Subject Section

TITAN: T Cell Receptor Specificity Prediction with Bimodal Attention Networks

Anna Weber¹,²,*, Jannis Born¹,² and María Rodríguez Martínez¹.

¹IBM Research, Zurich, Switzerland and
²ETH Zurich, Switzerland.

*To whom correspondence should be addressed.

Abstract

Motivation: The activity of the adaptive immune system is governed by T-cells and their specific T-cell receptors (TCR), which selectively recognize foreign antigens. Recent advances in experimental techniques have enabled sequencing of TCRs and their antigenic targets (epitopes), allowing to research the missing link between TCR sequence and epitope binding specificity. Scarcity of data and a large sequence space make this task challenging, and to date only models limited to a small set of epitopes have achieved good performance. Here, we establish a k-nearest-neighbor (K-NN) classifier as a strong baseline and then propose TITAN (Tcr epITope bimodal Attention Networks), a bimodal neural network that explicitly encodes both TCR sequences and epitopes to enable the independent study of generalization capabilities to unseen TCRs and/or epitopes.

Results: By encoding epitopes at the atomic level with SMILES sequences, we leverage transfer learning and data augmentation to enrich the input data space and boost performance. TITAN achieves high performance in the prediction of specificity of unseen TCRs (ROC-AUC 0.87 in 10-fold CV) and surpasses the results of the current state-of-the-art (ImRex) by a large margin. Notably, our Levenshtein-distance-based K-NN classifier also exhibits competitive performance on unseen TCRs. While the generalization to unseen epitopes remains challenging, we report two major breakthroughs. First, by dissecting the attention heatmaps, we demonstrate that the sparsity of available epitope data favors an implicit treatment of epitopes as classes. This may be a general problem that limits unseen epitope performance for sufficiently complex models. Second, we show that TITAN nevertheless exhibits significantly improved performance on unseen epitopes and is capable of focusing attention on chemically meaningful molecular structures.

Availability: The code as well as the dataset used in this study is publicly available at https://github.com/PaccMann/TITAN.

Contact: wbr@zurich.ibm.com

Supplementary information: Supplementary data will be available at Bioinformatics online.

1 Introduction

T-cells are an integral part of the adaptive immune system, whose survival, proliferation, activation and function are all governed by the interaction of their T-cell receptor (TCR) with immunogenic peptides (epitopes) presented on major histocompatibility complex molecules (MHC). A large repertoire of T-cell receptors with different specificity is needed to provide protection against a wide range of pathogens. This repertoire is generated using stochastic gene recombination and can theoretically produce diversities of $10^{15}$ – $10^{20}$ different receptors [Laydon et al., 2015] in an individual, each with unique binding capabilities. Due to this diversity, reliably predicting the binding specificity of a TCR from its sequence and understanding the mechanisms underlying TCR-pMHC interaction is highly challenging. At the same time, it has enormous potential to transform the field of immunology. A reliable prediction tool could unlock the wealth of information encoded in a patient's TCR repertoire, which reflects their immune history and could inform about past and current infectious diseases, vaccine effectiveness or autoimmune
reactions. Additionally, it could empower the application of therapeutic T-cells for cancer treatment, allowing the study of effectiveness and potential cross-reactivity risks in silico.

Recent advances in high-throughput sequencing techniques have led to the generation of an increasing amount of datasets linking TCR sequences to the epitopes they bind. However, the available data is still extremely sparse compared to the high dimensionality of the search space created by the TCR theoretical diversity. Moreover, current experimental settings typically link many TCRs to a single epitope, which leads to datasets that contain information about tens of thousands of TCRs, but only a few hundred different epitopes.

Nevertheless, several studies have attempted the prediction of TCR specificity from sequence using machine learning (for a review see [Mitch et al., 2019]). The most successful approaches so far are categorical epitope models, which exploit the relative abundance of TCR sequences to learn patterns of TCRs binding to the same epitope. Various machine learning concepts were applied to this task, including decision trees [Peckner et al., 2013], a range of different clustering approaches [Clavel et al., 2012], [Cid et al., 2017], [Otkin et al., 2019] and Variational Autoencoders [Sadhom et al., 2021]. Many of these can successfully predict which one of a small set of epitopes a given TCR will most likely bind to. However, these approaches are inherently incapable of predicting specificity to epitopes not contained in the training set (unseen epitopes), which fundamentally limits their applicability.

The next milestone towards this challenging goal are generic models, which explicitly encode both the TCRs and the epitopes. These have the potential to predict binding of any TCR–epitope pair, opening the door to the development of models that can generalize to both, unseen TCRs and epitopes. Current models show moderate performance on test data containing epitopes already encountered in training, but cannot extrapolate to unseen epitopes [Jurtz et al., 2018], [Springer et al., 2019], [Moris et al., 2020].

TITAN (Tcr epITope bimodal Attention Networks) exploits a bimodal neural network architecture to explicitly encode both TCR and epitope sequences. More specifically, TITAN uses convolutions to aggregate local information and fuses the modalities, using an interpretable attention mechanism from which binding probabilities are predicted.

Since interpretability is paramount in healthcare applications of machine learning, the use of context attention is central to our model, as it allows us to explain the choices of the algorithm and to analyze which amino acids or even atoms the model focuses on. These highlighted entities can be interpreted in the context of the underlying biochemical processes. We explore different encoding strategies for the epitopes such as SMILES [Weninger et al., 2019], a string-based, atom-level representation of molecules. SMILES are ubiquitously utilized in chemoinformatics for a wide range of applications, from predicting the chemical or pharmacological properties of molecules [Goh et al., 2017], [Masica et al., 2019] to generative modelling [Gómez-Bombarelli et al., 2019]. Using SMILES effectively results in an encoding formulation of the TCR–epitope binding problem as the more general compound protein interaction (CPI) task, thus enabling the usage of large databases of protein-ligand binding affinity for pretraining, e.g. BindingDB including >1M labelled samples [Gabriel et al., 2016].

2 Methods

2.1 Data

In order to assemble a larger and more diverse dataset, we combine data collected in the VDJ database [Bagaev et al., 2019] with a recently published COVID-19 specific dataset published by the ImmuneRACE project [Bagnes et al., 2020]. Since paired chain data is still rare, we restrict ourselves to TCR/chain sequences.

We use all human TCR/epitope sequences downloaded from the VDJ database (07/12/2020), i.e. 40,438 TCR sequences assigned to 191 peptides. Since this dataset is highly imbalanced, we exclude epitopes with less than 15 associated TCRs and downsampling to 400 TCRs/epitope and excluding epitopes with less than 15 associated TCRs to arrive at a dataset of 12,996 examples.

We refer to the combined dataset with samples from VDJ and the COVID-19 dataset as VDJ+C0VID-19.

Since these primary datasets contain only positive examples, we need to generate negative data. This can be achieved by shuffling the sequences, thereby associating TCRs with epitopes that they have not been shown to bind. Due to the low probability of a randomly drawn TCR binding a specific epitope, this manner of generating negative samples is established in the field [Moris et al., 2020], [Deicher et al., 2020]. It has also been shown to limit overestimation of performances in comparison to adding additional naive TCR sequences from other sources [Moses et al., 2020]. Furthermore, by shuffling the pairing of TCRs and epitopes, we can match the number of negative examples to that of positive examples for each TCR, avoiding unbalanced datasets. With this procedure, we build a training dataset of 46,290 examples, 50% of which are positive, encompassing 192 different epitopes.

To ensure a fair comparison to the state of the art model ImRex, we also downloaded the publicly available dataset that ImRex was trained on. This dataset is based on the VDI database and contains 13,404 samples for 118 different epitopes, with 50% negative samples. We use it to train all models for the final comparison (see section 2.4). To evaluate the performances, we use an independent test set generated from the McPAS database [Tickoxen et al., 2017] (downloaded on 03/11/2020). We exclude non-human TCRs and remove all samples with TCRs contained in the ImRex training data. Then we split the McPAS data into two test sets: one including all samples with peptides contained in the ImRex training set (seen epitope test set) and one with epitopes not contained in the ImRex dataset (unseen epitope test set). For both test sets, 50% negative data is generated by shuffling. The final seen epitope test set contains 9740 samples, the unseen epitope test set contains 1438 samples.

2.2 Models

2.2.1 Problem formulation

We are interested in learning a mapping $\Phi$ between the space of receptors $T$ and the space of epitopes $E$ to the the space of affinity scores $\mathcal{A}$, i.e. $\Phi : E \times T \rightarrow \mathcal{A}$. $\Phi$ is learned with a training data set $D = \{e_i, t_j, a_{ij}\}_{i,j=1}^N$.

More formally, for the training data $D = \{e_i, t_j, a_{ij}\}_{i,j=1}^N$, we choose $e_i$ and $t_j$ to be epitope and TCR sequences, respectively ($t_j$ is the CDR3 region). Let $(e_i, t_j)$ denote an unseen sample from the test dataset.
Overview of TITAN architecture. a) Our model ingests a TCR and an epitope sequence, which get encoded using BLOSUM62 for amino acid sequences or learned embeddings for SMILES. Then, 1D convolutions of varying kernel sizes are performed on both input streams before context attention layers generate attention weights for each amino acid of the TCR sequence given an epitope and vice versa. Finally, a stack of dense layers outputs the binding probability. Conceptually, this architecture is identical to the one proposed by Born et al. [2021a] (cf. Figure S3) but our visualization here is more fine-grained. b) The linchpin of the model is the bimodal context attention mechanism. It ingests the convolved TCR and epitope encodings, treating one as a reference and the binding partner as context. A series of transformations combines the modalities and yields an attention vector over the reference sequence (driven by the context) that can be overlayed with the molecule like a heatmap.

$D_{\text{Test}} = \{e_i, t_i\}_{i=1}^{N_{\text{Test}}}$. With the goal of predicting $\delta_j$ to approximate the unknown $\alpha_j$, we first retrieve the subset of training data $D_k$ containing the $k$ nearest neighbors using the distance measure

$$D(e_i, t_i, e_j, t_j) = \frac{\text{Lev}(e_i, e_j)}{|t_j|} + \frac{\text{Lev}(t_i, t_j)}{|e_i|}$$

(1)

where $|\cdot|$ denotes sequence length and $\text{Lev}(\cdot, \cdot)$ is the Levenshtein distance [Levenshtein, 1966], i.e., a string-based distance measure that measures the number of single-AA changes required to transform one sequence into the other. Then, the prediction $\delta_j$ is trivially computed by $\delta_j = \sum_{i=1}^{k} \omega_i$ with $\omega_i \in D_k$. We evaluate the model on all odd $K$ (to avoid ties), s.t. $1 \leq K \leq 25$, and choose the value for $K$ leading to the best ROC-AUC score for comparisons. Note that this model is non-parametric.

2.3 Model Architecture

Fig 1 shows an overview of the algorithmic steps of TITAN. To predict binding, we devise a bimodal architecture that separately ingests both the TCR and the peptide sequence.

The TCR sequences are encoded with the BLOSUM62 matrix [Henikoff and Henikoff, 1992], which is based on evolutionary similarity of amino acids and has been widely applied in TCR specificity prediction tools [Jurtz et al., 2018; Jokinen et al., 2019]. Either the full sequence or only the CDR3 region were used as input. Since the antigenic peptides are small molecules, we explore two options to encode them: one that uses an amino acid-wise encoding such as BLOSUM62, and one that uses an atom-level encoding with SMILES. All sequences were padded to the same length of 500 tokens. Advantageously, SMILES representations of a molecule are not unique; thus facilitating data augmentation [Bjornerud, 2017].

The remaining architecture is inspired by Macias et al. [2019] and almost identical to the compound-protein-interaction (CPI) model presented in Borth et al. [2021a]. In case of pretraining on CPI data (see below), the models are identical, otherwise the SMILES-encoding ligand input stream is replaced with an AA-encoding epitope stream. Three parallel channels with convolutions of kernel sizes 3, 5 and 11 are employed on the input sequences to combine information from local neighborhoods of varying spatial extent. A fourth channel has a residual connection without convolutions (see Fig 1a). For each of the four channels, we utilize two attention layers, where one modality is used as a context to compute attention scores over the other (see Fig 1b). This allows the model to use information from the binding partner, i.e., the context, to learn the importance of each token in the input sequence, i.e., the reference. The attention weights $\alpha_i$ are computed as:

$$\alpha_i = \frac{\exp (u_i)}{\sum_j \exp (u_j)}$$

(2)

where $u_i = \tanh(X_i W_1 + W_3(X_i W_2)) v^T$ and $X_i \in \mathbb{R}^{T \times K}$ is the context input, where $U$ and $K$ are sequence length and number of convolutional filters in the other modality, respectively. $W_1 \in \mathbb{R}^{H \times K}$, $W_3 \in \mathbb{R}^{T \times U}$ and $v \in \mathbb{R}^A$ are learnable parameters. Intuitively, both inputs are projected into a common attention space $\mathbb{R}^A$ with $A = 16$ and then summed up, which enables the layer to take the context into account for determining feature relevance. $v$ combines the information through a dot product, the output of which is fed to a softmax layer to obtain the attention weights $\alpha_i$, which are used to filter the inputs. Finally, both TCR and peptide information gets passed to two dense layers with 368 and 184 nodes, respectively, which output the binding probability.

2.4 Pretraining

By using SMILES encodings of the epitopes, predicting epitope receptor binding affinity can be seen as a CPI prediction task. We utilize BindingDB [Kishton et al., 2016], as of April 2020, to pretrain our model. To reduce the problem complexity and potential biases associated with the different experimental platforms to measure affinity, we binarize the binding data and define all entries in the database as binding, ignoring continuous affinity measurements. We generate an equal amount of negative examples by randomly assigning ligands to proteins [Borth et al., 2021a]. In order to avoid high discrepancy between sequence lengths, ligands with a length > 250 SMILES tokens and proteins larger than 1028 amino acids are discarded. This results in 325,688 ligands and 3351 proteins and a total of 471,917 pairs from which 98% (423,915) are used for training and the rest for validation. To adapt the model size to the larger available dataset, we changed the layer sizes for pretraining by Padding.
TCR sequences to 1028, setting the attention space $A = 256$, using four convolutional channels with kernel sizes 3, 7, 9 and 13 for epitopes and 3, 7, 13, 19 for TCR and using three final dense layers with 2048, 1024 and 512 nodes.

2.5 Data Splitting

We evaluate our models on a 10-fold cross-validation split. To determine the generalization capabilities of the models towards unseen TCRs and unseen epitopes separately, we use two different split methods. In the first, we ensure that each TCR is restricted to only one fold, which ensures that the validation datasets do not contain TCRs which were shown in training. The epitopes, however, are distributed randomly over the folds, so that most of the epitopes in the validation dataset were also shown during training. We refer to this split as the TCR split. Additionally, we generate a strict split, where we ensure that each TCR and each epitope is restricted to a single fold, ensuring that neither TCRs nor epitopes contained in the validation dataset were shown during training. To ensure the separation of TCRs and epitopes in their folds, we generate negative data by shuffling within each fold. A UMap [McInnes et al., 2018] visualization of all samples of the dataset is presented in Fig 2b and shows the ramifications of the splitting strategy. The feature space for the UMap dimensionality reduction was generated by embedding the amino acid sequences using a pretrained protein language model [Elnaggar et al., 2020].

2.6 Model Training

All described architectures were implemented in PyTorch 1.4 and used the pytorch package for data handling and preprocessing. The models optimized binary cross entropy loss with Adam [Kingma and Ba, 2015] ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1e-8$) and a learning rate of 0.0001. In the convolutional and dense layers we employed dropout ($p = 0.5$) and ReLU activation. All models were trained with a batch size of 512 on a cluster equipped with POWER8 processors and a single NVIDIA Tesla P100. The learning rate was tuned using the VDJ dataset and the remaining hyperparameters were chosen carefully based on previous experience.

For TCRs, we have the options to either focus solely on the hypervariable CDR3 loop, which is known to be the main peptide binding region, or to consider the full variable region of the TCRβ sequence, which includes the V, D and J segments and encompasses all three hypervariable loops CDR1, CDR2 and CDR3. Fig 3a shows that the use of the full sequence information boosts the model performance, indicating that valuable information is contained in regions outside of the CDR3 loop.

TCRMatch predicts TCR specificity using only sequence similarities to immobilized TCRs. To measure the performance of our models, we used the ROC-AUC metric. Table 1 shows the performance of different amino acid embedding methods on the TCR split. "weber" — 2022/2/8 — 2:14 — page 4 — #4
Table 2. 10-fold cross validation performance on TCR split. Mean and standard deviation of each model configuration on the VDJ dataset and the VDJ + COVID-19 dataset.

| Model                      | VDJ   | VDJ+COVID-19 |
|----------------------------|-------|--------------|
| K-NN (Baseline)            | 0.789 ± 0.01 | 0.78 ± 0.01  |
| AA CDR3                    | 0.75 ± 0.02  | 0.74 ± 0.007 |
| AA full                    | 0.76 ± 0.007 | 0.75 ± 0.007 |
| SMI CDR3                   | 0.73 ± 0.007 | 0.76 ± 0.008 |
| SMI full                   | 0.75 ± 0.006 | 0.77 ± 0.006 |
| Pretrained                 | 0.81 ± 0.01  | 0.84 ± 0.01  |
| Pretrained aug.            | 0.803 ± 0.01 | 0.861 ± 0.00 |
| Pretrained semifrozen      | 0.801 ± 0.01 | 0.844 ± 0.00 |
| Pretrained semifrozen aug. | 0.822 ± 0.01 | 0.868 ± 0.01 |

Regarding the pretraining, we tested two different settings, one where all weights could be adapted during fine-tuning, and a semifrozen setting, where only allowed weight changes in the TCR channel and the final dense layers of the epitope. This was done to prevent catastrophic interference of SMILES features due to the extremely low number of different epitopes in the fine-tuning dataset. Fig 3 shows that pretraining severely improves model performance in both settings. We further improved model performance by exploiting the non-uniqueness of SMILES strings to perform data augmentation. Augmentation is achieved by randomly generating a valid SMILES representation [Bjerrum, 2017] of the epitope on the fly at each training step using pytda [Born et al., 2019]. The best overall model performance was achieved by the semifrozen pretrained model with augmentation, with a mean ROC-AUC of 0.868 ± 0.005 and a mean balanced accuracy of 0.790 ± 0.005, clearly outperforming the K-NN baseline by a large margin.

The high performance on validation data that does not contain TCR sequences used in training shows that the model successfully generalizes to unseen TCRs. Comparing the performance across groups of TCRs with different similarities to the training data, we find that while the model performs better for TCRs that are highly similar to their closest partners in the training set, the performance on the TCRs with highest distance from the training data is still high (ROC AUC 0.844 ± 0.016, balanced accuracy 0.709 ± 0.008, see Fig S3).

3.2 Analysis of Attention Layers

We can investigate the decision processes of TITAN in the AA CDR3 setting – a setting that clearly outperforms previous work (see subsection 3.1) – over a number of exemplary CDR3 sequences.
We see that while there is some preference to focus on certain positions, the model adapts the attention to the different sequences. The mean inter-TCR variance per token is at $4.3 \cdot 10^{-5}$. Moreover, we find that the attention also adapts to the context (i.e., the epitope), for which we want to predict the interaction. The mean variance per token within the same TCR interacting with different epitopes is $1.3 \cdot 10^{-5}$. This demonstrates that the model is capable of adapting the attention layer to both the input and the context.

Fig. 4 shows the attention of the same model (AA CDR3 setting) on several exemplary epitopes. We can clearly see that the model chooses to focus heavily on the same positions on each epitope. The preferred positions are independent of the sequence of both the input epitope and the context TCR. Comparing the attention scores across epitopes, we find that both the inter-epitope and the intra-epitope variance of the attention are extremely small, at $4.9 \cdot 10^{-10}$ and $2.5 \cdot 10^{-9}$ respectively. This behavior indicates that TITAN fails to learn meaningful patterns in the epitope sequences from the limited diversity of epitopes in the dataset. We conjectured that the model finds a way to internally generate classes of epitopes represented by meaningless — but unique — vectors and predict specificity of TCRs towards these. This hypothesis is supported by the observation that compressing each epitope sequence to the chain of amino acids with attention scores $\alpha_i > 0.1$ yields only a very moderate reduction in the number of unique sequences. Crucially, these shortened epitope sequences are still unique for 185 out of 192 epitopes in the dataset (96%). By focusing on these fixed, invariant positions the model circumnavigates learning generic representations of epitopes and instead internally classifies epitopes, at the cost of losing the power to differentiate 19 epitopes.

Pretraining the models on the compound interaction task greatly increased the diversity of sequences that were seen by the epitope channel of the model. While most of the ligands in BindingDB are not peptides, and may therefore exhibit structures and chemical properties that differ strongly from the epitopes, the model may nevertheless infer general rules of chemical interaction from them. This enables the model to learn more meaningful attention weights, which may be the reason for the performance boost we observe for pretrained models. In Fig. 4, we show a visual representation of the attention scores over one exemplary epitope as a case study of TITAN in the pretrained semifrozen augmented setting. The chosen epitope CINGVCWTV was not shown during training. The attention scores therefore reflect the transferable knowledge the model has gained during pretraining. We see that the attention patterns align well with our expectations. The attention is high on many of the oxygens, as well as on the two thiol groups of the cysteines. Nevertheless, we see that while the attention layers may extract some chemically relevant structures, they fail to capture others. Specifically, large and hydrophobic amino acid sidechains tend to get low attention scores, although they may be of high importance for molecule interactions.

### 3.3 Performance on Strict Split

Fig. 3 compares the performances of the different TITAN settings on the strict epitope split, which measures the generalization capabilities to unseen epitopes. As expected, model performance drops severely across all settings. Moreover, we find that all TITAN settings perform similarly, with mean ROC-AUC scores around 0.6, while the K-NN baseline model shows a mean ROC-AUC of only $0.53 \pm 0.03$ for a choice of $k = 25$ (see Fig. 5 for comparison). Compared to the unseen TCR performance, we also see an increase in the standard deviation of the scores, with ROC-AUC scores of over 0.7 for some folds, and 0.5 for others. All mean ROC-AUC values are summarized in Table 3.

The best score overall is still achieved by the pretrained semifrozen augmented model, with a mean ROC-AUC of 0.62 ± 0.05 and a mean balanced accuracy of 0.61 ± 0.05. However, its superiority over other TITAN setting is not as large as in the TCR split scheme. We can also see that pretraining does not strongly improve model performance on unseen epitopes.

| Model | ROC-AUC |
|-------|---------|
| K-NN (Baseline) | 0.535 ± 0.03 |
| AA CDR3 | 0.60 ± 0.04 |
| AA full | 0.59 ± 0.04 |
| SMG full | 0.60 ± 0.06 |
| Pretrained | 0.584 ± 0.04 |
| Pretrained + Aug. | 0.591 ± 0.03 |
| Pretrained semifrozen | 0.577 ± 0.06 |
| Pretrained semifrozen + Aug. | 0.618 ± 0.06 |

Another surprising result is the comparably good performance of the TITAN AA CDR3 setting on unseen epitopes. The analysis of the attention
layer in Fig. 5 shows that in this setting, TITAN uses an attention mask to reduce the task to a TCR classification problem. This should prevent the model to generalize to new epitopes. A close look at Fig. 5 reveals that the best performance of the AA CDR3 model on the strict split is achieved during the first 10 to 20 epochs of training. During this time, the attention is still uniformly distributed over all input tokens, because many training epochs are required for the model to build the static attention mask described in section 3.2.

Fig. 5 also shows that while the ROC-AUC increases continuously over the training epochs in the TCR split cross-validation, it stagnates during training on the strict split. We hypothesize that the underlying factor causing this phenomenon is the violation of the i.i.d. assumption across training and validation data. This behavior was common during training on the strict split and again highlights the challenge for models to generalize to unseen epitopes from the sparsity of the current datasets.

While TITAN’s generalization capabilities to unseen epitopes are still limited, the performance is moderately good. This suggests that, despite their sparsity, the currently available datasets do contain enough information to enable generalization to unseen epitopes to a certain degree.

We find that all of our model settings outperform ImRex on both independent test sets by a large margin. Moreover, our K-NN baseline model clearly outperforms ImRex on seen epitopes.

Table 4. Comparison of TITAN with prior work. All but the ImRex model (shaded in gray) are contributions of this work. Models were trained on identical data and tested on an independent test set to ensure a fair comparison.

| Model          | seen epitopes | unseen epitopes |
|----------------|---------------|-----------------|
| ImRex          | 0.61          | 0.50            |
| K-NN (Baseline)| 0.79          | 0.37            |
| AA CDR3        | 0.83          | 0.69            |
| AA full        | 0.81          | 0.64            |
| SMI CDR3       | 0.85          | 0.72            |
| SMI full       | 0.86          | 0.64            |
| Pretrained     | 0.79          | 0.78            |
| Pretrained + Aug| 0.83         | 0.65            |
| Pretrained semifrozen | 0.77   | 0.69            |
| Pretrained semifrozen + Aug | 0.87 | 0.60 |

Moreover, we see that the base models (AA CDR3, AA full, SMI CDR3, SMI full) all perform better on the independent unseen epitopes test data than during 10-fold crossvalidation, while for the pretrained model settings, performance is comparable to the crossvalidation. For the unseen epitopes test set we observe a similar trend as above in the strict split crossvalidation. All models show performances clearly better than chance, with the pretrained semifrozen model with augmentation even achieving a ROC-AUC of 0.78. However, one needs to keep in mind, that the sample size for the unseen epitopes test is at only 1500, making especially high (or low) scores likely to be statistical outliers.

4 Discussion

In this work, we present TITAN, a generic, bimodal, sequence-based neural network for predicting TCR-epitope binding probability that significantly outperforms the state-of-the-art. We compare several settings of TITAN that differ in their inputs for TCRs and epitopes. While we restrict ourselves to the TCR3 chain, we find that inputting the complete variable TCR sequence boosts performance compared to scenarios where only the CDR3 region is used.

Notably, we are, to the best of our knowledge, the first to formulate TCR-epitope binding prediction as a subclass of the commonly investigated task of predicting compound-protein-interaction. The concomitant change of representing epitopes as SMILES (an atomic, more granular representation) instead of amino acid sequences improves the predictive power of TITAN. This reformulation not only enables data augmentation — by exploiting the non-uniqueness of SMILES (Bjerrum et al. 2017) — to enrich the training data, but also positions us to leverage large-scale datasets such as BindingDB for pretraining. Pretrained TITAN models achieve considerably higher scores than all TITAN base models.

Freezing the weights of the epitope input channel coupled with SMILES augmentation gives us the best performance with a mean ROC-AUC of 0.868 ± 0.005 on the TCR split.

To assess model performance in a more rigorous fashion, we designed a K-NN classifier based on the Levenshtein distance as a baseline model. Surprisingly, this simple model achieves a mean ROC-AUC score of 0.779 ± 0.007 on the TCR split crossvalidation. This model surpassed the performance of complex neural networks from previous publications (Moris et al. 2020) and is only outperformed by our pretrained...
TITAN models, which highlights the importance of including relevant baseline models.

As a final assessment of model performance, we compare TITAN to the current state of the art for general TCR specificity prediction, ImRex. To ensure fair comparison, we train our models on the ImRex training data and evaluate the performance on an independent test set derived from a different database. We demonstrate that both pretrained and base TITAN models, as well as the K-NN baseline, outperform ImRex by a large margin. The best result is again achieved by the pretrained semifrozen augmented TITAN model, with a ROC-AUC of 0.87 on epitopes included in the training data.

Finally, the main challenge for generic TCR–epitope interaction prediction remains the generalization to unseen epitopes. Here, we report two major breakthroughs. First, we demonstrate that TITAN exhibits moderate performance on unseen epitopes, where the best TITAN model, with a ROC-AUC of 0.87 on epitopes included in the training data. The result is again achieved by the pretrained semifrozen augmented TITAN model, as well as the K-NN baseline, outperform ImRex by a large margin. The best baseline models.

5 Acknowledgements

The authors acknowledge funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 813545 and 826121. We thank Matteo Manica and Joris Cadow for building up much of the model architecture used, as well as for many helpful discussions.

References

D. V. Bagaev, R. M. A. Vroomans, J. Samur, U. Sterrybo, C. Rius, G. Dolton, A. Greenfield-Watson, M. Attil, E. S. George, I. V. Zvyagin, N. Babel, D. K. Cole, A. J. Unitt, A. K. Sewell, D. M. Chudasama, F. Lacaud, and M. Shugay. VDD1b in 2019: database extension, new analysis infrastructure and a T-cell receptor motif comprehend. Nucleic Acids Research. 48:1057–1062, 2019. doi: 10.1093/nar/gkz874. URL: https://doi.org/10.1093/nar/gkz874.

J. Born, M. Manica, J. Cadow, G. Markert, N. A. Mill, M. Filipavicius, N. Ivanuk, A. Cardinale, T. Laino, and M. Rodríguez Martinez. Data-driven molecular design for discovery and synthesis of novel ligands: a case study on SARS-CoV-2: Machine Learning: Science and Technology. 2(2):026204, 2021a. ISSN: 2632-2133. doi: 10.1088/2632-2158/abaebe. URL: https://doi.org/10.1088/2632-2158/abaebe.

J. Born, M. Manica, A. Oskoee, J. Cadow, G. Markert, and M. Rodríguez Martinez. Pacmann2: De novo generation of hit-like anticancer molecules from transcriptomic data via reinforcement learning. (Science, 244):102269, 2021b. ISSN: 2589-0042. doi: https://doi.org/10.1016/j.sci.2021.102269. URL: https://www.cell.com/scto/content/early/2021/05/14/Science.2589-0042.c102269.

W. D. Chemister, A. Crickles, S. Mahajan, R. Visà, Z. Kosmolagi-Yazic, Z. Yan, J. A. Greenbaum, L. E. Jensen, M. Nielsen, S. Christley, et al. Tmatch: Predicting t-cell receptor specificity based on sequence similarity to previously characterized receptors. bioRxiv, 2020.

T. U. Consortium. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Research. 49(D1):D480–D489, 11 2020. ISSN: 0305-1048. doi: 10.1093/nar/gkaa1190. URL: https://doi.org/10.1093/nar/gkaa1190.

P. Dash, A. J. Fosse-Gurldarl, T. Herz, G. C. Wang, S. Sharma, A. Souquette, J. C. Crawford, E. B. Clemens, T. H. O. Nguyen, K. Kedzierska, N. L. La Gruta, P. Bradley, and G. P. Thomas. Quantifiable Predictive Features Define Epitope-Specific T Cell Receptor Repertoires. 547(7661):89–93, 2017. ISSN: 1476-4687. doi: 10.1038/nature2383. URL: https://doi.org/10.1038/nature2383.

N. De Neuter, W. Bitterman, C. Breinart, B. Cuyper, A. Meizis, P. Moris, A. Suls, V. Van Tendeloo, B. Oujimi, K. Laukens, and P. Meysman. The immunerace study: A prospective multicohort study of immune response action to covid-19 events with the immunecode™ open access database. Multidim. Syst. Signal Process. 31(2):473–500, 2020. doi: https://doi.org/10.1007/s11045-020-00410-3. URL: http://dx.doi.org/10.1007/s11045-020-00410-3.

J. N. Dines, T. J. Malley, E. Svejpa, H. M. Simmons, R. Taniguchi, M. Klinger, L. Baldo, and H. Robins. The immune assay: A prospective multicohort study of immune response action to covid-19 events with the immunecode™ open access database. medrxiv, 2020. doi: 10.1101/2020.08.17.20175158. URL: https://www.medrxiv.org/content/early/2020/08/17/20175158.

A. Elmagar, M. Heinzigner, C. Dallago, G. Rihasti, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, D. Bhowmik, et al. Potrans: Towards cracking the language of life’s code through self-supervised deep learning and high performance computing. arXiv preprint arXiv:2007.06225, 2020.

D. S. Fisher, Y. Wu, B. Schubert, and F. J. Theis. Predicting antigen specificity of single t cells based on tcr cdr3 regions. Molecular Systems Biology, 16(8):e9416, 2020. doi: https://doi.org/10.1038/s41396-019-02993-w.

A. Gaulton, A. Hersey, M. Nowotka, A. P. Bento, J. Chambers, D. Mendez, P. Matos, F. Atkinson, L. J. Bellis, E. Cibrián-Uhalte, M. Davies, N. Dolman, A. Karlsson, M. P. Magarillos, J. P. Overington, G. Papadatos, I. Smir, and J. A. Greenbaum, L. E. Jensen, M. Nielsen, S. Christley, et al. Tmatch: Predicting t-cell receptor specificity based on sequence similarity to previously characterized receptors. bioRxiv, 2020.
A. R. Leach. The ChEMBL database in 2017. *Nucleic Acids Research*, 45(D1):D945–D954, 2016. ISSN 0305-1048. doi: 10.1093/nar/gkw1074. URL: https://doi.org/10.1093/nar/gkw1074

S. Gielis, P. Morts, W. Bittmerius, N. De Neuter, B. Ogunjimi, K. Laukens, and P. Meysman. Detection of Enriched T Cell Epitope Specificity in Full T Cell Receptor Sequence Repertoires. 10-2820, 2019. ISSN 1664-3224. doi: 10.3389/fimmu.2019.02820. URL: https://www.frontiersin.org/article/10.3389/fimmu.2019.02820

M. K. Gibson, T. Liu, M. Butulija, G. Nicola, L. Hwang, and J. Chong. Bindingdb in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic acids research*, 44(D1):D1045-D1053, 2016.

J. Gramm, H. Huang, A. Nair, O. Halton, L. E. Wagner, F. Rubelt, X. Ji, A. Han, S. M. Kraus, C. Petras, N. Haas, C. S. L. Achakul, A. Sette, S. D. Boyd, T. J. Sccoli, D. M. Martinez, and M. M. Davis. Identifying Specificity Groups in the T Cell Receptor Repertoire. 547:7661:94–98, 2017. ISSN 1476-4687. doi: 10.1038/nature22976. URL: https://doi.org/10.1038/nature22976

G. B. Cui, N. O. Hadas, C. S. N. Jung, and A. Viallez. SMILES2Vec: An Interpretable General-Purpose Deep Neural Network for Predicting Chemical Properties. 8:pages, dec 2017. URL: http://arxiv.org/abs/1711.02034

R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheehan, J. Aguilera-Iparraguirre, T. D. Hietz, R. P. Adams, and A. Aspuru-Guzik. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Central Science*, 4(2): 268–276, Feb 2018. ISSN 2374-7951. doi: 10.1021/acscentsci.7b00572. URL: https://pubs.acs.org/doi/10.1021/acscentsci.7b00572

S. Henkoff and J. G. Henkoff. Amino acid substitution matrices from protein blocks. *Nucleic Acids Research*, 24(1):D1045–D1053, 2016.

E. Jokinen, M. Heinonen, J. Huismanen, S. Mustoja, and H. Lahelmaiski. TCRGP: Determining Epitope Specificity of T Cell Receptors. page 543232, 2019. doi: 10.1016/S443232. URL: http://pubs.acs.org/doi/10.1021/acs.jcheminf.9b00297

V. I. Jurtz, L. E. Jensen, A. K. Bentzen, M. C. Jepsen, S. Mahajan, R. Vita, K. K. Jensen, P. Marcatili, S. R. Hadrup, B. Peters, and P. Meysman. NoTCR: sequence-based prediction of TCR binding to peptide-MHC complexes using convolutional neural networks. *bioRxiv*, page 453706, oct 2018. doi: 10.1101/453706. URL: https://www.biorxiv.org/content/early/2018/10/02/453706

D. P. Kingma and J. L. Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR, dec 2015. URL: https://www.openreview.net/forum?id=Syk05oCzl

D. J. Laydon, C. R. Bangham, and B. Asquith. Estimating T cell repertoire diversity: Limitations of classical estimators and a new approach. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374(1765), aug 2019. ISSN 14712970. doi: 10.1098/rstb.2014.0291. URL: https://pubs.rstb.royalsocietypublishing.org/content/374/1765.10.1098/rstb.2014.0291.abstract

V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710, 1966.

M. Manica, A. Oskooei, J. Born, V. Subramanian, J. Sáez-Rodríguez, and M. Rodriguez-Martínez. Toward explainable anticancer compound sensitivity prediction via multimodal attention-based convolutional encoders. *Molecular pharmaceutics*, 16(12):4797-4806, 2019.

L. McIntyre, J. Healy, and J. Mervile. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

T. Mikkov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

P. Moris, J. De Paauw, A. Postovskaya, S. Gielis, N. De Neuter, W. Bittremieux, B. Ogunjimi, K. Laukens, and P. Meysman. Current challenges for unseen-epitope T cell interaction prediction and a new perspective derived from image classification. *Briefings in Bioinformatics*, 2020.

A. Mösch, S. Raffegerst, M. Weis, D. J. Schendel, and D. Frishman. Machine Learning for Cancer Immune Profiling Based on Epitope Recognition by T Cell Receptors, nov 2019. ISSN 1664021. URL: https://arxiv.org/abs/1911.01077

J.-W. Sidhom, H. B. Larman, D. M. Pardoll, and A. S. Baras. Deeptcr is a deep learning framework for revealing sequence concepts within t-cell repertoires. *Nucleic Acids Research*, 44(D1):D1045–D1053, 2016.

J. W. Sidhom, H. B. Larman, D. M. Pardoll, and A. S. Baras. Deeptcr is a deep learning framework for revealing sequence concepts within t-cell repertoires. *Nucleic Acids Research*, 44(D1):D1045–D1053, 2016.

P. Moris, J. De Paauw, A. Postovskaya, S. Gielis, N. De Neuter, W. Bittremieux, B. Ogunjimi, K. Laukens, and P. Meysman. Current challenges for unseen-epitope T cell interaction prediction and a new perspective derived from image classification. *Briefings in Bioinformatics*, 2020.