TCRGenesis: Generation of SIINFEKL-specific T-cell receptor sequences using autoregressive Transformer

Yang An[*†‡] **Felix Drost**[*†‡] **Adrian Straub**[‡]

Annalisa Marsico[†] **Dirk H. Busch**[‡] **Benjamin Schubert**[†‡]

[†]Helmholtz Center Munich [‡]Technical University of Munich

Abstract

Engineered T-cell therapies are a promising new approach for treating previously uncurable diseases. These therapies involve genetically modified T cells expressing custom T cell receptors (TCRs) that recognize antigens from cancer, virus-infected, or autoimmune cells. However, the identification or generation of suitable TCRs remains an unsolved challenge. Computational methods hold the potential to accelerate the development of TCRs binding towards target antigens. While the computational investigation of the TCR-epitope landscape has been mainly focused on binding prediction, synthetic TCR design has recently emerged as the next frontier. Here, we present a proof-of-concept study on generating full TCR sequences reactive to a fixed epitope *in silico*. Towards this, we utilized a unique dataset comprising thousands of TCRs experimentally validated as reactive towards the model epitope-MHC complex SIINFEKL/H2-K^b and a naive TCR background to train our autoregressive transformer model TCRGenesis. The model generated a repertoire of realistic TCRs as validated through various biophysical and sequence properties. Further, the sequences exhibited high binding scores according to a predictor specifically developed for evaluation. The generator inherently captured the rules governing binding towards SIINFEKL as its perplexity score assigned to real, unseen TCR sequences separates well between binding and non-binding TCRs, and the generated sequences resembled binders. This work marks one of the first steps in the full-sequence design of TCRs specific to an antigen *in silico*, which we envision will accelerate the development of future immunotherapies and personalized medicine through rapid and reliable TCR synthesis.

Engineered T-cell therapies are a promising new approach for treating previously uncurable diseases. These therapies involve genetically modified T cells expressing custom T cell receptors (TCRs) that recognize antigens from cancer, virus-infected, or autoimmune cells. However, the identification or generation of suitable TCRs remains an unsolved challenge. Computational methods hold the potential to accelerate the development of TCRs binding towards target antigens. While the computational investigation of the TCR-epitope landscape has been mainly focused on binding prediction, synthetic TCR design has recently emerged as the next frontier. Here, we present a proof-of-concept study on generating full TCR sequences reactive to a fixed epitope *in silico*. Towards this, we utilized a unique dataset comprising thousands of TCRs experimentally validated as reactive towards the model epitope-MHC complex SIINFEKL/H2-K^b and a naive TCR background to train our autoregressive transformer model TCRGenesis. The model generated a repertoire of realistic TCRs as validated

[*]Equal contribution
{yang.an,felix.drost,annalisa.marsico,benjamin.schubert}@helmholtz-munich.de
{adrian.straub,dirk.busch}@tum.de

AI for New Drug Modalities at NeurIPS 2024.

through various biophysical and sequence properties. Further, the sequences exhibited high binding scores according to a predictor specifically developed for evaluation. The generator inherently captured the rules governing binding towards SIINFEKL as its perplexity score assigned to real, unseen TCR sequences separates well between binding and non-binding TCRs, and the generated sequences resembled binders. This work marks one of the first steps in the full-sequence design of TCRs specific to an antigen *in silico*, which we envision will accelerate the development of future immunotherapies and personalized medicine through rapid and reliable TCR synthesis.

1 Introduction

T cells play a crucial role in the adaptive immune system. Upon recognition of foreign linear antigen fragments, so-called epitopes, they elicit an immune reaction to clear tumorous or infected cells and pathogens from our bodies. The recognition is enabled through their T-cell receptor (TCR), a specialized cell-surface receptor binding towards epitopes presented on Major Histocompatibility Complexes (MHC). TCRs are highly specific towards their targets and enable T cells to see "inside" a cell using the MHC as a window. Therefore, T cells with engineered TCRs hold immense potential for highly versatile and reliable immunotherapies against autoimmune diseases [43], viral infections [42], and tumors [12]. However, identifying suitable TCRs within the vast and diverse repertoire represents a difficult challenge and until now requires extensive experimental repertoire screenings. Especially for cancer, this depicts a major roadblock in therapy development as the somatic mutations are often patient-specific, requiring personalized functional TCRs. As a solution, *in silico* design of epitope-specific TCRs would alleviate the cumbersome screening process, thereby reducing cost and time, crucial in treating cancer patients.

In recent years, computational TCR models aimed to identify the epitope specificity from the TCR sequence. A straightforward approach is to query annotated reference atlases of disease-specific [8] or general TCR-epitope pairs [52, 2] and match similar TCRs with unknown specificity. The similarity between TCRs is estimated through weighted edit distances [33] or overlapping K-mers [6]. Lately, deep learning models have been proposed to project TCR sequences into numeric embeddings, reflecting its specificity through unsupervised [46] or supervised [10] representation learning. Furthermore, models have been trained to directly predict specificity towards a pre-defined epitope using classical machine learning algorithms such as Gaussian Processes [24] or Random Forests [18]. These predictors and database queries treat the epitope solely as a category and only consider the TCR sequence. However, both approaches have the disadvantage that the target epitope must be fixed *a priori*, and several TCRs with validated binding must be known. Therefore, several deep-learning models were proposed that incorporate the epitope sequence. Instead of learning the features characteristic of a TCR specific to a single target, they aim to model the interactions within the TCR-epitope complex. A plethora of methods experimented with different network types such as convolutional neural networks [53], recurrent neural networks [15], and attention-based networks [35, 37, 3], alongside differing their input encodings such as learned embeddings [35, 37, 3, 15] and physio-chemical informed residue representations [53, 15]. Nevertheless, differences in performance could often be attributed to varying training data and testing criteria. Despite continued development in the last years, TCR-epitope predictors were reported to fail for out-of-distribution predictions on unobserved epitopes [19].

Another line of work aims at the direct generation of TCRs. First computational approaches include IGoR [32] and OLGA [44] which simulated the V(D)J-Recombination process [39]: one germline segment for V-, (D)-, and J-genes respectively were randomly sampled and joined (combinatorial diversity), followed by random insertions and deletions at their joints (junctional diversity). Based on these results, evolutionary selection can be incorporated into these processes by correcting the generation probabilities [45]. soNNia [22] filters the V(D)J usage according to experimentally validated target repertoires through a neural network. Subsequently, deep learning model for TCR generation followed and several models have been proposed to generate CDR3β sequences through unconditional [7] and conditional VAEs [26], autoregressive GRU [23], and reinforcement learning [4, 28]. Fast et al. [13] used an LSTM to generate CDR3α and -β given start V and J-genes and a target. However, these approaches focus on the Complementarity-determining region 3 (CDR3) of the β-chain, the most important [47] but not sufficient region for binding, instead of the full and paired receptor sequence. Therefore, they omit critical parts of the sequence required for binding and expression in cells. Additionally, generating TCR sequences for any specified epitope would be

more practical and broadly applicable, the currently available data is insufficient for generalization as evidenced by evaluating binding predictors on out-of-distribution samples [19]. Beyond difficulties in training a pan-epitope generative model, *in silico* evaluation of generated sequences using general predictors may not translate to experimentally validated results.

In this work, we present TCRGenesis, a deep learning approach for generating epitope-specific TCRs through an autoregressive transformer model. The model leverages a unique in-house dataset comprising 6,880 TCRs specific to the ovalbumin-derived model epitope SIINFEKL bound uniquely to the MHC H2-K^2 and tens of thousands of naive background TCRs with unknown specificity. The model generated realistic, full-sequence TCRs with matched α- and β-chains that resemble the biophysical and sequence properties of the experimentally validated repertoire. Since general TCR-epitope predictors proved unsuccessful for the SIINFEKL epitope in our experiments, binding specificity was validated through a predictor specifically trained to estimate binding to this target. This predictor provided us additional insights into the TCR sequence elements required to model the TCR-epitope interaction. Our synthetic TCR sequences closely resembled the real SIINFEKL-binders while differing from naive sequences. TCRGenesis inherently learned to distinguish between binders and non-binders. We envision that our model serves as a proof-of-concept study to generate the whole sequence of antigen-specific TCRs *in silico*, which ultimately can be employed for personalized immunotherapies against autoimmune diseases, infections, and cancer.

Our main contributions can be summarized as follows:

1. To the best of our knowledge, we trained the first generative model on full and paired TCR sequences binding towards a single epitope.
2. Thorough evaluations were performed to determine the model's capability to generate valid TCRs using ANARCI [11] and biophysical properties.
3. The binding capability of the generated TCR sequences was assessed using a reliable predictor, trained specifically on TCR-SIINFEKL binding only.

2 Method

Formally, the task is to generate *de novo* TCR sequences $\mathbf{x} = (x_1, ..., x_L)$, where L is the variable length of the sequences. These sequences should display high binding affinity towards SIINFEKL while preserving its overall biological functionality and expressability. Given a dataset $\mathbf{X_S}$ of known SIINFEKl-binding TCRs, a generative model $\mathbf{g}_\theta(\mathbf{x})$ parameterized by θ is trained to approximate the true distribution $p(\mathbf{x_S})$ underlying these sequences. Subsequently, the model is used to sample novel TCR sequences with similar binding properties.

TCRs are heterodimers formed by an α and a β chain represented as a sequence of amino acids. To encode them for our models, the sequences were tokenized on an amino-acid level, treating each residue as a discrete element from a vocabulary of the 20 canonical amino acids. Both chains were then concatenated with a separation token '-' between them to preserve the distinction of the two chains. Start and end tokens were added and the sequences were padded to a fixed length to enable parallelization during model training.

2.1 Predictor

For *in silico* evaluation of the generated sequences, we first trained an ensemble [25] of TCR-specificity predictors on a binary sequence classifcation task. Our predictor models consist of a pretrained ESM-2 [29] transformer encoder for feature extraction. ESM-2 was trained on 250 Mio. protein sequences and has been shown to inherently encode structural information in their embeddings. CLS token pooling was performed to extract a feature vector of fixed length. The classification head consists of blocks of linear, batch normalization [21], ReLU activation function, and dropout [49] layers. The last linear layer outputs a single logit, normalized using a sigmoid activation function to calculate a binding score. This model was fine-tuned using LoRA [20] to optimize the Binary Cross Entropy loss on our dataset. The averaged output $BS(\mathbf{x})$ of the ensemble is then used to quantify the binding capability to SIINFEKL of a TCR sequence $\mathbf{x}$.

2.2 Generator

Transformer encoders are effective for feature extraction but less so for sample generation. Therefore, we opted for the pre-trained, autoregressive ProGen2 model [38] as the generative model. We fine-tuned ProGen2 to model the distribution of SIINFEKL-binding sequences by using the next-token prediction task with cross-entropy loss. For sequence generation, we employed temperature sampling at $T = 1.0$.

3 Experiments

3.1 Data

To train our discriminative and generative models, we used an in-house dataset consisting of 6,880 unique mouse TCR sequences determined experimentally as binding towards SIINFEKL. A negative background set of 35,307 TCRs was derived from naive repertoires of mice that had not been exposed to ovalbumin. Despite the possibility, that SIINFEKL-binding TCRs may be contained in the naive repertoire, the probability of this occurring is marginally small. Consequently, the naive TCRs were treated as non-reactive with respect to SIINFEKL.

To prevent data leakage, we used scirpy [50] to group TCRs into clonotype clusters. Two sequences were grouped if they had identical sequences in either the α- or β-chain. Since this partitions the TCRs into big conglomerates, the groups were then further divided using Leiden clustering [51] forming clonotypes. The dataset was split into six subsets, stratified by specificity and grouped by the aforementioned clonotype definition. Due to grouping, the subsets have nearly but not exactly equal size. Five of the subsets were used for cross-validation training of the predictor models, while one was held out as a test set. The generative model was trained on four subsets, while one served as validation and one as test set.

3.2 Training

Optuna [1] was used for hyperparameters optimization. All hyperparameter optimization runs had a budget of 48 hours on a single Nvidia A100 for both the discriminative and generative models. The final hyperparameters can be found in Table 3 and 4. The predictors were optimized for their performance to discriminate between binding and non-binding TCRs, measured on their respective validation set using Area under ROC curve (AUROC).

While binding prediction has a defined aim, the generation of TCRs has multiple partially conflicting objectives, with redundant solutions including mode collapse or memorization of training sequences. To balance between different aspects of the generative model, we ran the hyperparameter optimization to maximize a heuristic score $S_{HPO}(\mathbf{X_{gen}}$ on a set of 128 generated samples $\mathbf{X}_{gen}$ at the end of each training epoch:

$$S_{HPO}(\mathbf{X}_{gen}) = N(\mathbf{X}_{gen}) * D(\mathbf{X}_{gen}) * 0.5 * (1/PP(\mathbf{X}_{gen}) + BS(\mathbf{X}_{gen})) \tag{1}$$

where the novelty $N(\mathbf{X}_{gen})$ indicated the ratio of generated samples having at least one residue-level difference to the training dataset to prevent memorization and overfitting. For prevention of mode collapse, diversity $D(\mathbf{X}_{gen})$ was defined as the fraction of samples after deduplication to the total number of generated samples. The perplexity of a protein sequence inferred from generative models has been shown to correlate with expressibility and fitness [41, 16]. Hence, we used our generative model at the current training state to predict a perplexity score $PP(\mathbf{X}_{gen})$ of generated samples as an indication of mutation effects. As all other metrics range between zero and one with higher values indicating higher preference, we took the inverse of the perplexity. The inverse perplexity further served as an out-of-distribution score and therefore a proxy of trustworthiness for our binding predictors. Finally, the mean binding score $BS(\mathbf{X}_{gen})$ derived from our trained predictor ensemble indicated the binding of the generated sequences to SIINFEKL.

3.3 Predictor Results

As a first step, we required a predictor to classify whether a given TCR is specific to the SIINFEKL epitope to perform an efficient computational evaluation of our generated sequences. In recent

Method	Network	AUROC	APS	F1-Score	Accuracy
iTCep [53]	CNN	52.1	14.4	22.7	53.8
TULIP-TCR [35]	Attention	52.4	14.0	23.7	47.7
BERTrand [37]	Attention	52.9	14.9	22.6	66.0
TCellMatch [15]	RNN	53.2	14.6	24.0	54.9
ATM-TCR [3]	Attention	57.0	16.8	24.5	70.9
Ours - CDR3α	Attention	70.1±1.4	31.4±1.5	25.1±7.6	85.8±1.0
Ours - CDR3β	Attention	81.6±0.2	52.1±2.2	46.6±5.3	87.6±0.7
Ours - CDR3$\alpha\beta$	Attention	85.3±0.6	62.6±1.4	55.3±2.1	90.1±0.5
Ours - Full	Attention	89.6±1.1	73.9±0.6	66.6±0.7	92.0±0.4
Ours - Ensemble	Attention	**92.0**	**79.8**	**72.2**	**93.7**

Table 1: Comparison in discriminative performance between five pan-epitope predictors and our SIINFEKL-specific predictors trained on different input regions. All values are given in %. The standard deviation of our models is obtained on the same test set for K=5 training splits.

years, several methods have been proposed to estimate binding based on the TCR and the epitope sequence, but have been shown to generalize poorly to novel target epitopes [19]. Further, they have been predominantly trained on human TCRs using the CDR3β sequence as their sole input, as this represents the majority of the data currently present in public databases [52, 2]. We evaluated 13 of these predictors on the test set using ePytope-TCR [9] and here report the five predictors with the highest obtained AUC scores (Table 1). We observed a performance only slightly better than random predictions at an AUC between 0.5 and 0.6 and an average precision score lower than 0.2.

We found the overall performance of these models was not sufficient to evaluate a generative model. Therefore, we implemented an ESM-2-based predictor for which we report the performance on a test set of five models with varying training splits. The unique character of our dataset, containing full and paired sequence information, allowed us to quantify the influence of the different TCR regions on the binding prediction (Table 1). As expected, the CDR3β sequence was more informative for binding prediction than the CDR3α sequence leading to an increase of 0.115 in Area Under the Receiver Operating Characteristic Curve (AUROC) and 0.207 in Average Precision Score (APS) as the β-chain was observed to be in closer contact to the epitope [48]. Combining CDR3-α and -β and utilizing full sequence information further improved the performance to an AUC of almost 0.9 at an APS greater than 0.7. Ultimately, the best scores were reached across all metrics, when combining the five models trained on the full sequence information into an ensemble model.

This improvement of our model was not surprising as it was developed and trained on the binding of TCRs towards SIINFEKL, which is only contained to a limited amount in public databases used to train general TCR-epitope predictors. However, a reliable predictor was necessary to allow for a cost- and time-efficient evaluation in the development process of our generative models without exhaustive wet-lab experiments.

3.4 Generator Results

After hyperparameter optimization, $2^{14} = 16,384$ TCR sequences were generated to evaluate the generative capabilities of our model. First, we removed duplicate sequences, which reduced the number of unique sequences to 12,895. Since our focus is on *de novo* generation, we also excluded sequences matching SIINFEKL-binders from our dataset, resulting in 12,538 novel and unique sequences. We further discarded synthetic sequences with more or less than two chains (N=12,521). Generating repetitions, hence too many chains, is a common failure mode observed in large language model generation [17].

Next, we used ANARCI [11] as an external tool to validate generated TCRs. ANARCI aligns input sequences to a Hidden Markov Models (HMM) representing reference TCRs from IMGT [31]. Sequences with hmmer [14] bit-score below the default threshold of 80 were discarded. Eleven sequences could not be aligned with any sequence from IMGT, leaving 12,510 sequences. Further, ANARCI numbers the recognized and aligned sequences with the IMGT scheme [27]. This enables the identification of regions within the sequence, including the CDR3. Generated sequences with

lengths in α- and β-chains or identified regions (CDR1-3, FR1-4) outside the length interval of real binders were also filtered out (N=12,464) leaving 76,07% of the generated TCRs after filtering.

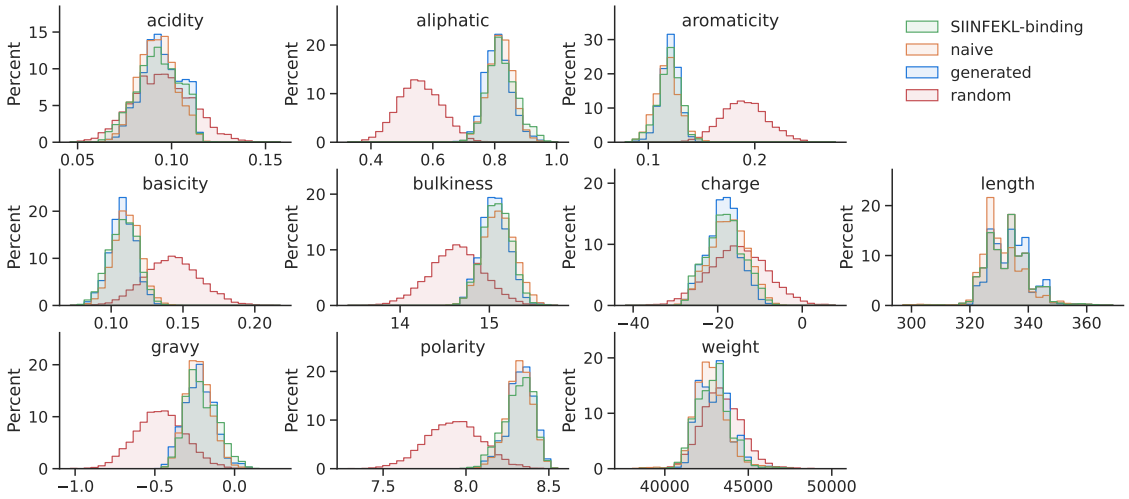
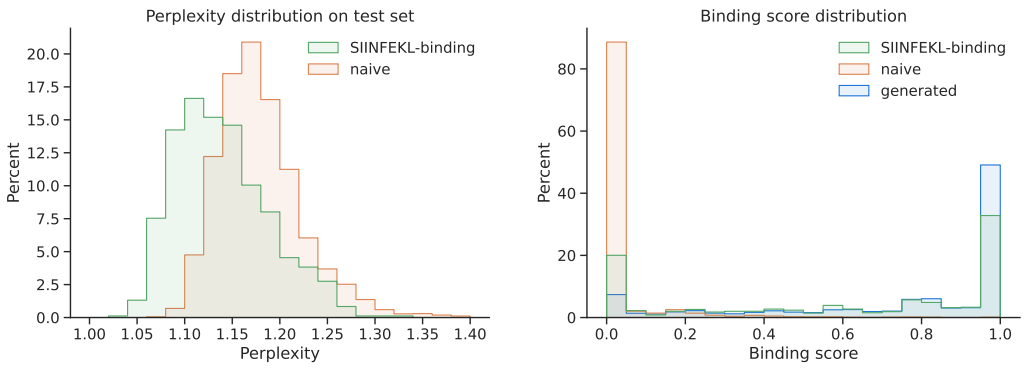


Figure 1: Comparison of nine biophysical and sequence length distributions between SIINFEKL-binding, naive, generated, and random sequences.

Following these manual and external filter steps, we confirmed that the generative model captures the distribution of real TCRs. To this end and in line with [7], we calculated the length distribution and nine biophysical properties of each TCR sequence. As a baseline, we generated random amino acid sequences of the same length distribution. For all properties, our generated distributions overlap well with real SIINFEKL-binding and naive TCRs, while being distinct from the random sequences (Figure 1). To quantify the differences in distributions, we calculated the Wasserstein distance between our generated and random sequences, as well as our generated and the SIINFEKL binder. The ratio between both were taken to create values in the same range. Higher values indicate a higher distribution overlap between generated and binder compared to generated and random sequences: acidity: 2.4, aliphatic: 20.1, aromaticity: 49.7, basicity: 31.0, bulkiness: 23.5, charge: 7.1, gravy: 15.9, polarity: 29.2, weight: 4.3.

While ANARCI and the biophysical properties confirmed that our model successfully generated realistic TCRs, these metrics are not capable of estimating binding towards SIINFEKL. Consequently, we analyzed if the generative model captured patterns characteristic of SIINFEKL-binding TCRs. Ideally, low perplexity to unseen, in-distribution sequences should be assigned by our model. On the test set of SIINFEKL-binding TCRs, the perplexity reached a value of 1.19, which is near the theoretical lower bound of 1.0. This suggests that the model effectively learned the distribution of these binding sequences. Further analysis showed that when comparing the inferred perplexity of unseen binder and naive TCR sequences, a separation between both sets can be observed (Figure 2a). By treating the perplexity as an indication for negative binding affinity, the generative model achieved an AUROC of 73.4 and APS of 48.5 despite not explicitly being trained for binding prediction. These scores outperform the five benchmarked predictors and are between our CDR3α and CDR3β models.

As a more reliable metric, we used our predictor ensemble to estimate if the generated sequences could bind to SIINFEKL. The synthetic sequences were predicted to have an average binding score of 0.75 ± 0.33, higher than binder sequences from the test set with 0.59 ± 0.39 and significantly above the scores of the naive sequences 0.03 ± 0.11. The Wasserstein distance measuring the overlap of prediction scores between the generated and real test binders was 0.15, demonstrating a strong similarity, while the distance between the generated and naive sequences was much larger at 0.72, reflecting the expected divergence towards non-binding TCRs (Figure 2b).



(a) (b)

Figure 2: (a) Using the per-sequence perplexity values predicted by our generative model, we were able to distinguish between real TCRs binding to SIINFEKL and naive TCRs from the test set. (b) The generated TCR sequences had a high overlap in predicted binding score with real binders from the test set.

To further demonstrate that our generated sequences are not only similar to any TCRs but also likely to bind, we employed the edit distance as a straightforward and human-interpretable metric for distance calculation. Edit distance measures the steps required to transform one sequence to another, while each of the operations - substitution, deletion, and insertion of an amino acid - adds one unit of distance. After calculating the edit distances of each sequence between the three sets - real SIINFEKL-binder, naive, and generated sequences, we took from each query sequence the value with the minimum edit distance (minED) from the target set. As minED is likely to decrease with higher numbers of samples in the target set, we randomly sampled 5,000 sequences from each set.

The edit distance between the generated and binding sequences was notably smaller (full sequence: 10.1$\pm$7.2, CDR3: 4.6$\pm$2.2) compared to the edit distance between generated and non-binders (full sequence: 17.8$\pm$6.4, CDR3: 6.3$\pm$1.4). Additionally, the edit distance distribution between generated and naive TCRs closely mirrored the distance distribution between SIINFEKL-binder vs. naive TCRs (full: 19.1$\pm$8.8 CDR3: 6.3$\pm$1.3, Figure 3a,3b). Combined with the binding score results from our predictor ensemble, this indicates that our model learned the pattern of SIINFEKL-binding TCRs beyond redundant single-point mutations. Although a greater distance of our generated sequence to known binding CDR3 sequences moderately correlate in negative direction (Spearman correlation = -0.477, p-value < 0.001) with the predicted binding score, TCRGenesis was still able to generate a high proportion of predicted binders even at minED within CDR3 up to 12 (Figure 4).

Another way to visualize the relationship of the generated TCRs to the landscape of real TCRs is through Uniform Manifold Approximation and Projection (UMAP) for dimensionality reduction. All pairwise edit distances of the previously subsampled sets (5000 samples from each of the three groups) were used as input to UMAP (Figure 3c, parameter: $n\_neighbors = 1000$, $min\_dist = 5.0$, $spread = 5.0$). Overall, the SIINFEKL-binding and generated sequences tended to form small, overlapping clusters distributed widely across the latent space, while the naive TCR sequences were more dispersed throughout the whole space. This pattern reflects the nature and relationship between TCRs and epitopes: While similar TCRs share similar binding profiles and each TCR is highly specific to its cognate epitopes, a single epitope can be recognized by many TCRs, often with distinct motifs.

To conclude the success rate of generating unique, novel sequences with high fitness and strong binding potential to SIINFEKL, we applied filters based on perplexity thresholds and binding score thresholds. For perplexity as a proxy for fitness, we chose the 90th percentile on SIINFEKL-binders from the test set to determine in-distribution membership. Despite this strict cutoff, only a single sequence was removed. We believe, that our previous filters were rigorous enough to filter out sequences unlikely to be expressed in nature. As the final step, we chose the binding score value, where our predictor performs with 90% precision on the test set. 3,162 TCRs were removed as

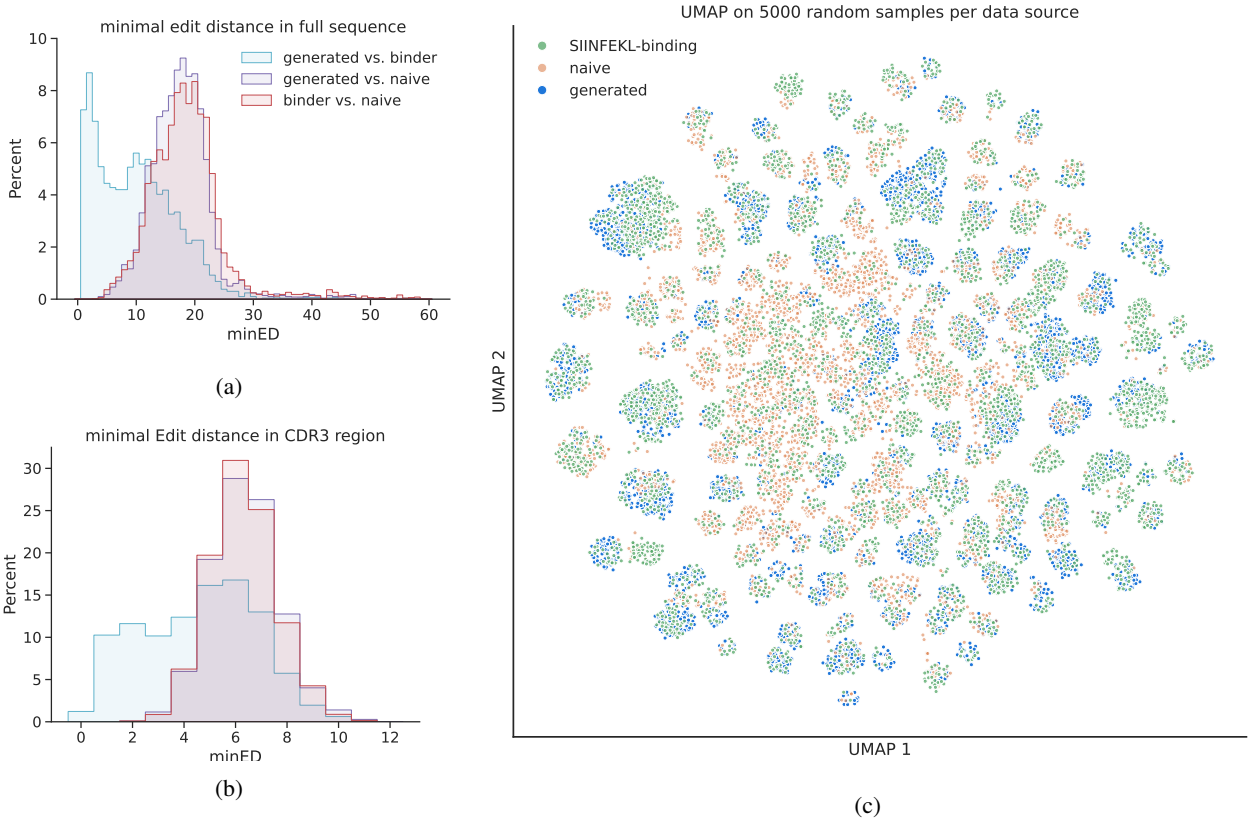


Figure 3: (a, b) Edit distance of each sequence from the query set (5000 random samples) to the closest sequence in the target set (5000 random samples) in (a) paired full sequences (b) paired CDR3. (c) UMAP [34] using 5000 random samples from each data source.

potential non-binders, leaving 9,301 high-confidence samples. In conclusion, from the 16,384 generated sequences, 57.1% were deemed successful, SIINFEKL-specific TCRS.

Finally, we compared our model with the baseline methods soNNia[22] and TCRpeg [23]. soNNia is a parametric statistical model of the V(D)J recombination process. After the generation, samples were filtered based on the V(D)J frequencies in our dataset. TCRpeg follows a similar paradigm to our method and uses an autoregressive GRU [5] model with word2vec [36] embeddings as input, originally trained on CDR3β only. For a fair comparison, we adapted it to use the identical full and paired sequence data and retrained TCRpeg with default hyperparameters on our dataset for 1,000 epochs. For both models, we generated the same amount of 16,384 TCR sequences. soNNia and TCRpeg both reached near-perfect scores of 1.0 for novelty and diversity. While 97% (16,119) of soNNia generated TCRs were valid TCRs, defined as having exactly two chains, were recognized by ANARCI, and had lengths within the range of real TCRs, TCRpeg generated sequences were more frequently deemed invalid, leaving 13,265 valid sequences. The binding scores of valid soNNia-generated sequences were only slightly higher than naive TCRs ($BS(\mathbf{X}) = 0.06 \pm 0.16$ leaving 439 samples potentially binding to SIINFEKL. For TCRpeg, the binding scores were significantly higher with 0.52 ± 0.39, and 6,355 passed the binding score threshold. This concludes that soNNia had a high rate of generating sequences resembling TCRs, however, matching V(D)J preferences alone is not sufficient for designing SIINFEKL-binding TCRs. The original TCRpeg model was used by their author on CDR3β sequences which are typically between 6 and 23 long [30]. Given the concatenated, full α-and β-chains with mean length 331 ± 7.8, only 81% of generated sequences were recognized as TCRs. These results (Table 2) show the need for transformer models, which perform better in modeling long-range dependencies. Further, more careful hyperparameter optimization could improve the generation quality of TCRpeg. Overall, our model was able to generate a higher

ratio of sequences (57.1%) that were evaluated as valid TCRs and have high confidence of binding towards SIINFEKL, compared to TCRpeg (39.0%) and soNNia (2.7%).

Name	Network	Diversity ↑	Novelty ↑	valid TCR ↑	Binding score ↑	minED ↓
SoNNia [22]	Recombination	**1.0**	**1.0**	0.97	0.06 ± 0.16	35.1
TCRpeg [23]	GRU	1.0	1.0	0.81	0.52 ± 0.39	17.7
Ours	Transformer	0.79	0.97	**1.0**	$\mathbf{0.75 \pm 0.33}$	**10.1**

Table 2: Comparison of our method in generative performance with TCRpeg and soNNia. The best scores are highlighted in bold. For all methods, 16,384 full TCR sequences were generated. A TCR is defined as valid, if they have exactly two chains, are recognized by ANARCI [11], and are within the real length range.

4 Conclusion and Future Work

In this study, we investigated the *in silico* generation of TCRs specific to a single epitope through an autoregressive transformer model. For this purpose, we utilized an in-house dataset containing several thousand full-length, α-β-paired TCR sequences binding towards the model epitope/MHC-complex SIINFEKL/H2-K^b and naive background TCRs. This work represents a proof-of-concept for computational full sequence TCR synthesis and optimization through generative deep learning.

To evaluate our generative model, we required a reliable estimate of the binding capabilities of synthetic TCR sequences. However, we did not observe sufficient performance on the SIINFEKL estimate when testing general pre-trained TCR-epitope prediction approaches. As previously reported [19], these predictors often fail to generalize to unseen targets not contained in public TCR-epitope databases. Hence, we developed a predictor based on the ESM-2 [29] transformer encoder to classify whether a TCR binds towards SIINFEKL which outperformed pan-epitope-TCR predictors by a large margin due to the training dataset. The unique characteristics of this dataset further allowed us to quantify the influence of different regions in the TCR sequence on the binding prediction. While the majority of general TCR-epitope predictors focus on the CDR3β sequence alone, we showed that paired with the α-chain and full sequence information leads to a significant increase in predictive performance and, as a result, should be taken into account when developing prediction methods.

As a generative model, we fine-tuned the autoregressive ProGen2 transformer [38] on SIINFEKL-reactive TCR sequences. The resulting synthetic TCRs had a similar profile to the experimentally identified TCRs on a wide variety of biophysical and sequence properties. The model also inherently learned the characteristics separating SIINFEKL-reactive and naive TCRs as its perplexity value on TCR sequences separated both classes at an AUROC of 73.4. The generated repertoire greatly resembled the positive TCRs indicated by a lower nearest neighbor edit distance compared to the naive repertoire. Similar to the positive test data, the SIINFEKL-binding score of our predictor is highly elevated for the generated sequences in contrast to non-binders following the binding repertoire. Overall, our generative model was able to create several thousand TCR sequences that passed several TCR sanity checks and achieved good scores in perplexity and predictive binding scores, higher than the compared baselines soNNia [22] and TCRpeg [23].

To further improve the binding properties of the synthetic TCRs, we plan to include Direct Preference Optimization [40] to bias the generation towards specificity against SIINFEKL. Upon further improvement in computational methods and metrics, the TCRs must be ultimately validated experimentally by expressing representative candidates and measuring their specificity through low-throughput methods. This offers us the intriguing possibility to find dependencies of computational metrics such as binding score and perplexity to the TCRs' avidity, which could ultimately be used to optimize T cell activation properties *in silico*. So far, all positive TCRs in this study were reactive towards the same epitope. Having the capability to generate thousands of TCR sequences binding towards a single epitope enables researchers to improve properties beyond binding affinity, including reduction of cross-reactivity and improving immunogenicity. Nevertheless, the approach needs to be further validated on different target epitopes. Especially, the dependency between amounts of training data and performance of the predictive and generative models will be insightful for guiding experimental researchers in the number of positive TCRs required for *in silico* generation. Additionally, synergistic

effects might improve the performance of both models when training on multiple epitopes simultaneously. Ultimately, the generation toward novel targets could be obtained by conditioning the model on an epitope representation. However, it is unlikely that any approach can truly generalize due to constrictions in publicly available data as seen in general TCR-epitope predictors.

This work represents a proof-of-concept study to computationally generate full TCR sequences recognizing a given epitope. This challenge is a crucial step towards *in silico* design and optimization of TCRs. In contrast to previous studies, we focused on full sequence generation of both α- and β-chain, as only these synthetic TCRs can be directly expressed in cells. In our view, this property is crucial for future applications in clinical use. We envision that our method showcases the possibilities of deep learning-driven TCR design. Ultimately, these works will allow us to generate antigen-specific TCRs against pathogens and tumors fast and reliably for applications in immunotherapies and personalized medicine.

Acknowledgements

Y.A. and F.D. are supported by the Helmholtz Association under the joint research school Munich School for Data Science – MUDS. F.D. acknowledges financial support from the Joachim Herz Stiftung. D.H.B. was supported by the Deutsche Forschungsgemeinschaft (DFG) SFB-TRR 338/1 2021 -452881907 (project A01). A.M. acknowledges support by the BMBF Cluster4Future program CNATM.

References

[1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631, 2019.

[2] Dmitry V Bagaev, Renske MA Vroomans, Jerome Samir, Ulrik Stervbo, Cristina Rius, Garry Dolton, Alexander Greenshields-Watson, Meriem Attaf, Evgeny S Egorov, Ivan V Zvyagin, et al. Vdjdb in 2019: database extension, new analysis infrastructure and a t-cell receptor motif compendium. *Nucleic acids research*, 48(D1):D1057–D1062, 2020.

[3] Michael Cai, Seojin Bang, Pengfei Zhang, and Heewook Lee. Atm-tcr: Tcr-epitope binding affinity prediction using a multi-head self-attention model. *Frontiers in immunology*, 13:893247, 2022.

[4] Ziqi Chen, Martin Renqiang Min, Hongyu Guo, Chao Cheng, Trevor Clancy, and Xia Ning. T-cell receptor optimization with reinforcement learning and mutation polices for precision immunotherapy. In *International Conference on Research in Computational Molecular Biology*, pages 174–191. Springer, 2023.

[5] Kyunghyun Cho, Bart van Merriënboer, Çağlar Gulçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, 2014.

[6] William D Chronister, Austin Crinklaw, Swapnil Mahajan, Randi Vita, Zeynep Koşaloğlu-Yalçın, Zhen Yan, Jason A Greenbaum, Leon E Jessen, Morten Nielsen, Scott Christley, et al. Tcrmatch: predicting t-cell receptor specificity based on sequence similarity to previously characterized receptors. *Frontiers in immunology*, 12:640725, 2021.

[7] Kristian Davidsen, Branden J Olson, William S DeWitt III, Jean Feng, Elias Harkins, Philip Bradley, and Frederick A Matsen IV. Deep generative models for t cell receptor protein sequences. *Elife*, 8:e46935, 2019.

[8] Jennifer N Dines, Thomas J Manley, Emily Svejnoha, Heidi M Simmons, Ruth Taniguchi, Mark Klinger, Lance Baldo, and Harlan Robins. The immunerace study: a prospective multicohort study of immune response action to covid-19 events with the immunecode™ open access database. *medRxiv*, pages 2020–08, 2020.

[9] Felix Drost, Anna Chernysheva, Mahmoud Albahah, Katharina Kocher, Kilian Schober, and Benjamin Schubert. Benchmarking of t-cell receptor-epitope predictors with epytope-tcr. *bioRxiv*, pages 2024–11, 2024.

[10] Felix Drost, Lennard Schiefelbein, and Benjamin Schubert. metcrs-learning a metric for t-cell receptors. *bioRxiv*, pages 2022–10, 2022.

[11] James Dunbar and Charlotte M Deane. Anarci: antigen receptor numbering and receptor classification. *Bioinformatics*, 32(2):298–300, 2016.

[12] Elvira D'Ippolito, Karolin I Wagner, and Dirk H Busch. Needle in a haystack: the naïve repertoire as a source of t cell receptors for adoptive therapy with engineered t cells. *International Journal of Molecular Sciences*, 21(21):8324, 2020.

[13] Ethan Fast, Manjima Dhar, and Binbin Chen. Tapir: a t-cell receptor language model for predicting rare and novel targets. *bioRxiv*, 2023.

[14] Robert D Finn, Jody Clements, and Sean R Eddy. Hmmer web server: interactive sequence similarity searching. *Nucleic acids research*, 39(suppl_2):W29–W37, 2011.

[15] David S Fischer, Yihan Wu, Benjamin Schubert, and Fabian J Theis. Predicting antigen specificity of single t cells based on tcr cdr 3 regions. *Molecular systems biology*, 16(8):e9416, 2020.

[16] Jonathan Frazer, Pascal Notin, Mafalda Dias, Aidan Gomez, Joseph K Min, Kelly Brock, Yarin Gal, and Debora S Marks. Disease variant prediction with deep generative models of evolutionary data. *Nature*, 599(7883):91–95, 2021.

[17] Zihao Fu, Wai Lam, Anthony Man-Cho So, and Bei Shi. A theoretical analysis of the repetition problem in text generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12848–12856, 2021.

[18] Sofie Gielis, Pieter Moris, Nicolas De Neuter, Wout Bittremieux, Benson Ogunjimi, Kris Laukens, and Pieter Meysman. Tcrex: a webtool for the prediction of t-cell receptor sequence epitope specificity. *BioRxiv*, 373472, 2018.

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

[20] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[21] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arxiv 2015. *arXiv preprint arXiv:1502.03167*, 2015.

[22] Giulio Isacchini, Aleksandra M Walczak, Thierry Mora, and Armita Nourmohammad. Deep generative selection models of t and b cell receptor repertoires with sonnia. *Proceedings of the National Academy of Sciences*, 118(14):e2023141118, 2021.

[23] Yuepeng Jiang and Shuai Cheng Li. Deep autoregressive generative models capture the intrinsics embedded in t-cell receptor repertoires. *Briefings in Bioinformatics*, 24(2):bbad038, 2023.

[24] Emmi Jokinen, Jani Huuhtanen, Satu Mustjoki, Markus Heinonen, and Harri Lähdesmäki. Predicting recognition between t cell receptors and epitopes with tcrgp. *PLoS computational biology*, 17(3):e1008814, 2021.

[25] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

[26] Allen Y Leary, Darius Scott, Namita T Gupta, Janelle C Waite, Dimitris Skokos, Gurinder S Atwal, and Peter G Hawkins. Designing meaningful continuous representations of t cell receptor sequences with deep generative models. *Nature Communications*, 15(1):4271, 2024.

[27] Marie-Paule Lefranc, Christelle Pommié, Manuel Ruiz, Véronique Giudicelli, Elodie Foulquier, Lisa Truong, Valérie Thouvenin-Contet, and Gérard Lefranc. Imgt unique numbering for immunoglobulin and t cell receptor variable domains and ig superfamily v-like domains. *Developmental & Comparative Immunology*, 27(1):55–77, 2003.

[28] Yicheng Lin, Dandan Zhang, and Yun Liu. Tcr-gpt: Integrating autoregressive model and reinforcement learning for t-cell receptor repertoires generation. *arXiv preprint arXiv:2408.01156*, 2024.

[29] 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.

[30] Long Ma, Liwen Yang, Bin Shi, Xiaoyan He, Aihua Peng, Yuehong Li, Teng Zhang, Suhong Sun, Rui Ma, and Xinsheng Yao. Analyzing the cdr3 repertoire with respect to tcr—beta chain vdj and vj rearrangements in peripheral t cells using hts. *Scientific Reports*, 6(1):29544, 2016.

[31] Taciana Manso, Géraldine Folch, Véronique Giudicelli, Joumana Jabado-Michaloud, Anjana Kushwaha, Viviane Nguefack Ngoune, Maria Georga, Ariadni Papadaki, Chahrazed Debbagh, Perrine Pegorier, et al. Imgt® databases, related tools and web resources through three main axes of research and development. *Nucleic acids research*, 50(D1):D1262–D1272, 2022.

[32] Quentin Marcou, Thierry Mora, and Aleksandra M Walczak. High-throughput immune repertoire analysis with igor. *Nature communications*, 9(1):561, 2018.

[33] Koshlan Mayer-Blackwell, Stefan Schattgen, Liel Cohen-Lavi, Jeremy C Crawford, Aisha Souquette, Jessica A Gaevert, Tomer Hertz, Paul G Thomas, Philip Bradley, and Andrew Fiore-Gartland. Tcr meta-clonotypes for biomarker discovery with tcrdist3 enabled identification of public, hla-restricted clusters of sars-cov-2 tcrs. *Elife*, 10:e68605, 2021.

[34] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[35] 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.

[36] Tomas Mikolov. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[37] Alexander Myronov, Giovanni Mazzocco, Paulina Król, and Dariusz Plewczynski. Bertrand—peptide: Tcr binding prediction using bidirectional encoder representations from transformers augmented with random tcr pairing. *Bioinformatics*, 39(8):btad468, 2023.

[38] Erik Nijkamp, Jeffrey A Ruffolo, Eli N Weinstein, Nikhil Naik, and Ali Madani. Progen2: exploring the boundaries of protein language models. *Cell systems*, 14(11):968–978, 2023.

[39] Mikhail V Pogorelyy, Anastasia A Minervina, Dmitriy M Chudakov, Ilgar Z Mamedov, Yuri B Lebedev, Thierry Mora, and Aleksandra M Walczak. Method for identification of condition-associated public antigen receptor sequences. *Elife*, 7:e33050, 2018.

[40] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

[41] Adam J Riesselman, John B Ingraham, and Debora S Marks. Deep generative models of genetic variation capture the effects of mutations. *Nature methods*, 15(10):816–822, 2018.

[42] Cliona M Rooney, CYC Ng, S Loftin, CA Smith, Congfen Li, Robert A Krance, Malcolm K Brenner, and HE Heslop. Use of gene-modified virus-specific t lymphocytes to control epstein-barr-virus-related lymphoproliferation. *The Lancet*, 345(8941):9–13, 1995.

[43] Isabelle Serr, Felix Drost, Benjamin Schubert, and Carolin Daniel. Antigen-specific treg therapy in type 1 diabetes–challenges and opportunities. *Frontiers in immunology*, 12:712870, 2021.

[44] Zachary Sethna, Yuval Elhanati, Curtis G Callan Jr, Aleksandra M Walczak, and Thierry Mora. Olga: fast computation of generation probabilities of b-and t-cell receptor amino acid sequences and motifs. *Bioinformatics*, 35(17):2974–2981, 2019.

[45] Zachary Sethna, Giulio Isacchini, Thomas Dupic, Thierry Mora, Aleksandra M Walczak, and Yuval Elhanati. Population variability in the generation and selection of t-cell repertoires. *PLOS Computational Biology*, 16(12):e1008394, 2020.

[46] John-William Sidhom, H Benjamin Larman, Drew M Pardoll, and Alexander S Baras. Deeptcr is a deep learning framework for revealing sequence concepts within t-cell repertoires. *Nature communications*, 12(1):1605, 2021.

[47] Nishant K Singh, Timothy P Riley, Sarah Catherine B Baker, Tyler Borrman, Zhiping Weng, and Brian M Baker. Emerging concepts in tcr specificity: rationalizing and (maybe) predicting outcomes. *The Journal of Immunology*, 199(7):2203–2213, 2017.

[48] Ido Springer, Hanan Besser, Nili Tickotsky-Moskovitz, Shirit Dvorkin, and Yoram Louzoun. Prediction of specific tcr-peptide binding from large dictionaries of tcr-peptide pairs. *Frontiers in immunology*, 11:1803, 2020.

[49] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[50] Gregor Sturm, Tamas Szabo, Georgios Fotakis, Marlene Haider, Dietmar Rieder, Zlatko Trajanoski, and Francesca Finotello. Scirpy: a scanpy extension for analyzing single-cell t-cell receptor-sequencing data. *Bioinformatics*, 36(18):4817–4818, 2020.

[51] Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):1–12, 2019.

[52] Randi Vita, Swapnil Mahajan, James A Overton, Sandeep Kumar Dhanda, Sheridan Martini, Jason R Cantrell, Daniel K Wheeler, Alessandro Sette, and Bjoern Peters. The immune epitope database (iedb): 2018 update. *Nucleic acids research*, 47(D1):D339–D343, 2019.

[53] Yu Zhang, Xingxing Jian, Linfeng Xu, Jingjing Zhao, Manman Lu, Yong Lin, and Lu Xie. itcep: a deep learning framework for identification of t cell epitopes by harnessing fusion features. *Frontiers in Genetics*, 14:1141535, 2023.

A Supplemental Material

Pretrained	cls-head activation	cls-head batch norm	cls-head dropout	cls-head hidden neurons	cls-head hidden layers
esm2_t12_35M_UR50D	ReLU	True	0.15	32	1

LoRA alpha	LoRA r	LoRA dropout	batch size	learning rate	early stopping	unfreeze epoch
128	8	0.15	32	0.0001	100	25

Table 3: Hyperparameters of our predictive models

Pretrained	batch size	learning rate	early stopping
progen2-small	64	0.000562	20

Table 4: Hyperparameters of our generative model

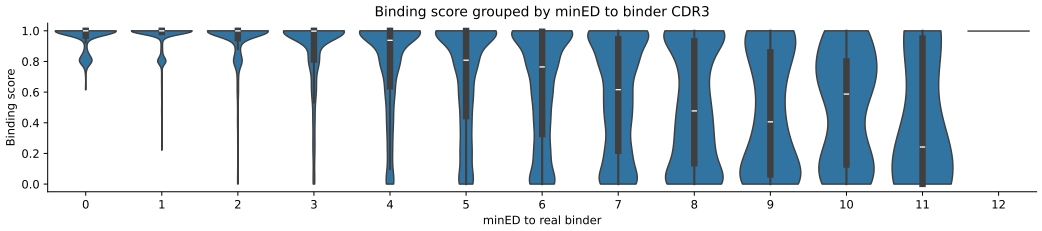


Figure 4: Binding score grouped by minED to SIINFEKL-binding CDR3 sequences. While minED and binding score correlate negatively, TCRGenesis is still able to generate binding TCRs with high minED to known binder CDR3.