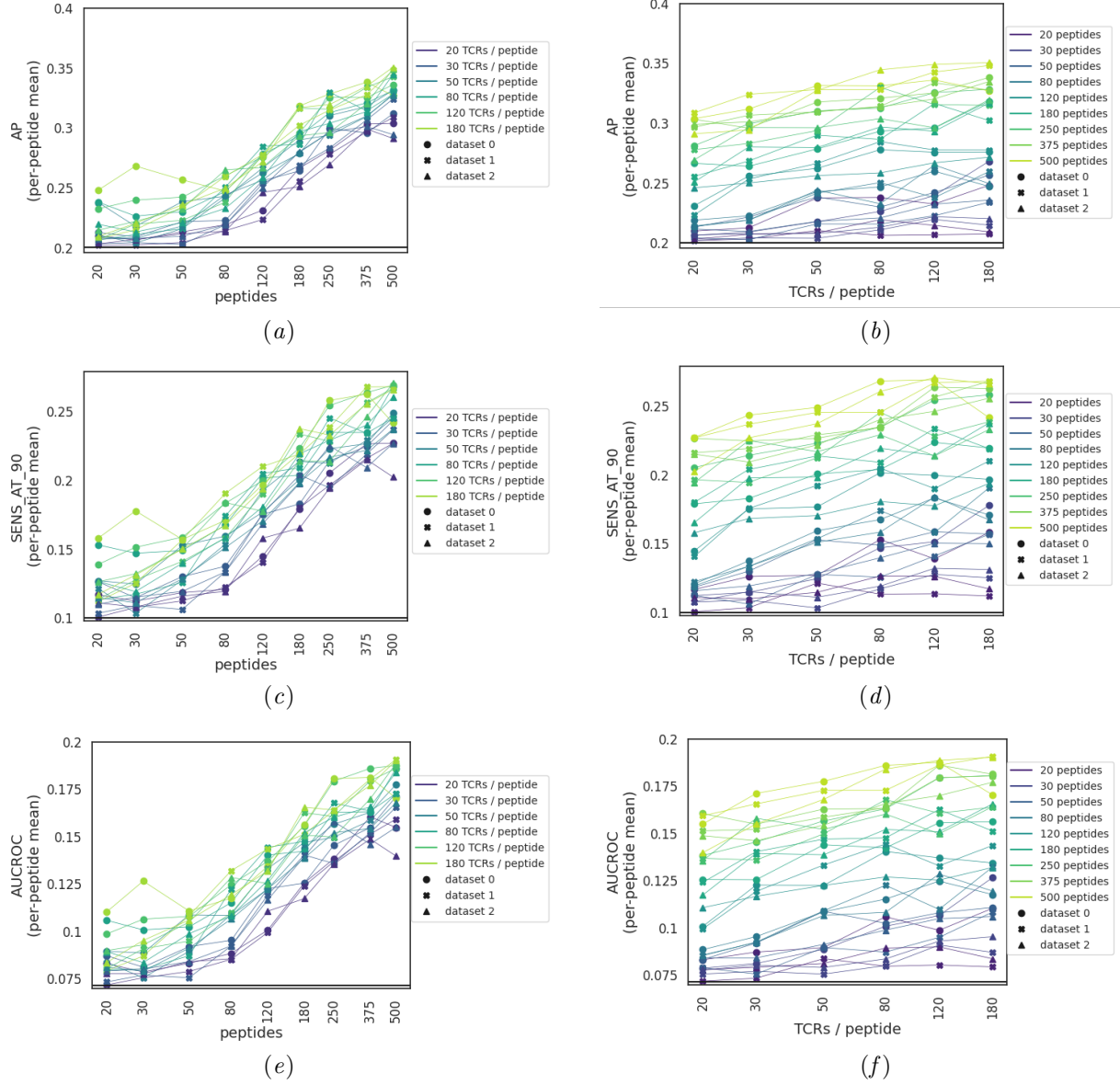| Dataset	CDR1α	CDR2α	CDR3α	CDR1β	CDR2β	CDR3β
Train-0	6 (5–7)	7 (6–8)	13 (11–16)	5 (5–5)	6 (6–6)	14 (12–16)
Train-1	6 (5–7)	7 (6–8)	13 (11–16)	5 (5–5)	6 (6–6)	14 (12–16)
Train-2	6 (5–7)	7 (6–8)	13 (11–16)	5 (5–5)	6 (6–6)	14 (12–16)
Validation	6 (5–7)	7 (6–8)	13 (11–16)	5 (5–5)	6 (6–6)	14 (12–16)
Holdout	6 (5–7)	7 (6–8)	13 (11–16)	5 (5–5)	6 (6–6)	14 (12–16)

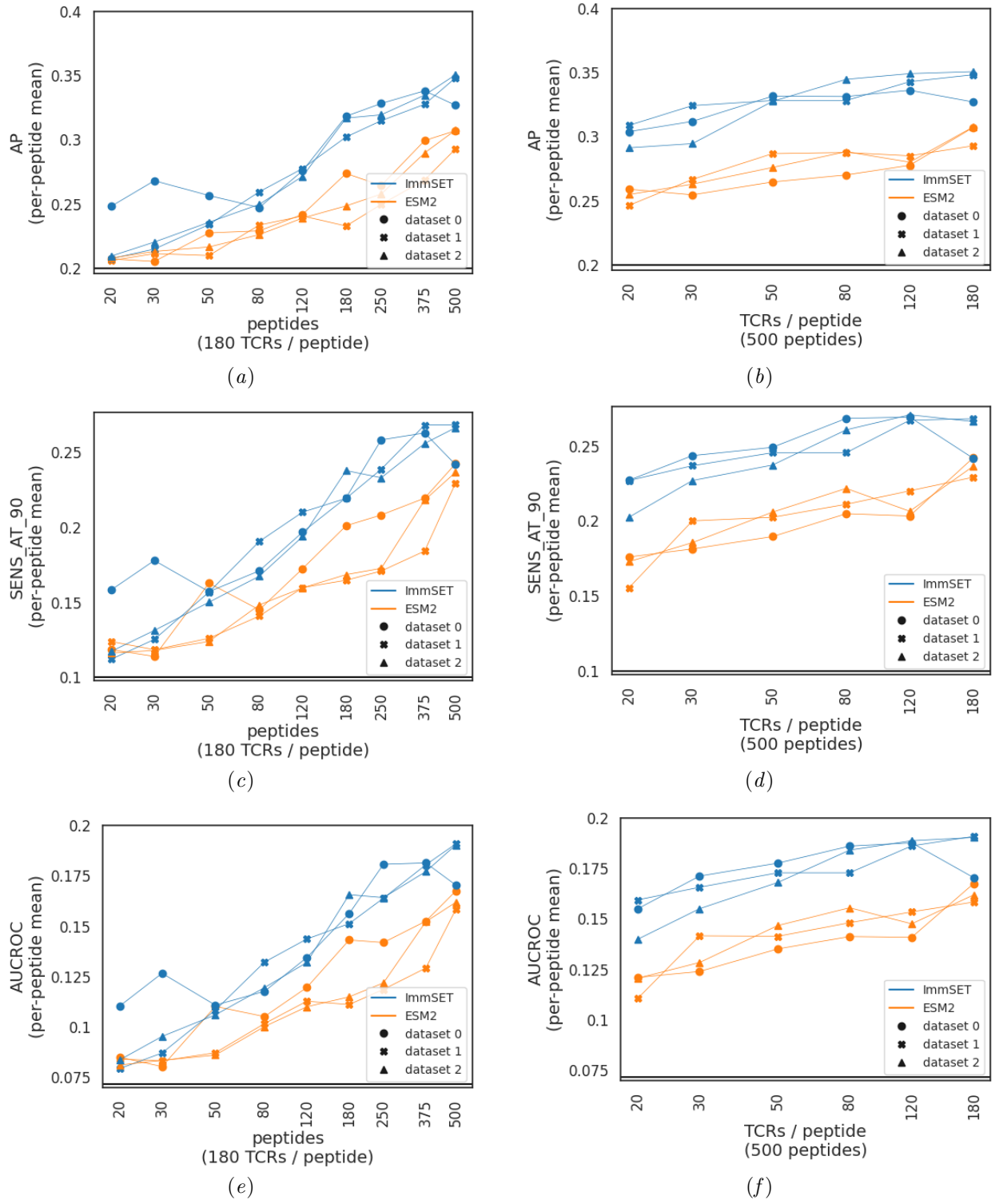Figure 9: ImmSET scaling analysis under different performance metrics

Figure 10: Comparison of ImmSET and ESM2 under different performance metrics

Appendix G. ESM2 and ImmSET scaling analysis results and parameters

To fit the trends of AUROC variation in both ESM2 and ImmSET as a function of dataset size and composition, we modeled performance as a power law in both the number of training peptides P and the number of TCRs per peptide T:

$$AUROC \propto P^{\alpha} T^{\beta} \qquad (6)$$

The α parameter describes how much AUROC improves with increasing peptides in the training set, while β describes the AUROC improvement with increasing TCRs per peptide.

For each model, we fit this trend across all available trained model results (N = 162 for ImmSET, comprising 9 values of P crossed with 6 values of T, in 3 replicates, and N = 42 for ESM2, comprising only the extrema of the training data grid, also in 3 replicates). The fits are obtained using a linear mixed effects model via the python `statsmodels` library (Seabold and Perktold, 2010), fitting the log of AUROC as a linear combination of the log of T and the log of P:

$$\log(\text{AUROC}) = c + \alpha \log P + \beta \log T + \varepsilon \qquad (7)$$

The parameters of the resulting fits are reported, with 95% confidence intervals in Table 7 and visualized in Figure 11. For both ESM2 and ImmSET, scaling with both T and P is significantly nonzero. Both models also exhibit stronger scaling with P than T, with α more than twice the value of β in both cases. The scaling parameters of the two models overlap substantially in their confidence intervals and – based on the experimental data available to this study – both models appear to scale comparably.

Figure 11: Best fits and 95% confidence intervals of scaling parameters for both ESM2 and ImmSET

Appendix H. Ablation study

We validated ImmSET's design decisions through a peptide-level bootstrap ablation study (Figure 12), where each variant is trained and evaluated using the full ImmSET tokenization scheme (special tokens delimiting peptide, CDRs, and HLA regions). Notably, we can see from the ablation study the importance of randomization for running inference with partial inputs when using the auxiliary losses.
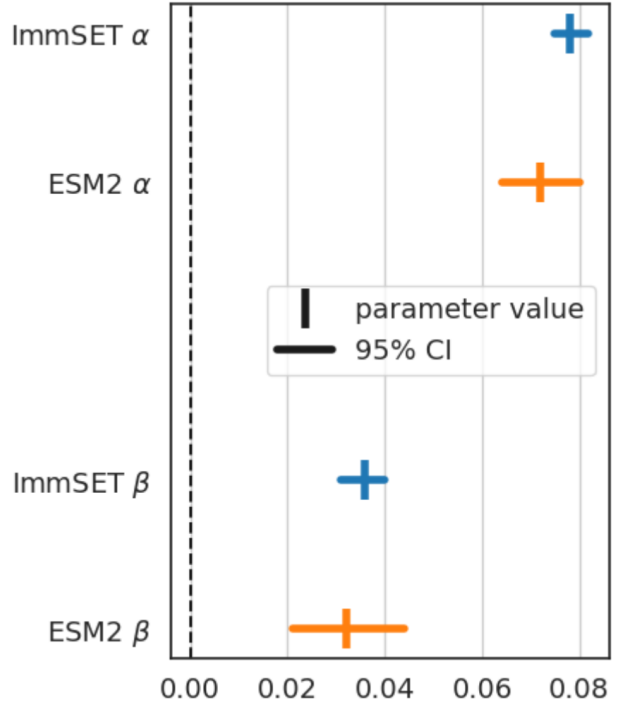
Table 7: We report the parameters of a best-fit power law relating AUROC to training data size and composition, quantified by P (number of peptides) and T (number of TCRs per peptide) for both ImmSET and ESM2, along with 95% confidence intervals of each scaling parameter.

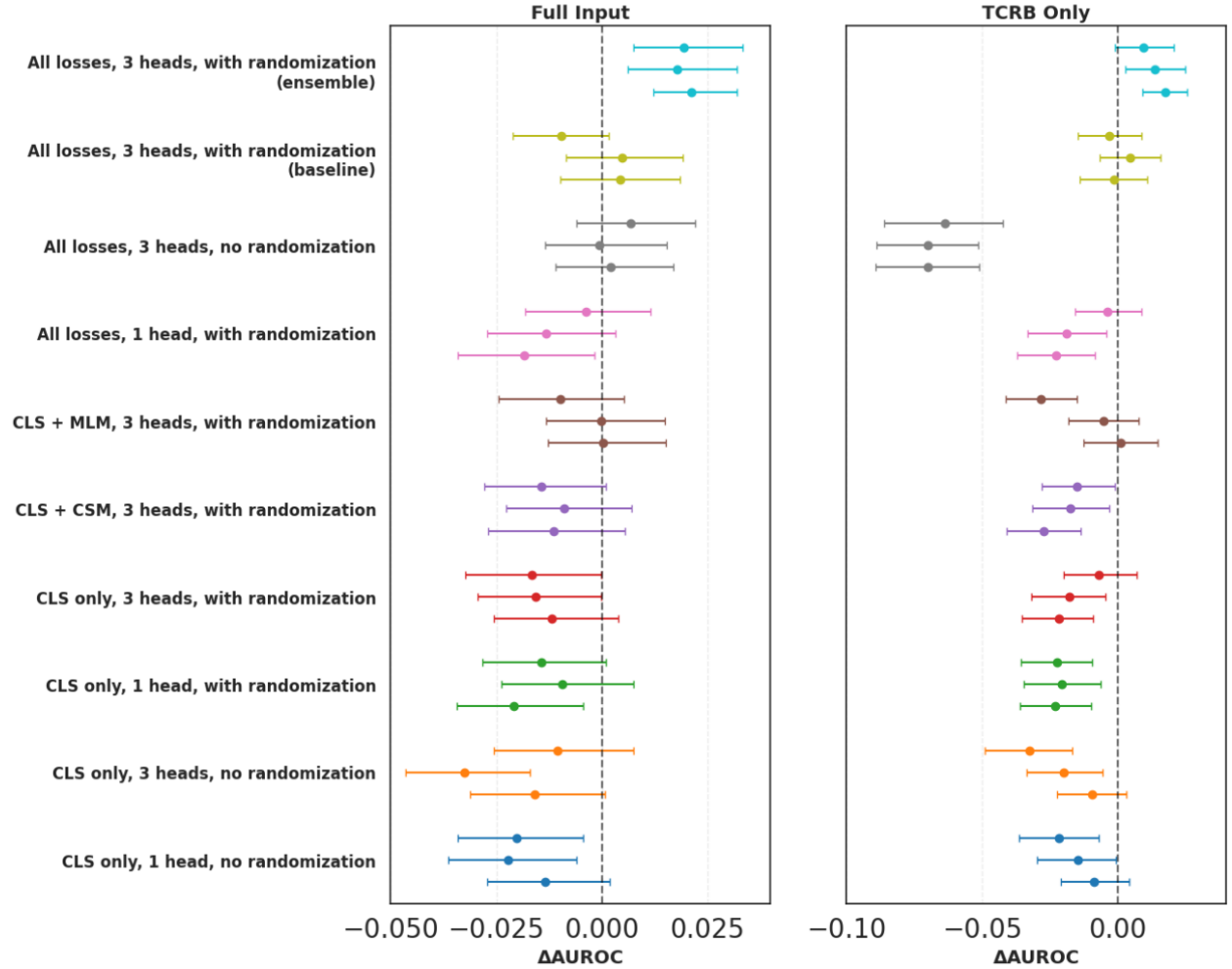| Metric	Model	α (P^α)	95% CI	β (T^β)	95% CI
AUROC	ImmSET	0.078	(0.075, 0.082)	0.036	(0.031, 0.040)
AUROC	ESM2	0.072	(0.064, 0.079)	0.032	(0.021, 0.044)

Figure 12: Ablation study. All models use the special ImmSET tokenization

All models are trained with identical data, training protocols, and tokenizer setup to ensure a fair comparison. Points in Figure 12 represent mean ΔAUROC, and error bars denote 95% confidence intervals estimated via peptide-level bootstrap resampling Algorithm 1. This evaluation strategy captures antigen-level variability and provides a more faithful measure of generalization performance.

We note that the distribution of per-peptide AUROC values is typically very broad. As a result, standard bootstrap procedures that resample rows across the entire dataset tend to underestimate the true variability of performance estimates. Peptide-level bootstrapping avoids this pitfall by resampling at the peptide level, thereby reflecting the dominant source of uncertainty. We therefore recommend, for future studies, reporting peptide-based bootstrap confidence intervals for AUROC (or other performance metrics, including ΔAUROC) rather than quoting a single point estimate. An alternative is to present the full distribution of per-peptide AUROC values, for example with a boxplot, which makes peptide-to-peptide heterogeneity directly visible.

In the main text we emphasize multiple independent training runs, which provide a strong assessment of robustness but come at a higher computational cost. For ablation studies, peptide-level bootstrapping offers a complementary view that highlights peptide-level variability with much lower compute requirements, making it particularly well suited for comparative analyses.

Appendix I. ImmSET ensemble

To achieve peak performance, we adopt an ensemble strategy (see Figure 13). Specifically, three ImmSET instances are initialized with different seeds and trained on the same dataset with varied sampling orders. Each run yields three parameter checkpoints, resulting in a total of nine parameter sets. At inference, predictions from all nine models are averaged to produce the final output.

The ensemble preserves the scaling behavior of individual ImmSET instances while providing a consistent performance gain (Figure 14).

Algorithm 1: Peptide-Based Bootstrapping for AUROC Comparison

---

Input: Dataset D of labeled TCR–peptide pairs, number of bootstrap iterations N, two models M_1 and M_2

Output: Mean ΔAUROC and 95% confidence interval

for $b \leftarrow 1$ **to** N **do**

    Identify unique peptides $\mathcal{P}$ in D;

    Sample $|\mathcal{P}|$ peptides with replacement from $\mathcal{P}$ to form $\mathcal{P}'$;

    Initialize bootstrap dataset $D_b \leftarrow \emptyset$;

    foreach *peptide* $p \in \mathcal{P}'$ **do**

        Let R_p be all rows in D corresponding to p;

        Sample $|R_p|$ rows with replacement from R_p to obtain R'_p;

        Add R'_p to D_b;

    end

    Compute $\text{AUROC}_1(D_b)$ using model M_1;

    Compute $\text{AUROC}_2(D_b)$ using model M_2;

    $\Delta_b \leftarrow \text{AUROC}_1(D_b) - \text{AUROC}_2(D_b)$;

    Store Δ_b in results list;

end

Report mean of $\{\Delta_b\}$ as ΔAUROC;

Report 2.5th and 97.5th percentiles of $\{\Delta_b\}$ as the 95% confidence interval;
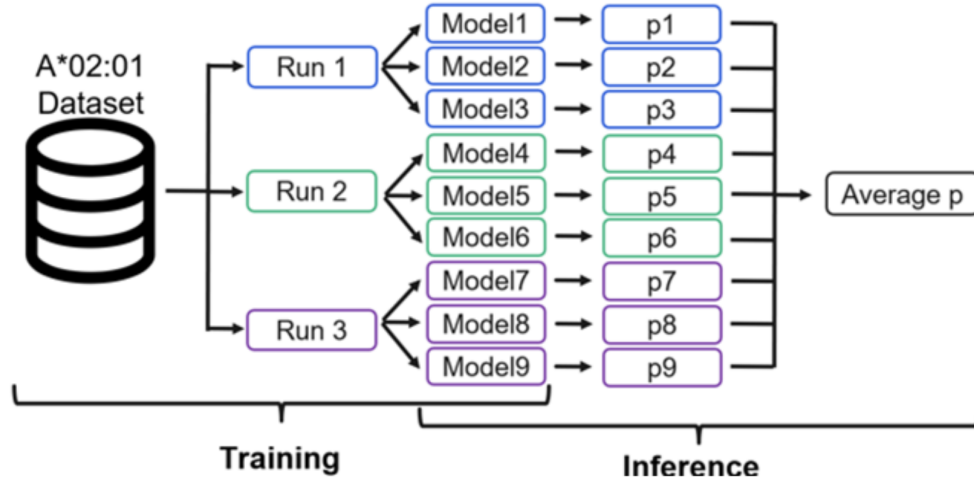

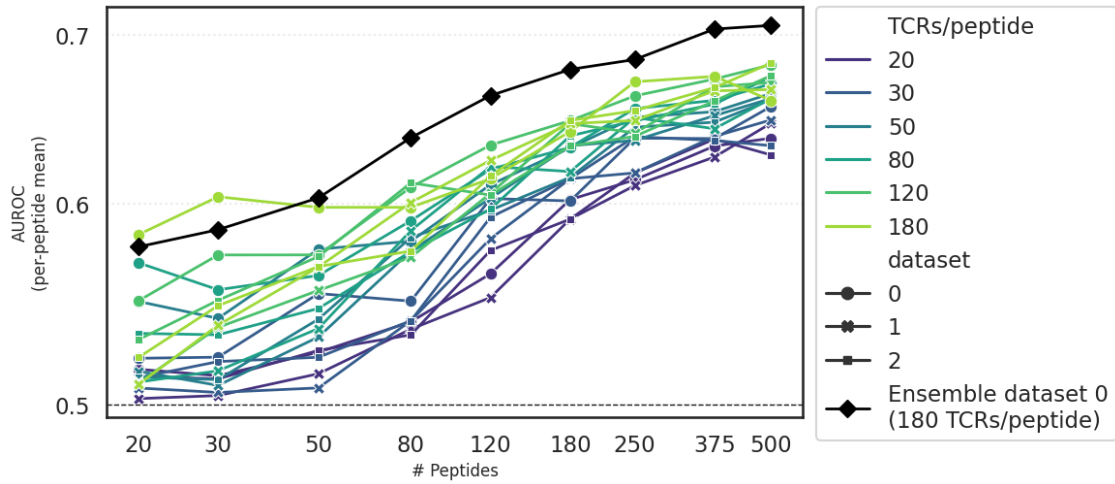Figure 13: Schematic of ImmSET ensemble

Figure 14: Ensemble scaling

Appendix J. AlphaFold-comparison test dataset composition

To compare ImmSET against AlphaFold2- and AlphaFold3-based pipelines, we used the benchmark dataset from the IMMREP2025 (Organizers, 2025) competition – a public competition comparing different models for their ability to classify TCR-pMHC interaction data. The AlphaFold3 results were provided by Phil Bradley, the winner of that competition. The ImmSET and AlphaFold2-based predictions were run in-house.

The evaluation dataset (Noakes et al., 2025) itself comprises a total of 1,000 unique TCRs, distributed equally across 20 peptide-MHC targets (50 TCRs each) from 2 HLA contexts (10 peptides each). Each TCR is specific to exactly one of the pMHC targets and experimentally confirmed to not be specific to the other pMHCs in the benchmark. In this evaluation, we use exclusively within-HLA negatives, meaning that TCRs with a ground truth specificity for an A*02:01-restricted peptide are used as negatives for other A*02:01 restricted epitopes but not for B*40:01-restricted epitopes. This is done to ensure that resulting performance is the product only of learned TCR-pMHC interaction patterns and not more general TCR-MHC preferences (Sharon et al., 2016).

The peptides in this benchmark were selected to share little similarity with any previously published T cell epitopes as found in public data repositories (Vita et al., 2025; Bagaev et al., 2022; Tickotsky et al., 2017; Burley et al., 2025; Nolan et al., 2020). We enforced the same separation from these peptides in our training and validation datasets. Each peptide is 4 or more Levenshtein distance from, and shares no substring longer than 5 amino acids with, any published peptide or any peptide in our train and validation data. This separation ensures that these peptides provide a realistic test of model generalization to new pMHC targets.

Appendix K. AlphaFold-based TCR-pMHC prediction

AlphaFold2: We compared ImmSET against two different AlphaFold2-based TCR-pMHC prediction pipelines, each with distinct adaptations of AlphaFold2 for the particular problem space of TCR-pMHC structure prediction: TCRmodel2 (Yin et al., 2023) and TCRDock (Bradley, 2023).

Predictions from the TCRdock pipeline were performed as described in its publication (Bradley, 2023). Briefly, for each TCR-pMHC target, three structures were predicted with the AlphaFold monomer "model_2_ptm" parameter set with a diverse set of hybrid structural templates sampling native TCR-pMHC docking geometries provided for each round of inference. The most confident predicted structure by predicted aligned error (PAE) was selected for each target TCR-pMHC complex. Classification scores were derived from the PAE score averaged across all TCR and pMHC residue pairs (TCR-pMHC PAE), then normalized for peptide-intrinsic and TCR-intrinsic effects on model confidence.

Predictions from the TCRmodel2 pipeline were performed using AlphaFold-Multimer v2.3 (Evans et al., 2021) and implementing the adaptions described in the TCRmodel2 publication (Yin et al., 2023). Five structures were predicted per target TCR-pMHC complex, one with each of five available AlphaFold-Multimer model weights, and the best ranked prediction by TCRmodel2's "model confidence" score was considered. Classification was performed using an ipTM score (interface predicted template modeling scores, a confidence score for the relative positions of the interacting chains) modified to focus only the interface between the TCR and the pMHC.

AlphaFold3: The AlphaFold3 predictions on the IMMREP25 benchmark were provided to us by the winner of that competition. The results compared in this study reflect AlphaFold3 run on the target TCR-pMHC complexes with two modifications – both reflective of how the model was run in winning the IMMREP25 competition. First, the model is modified to include inter-chain template features, rather than the default of only intra-chain template features. Second, the model is provided example TCR-pMHC docking examples as templates constructed by sampling from the canonical binding modes of known TCR-pMHC complexes, as done for AlphaFold2 in TCRDock (Bradley, 2023).

Classification scores are finally derived from the model's reported pLDDT values, and are normalized across all TCRs and peptides in the benchmark set so that the average prediction across all TCRs (for a given peptide) and across all peptides (for a given TCR) are fixed to 0. This normalization is designed to remove per-peptide and per-TCR biases; a similar method is described in TCRDock.

Appendix L. Generalization to other alleles

We constructed a holdout dataset for other alleles following the same procedure applied to A*02:01. When evaluating an ImmSET ensemble model trained exclusively on A*02:01 data, we observed a systematic improvement in performance in all of the alleles using each of the 3 A*02:01 training datasets. (Figure 15).

This result is noteworthy because cross-allele transfer is not expected in general, as different HLAs vary from one another in both structural similarity and presented peptide motifs.

Appendix M. Model inference time

Table 8 reports per-TCR–pMHC inference times for ImmSET, ESM2, and structure-based baselines using a single Nvidia T4 GPU. ImmSET and ESM2 instances achieve millisecond-scale latency, with ImmSET ensembles (used for comparison against structural models) remaining under 10 ms.

In contrast, TCRdock and TCRmodel2 require minutes per sample under full structural inference, which is the configuration used for evaluation.

For reference, a single inference round corresponds to approximately 95 s for TCRdock and 5 min for TCRmodel2, though these shorter runs are not used in our reported results.

No timing results are provided for AlphaFold3 as these were provided by the IMMREP25 winner on their own compute and we have access only to the overall predictions.

Appendix N. Biological Interpretability

While detailed analysis of biological interpretability would best be done using a pan-allelic version of the model trained across multiple HLAs, we have conducted the following experiment as a first look.

We studied the impact of single amino acid changes to the peptide on model score. From our holdout dataset of 134 unique peptides, we extracted the subset of peptides of length 9 (the most common length for A*02:01 epitopes, comprising 125 of our holdout peptides). For each 9-mer, we then derived all single amino acid variants of the base peptide. Using an ImmSET ensemble model trained on dataset 0 with 500 peptides and 180 TCRs/peptide as described previously, we scored each positive TCR in the holdout

(a) dataset 0
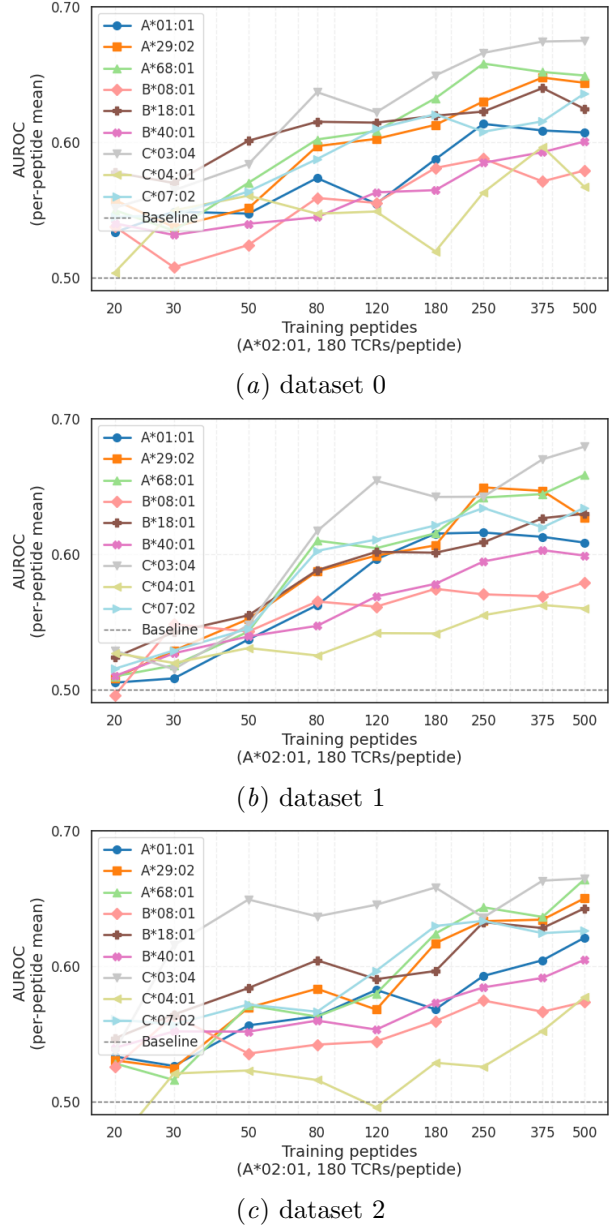

(b) dataset 1

(c) dataset 2

Figure 15: ImmSET cross-HLA generalization.

Table 8: Per TCR-pMHC inference times for ImmSET and structure-based baselines. For evaluation, we report total runtimes (full structural inference), while noting that a single round of TCRdock and TCRmodel2 corresponds to ~95 s and ~5 min, respectively.

Model	Inference time (per TCR–pMHC)
ImmSET (instance)	1.1 miliseconds
ImmSET (ensemble)	10 miliseconds
ESM2	5 miliseconds
TCRdock	4.5 minutes
TCRmodel2	25 minutes

set against all single amino acid variants of its true target.

This allowed us to asses how sensitive ImmSET is to peptide mutations as a function of amino acid properties and position. As seen in Figure 16, we observe that changes in the peptide charge, particularly at the 8th residue, led to large changes in model score.

A survey of the solved class I TCR-pMHC crystal structures suggests that the CDR1 and CDR2 of the TCRβ chain predominantly contact the peptide at position 8. This observation, coupled with the above observation that charge changes at position 8 have a large impact on ImmSET's score, led us to the hypothesis that there may be a charge-charge interaction between peptide position 8 and the TCRβ CDR1 and CDR2. Under this hypothesis, we attempted to build minimal models using only this interaction to predict TCR-pMHC specificity. Briefly, we used 5 features: the number of positive charges in TCRβ CDR1 and CDR2, the number of negative charges in TCRβ CDR1 and CDR2, and the charge of the peptide at position 8. We then fit a random forest model using these features to predict TCR-pMHC specificity using the "dataset 0" training datasets ranging from 20 to 500 peptides, each with 180 TCRs per peptide, and evaluate on the 125 holdout 9-mers and summarize results in Table 9. Since this minimal model treats only 9-mers, the training peptide counts here reflect only the 9-mer peptides in each of these datasets.

The performance of these minimal models is weaker than the overall ImmSET model but is significantly non random. This demonstrates one example in which introspection of the complicated overall model can lead to specific physiochemical hypothesis which in turn can lead to rule-based models of TCR-pMHC

interaction. These minimal models saturate performance at low numbers of training peptides, suggesting that such patterns could be learned on much smaller datasets, provided strong initial hypotheses and the data is carefully curated.

Future work can explore the question of whether ImmSET's attention maps can be converted into TCR-pMHC contact predictions, following the methods described in Rao et al. (2021) Appendix A, ideally with a pan-allelic version of the model trained across multiple HLAs where the analysis could study both TCR-peptide and TCR-MHC contacts.
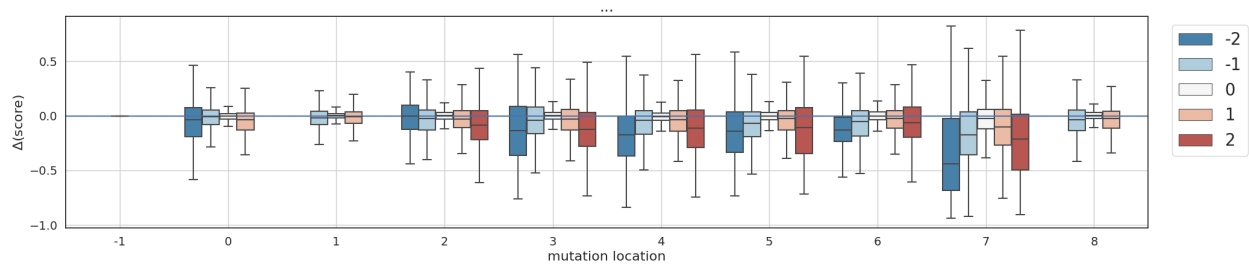
Figure 16: Change in model score as a function of the change in charge and mutation position

Table 9: Performance of minimal model.

# training peptides	Dataset	mean holdout AUROC
19	cfg0_p20_t180	0.549
28	cfg0_p30_t180	0.560
46	cfg0_p50_t180	0.560
70	cfg0_p80_t180	0.559
107	cfg0_p120_t180	0.563
158	cfg0_p180_t180	0.561
223	cfg0_p250_t180	0.564
339	cfg0_p375_t180	0.567
453	cfg0_p500_t180	0.569