Predicting T-Cell receptor specificity

Adnan Abdulle
Department of Engineering Mathematics
University Of Bristol
Bristol, United Kingdom

[Redacted]
Department of Engineering Mathematics
University Of Bristol
Bristol, United Kingdom

[Redacted]
Department of Engineering Mathematics
University Of Bristol
Bristol, United Kingdom
[redacted]

[Redacted]
Department of Engineering Mathematics
University Of Bristol
Bristol, United Kingdom

ABSTRACT

T-cell receptors play a vital part in the human immune system's response to epitopes. The literature reveals that there tools available to enable the clustering T-cell receptors (TCRs), and the predicting of epitopes that bind to them. It however suggests that there are several challenges, including high TCR variability. Data on TCR sequences was collected from the VDjdb database. Distance matrix representations were then created for the alpha, beta, and paired alpha-beta chains using tcrdist3. PCA, followed by t-SNE was used for dimensionality reduction of the dataset. This was then compared with a deep learning approach. It was found that the deep learning approach was more effective at separating the data compared with dimensionality reduction using PCA and t-SNE. The TCRs were also clustered based on specificity using the GIANA algorithm which produced promising results. A prediction algorithm was also explored, to predict epitopes based on TCR sequences. This was done using two k-nearest neighbour classifiers. The model performs competitively on the epitope prediction task, and insight is provided into the shortcomings of the algorithm.

I. Introduction

The study of TCR specificity is at the forefront of immunological research, bridging computational biology and immunotherapy. TCRs are central to adaptive immune response, recognizing and eliminating pathogens infected cells. Evolution between pathogens and the immune system has necessitated the development of a vast repertoire of TCRs, capable of recognizing a specific antigen epitope. This diversity is achieved from a unique mechanism known as VDJ recombination, which assembles diverse TCRs from segmented genetic elements. Despite the importance of TCR specificity in understanding immune response, predicting TCR-epitope interactions remains a challenge due to the diversity of TCR sequences and the complex nature of their interactions with antigens. Advances in sequencing technologies and resources like the VDJ database have catalysed computational approaches to decipher TCR specificity, offering the potential to transform disease diagnosis and treatments.

This report outlines a methodology for predicting TCR specificity. It describes the creation of a distance matrix representation for TCR sequences, the application of dimensionality reduction methods on TCR sequence data. It explores TCR diversity and specificity through clustering and visualization techniques. The analysis extends to the development and validation of an algorithm to predict antigen specificity from TCR sequences. Through this work, the report aims to gain a better understanding of antigen specificity and TCR interactions.

II. Literature Review

Vujovic et al [1] explores TCR clustering methodologies like TCRdist, CDRdist, and GLIPH, alongside techniques such as X-ray crystallography and ELISPOT (Enzyme Linked ImmunoSpot), to understand antigen specificity. X-ray crystallography offers insights into TCR-peptide Major Histocompatibility Complex (p-MHC) binding sites, while ELISPOT measures T cell activation without revealing TCR sequences. ImmunoMap visualizes TCR repertoire diversity, further enhancing T-cell responses.

TCRdist [2] and CDRdist [3] focus on analysing sequence similarities, particularly on conserved motifs in the CDR3 region, which are pivotal for TCR binding. The study also recommends using short forms of amino acids (k-mers) to evaluate TCR similarity, which reduces informational noise and avoids the complications associated with TCRs of different lengths. Additionally, the GLIPH (Grouping of Lymphocyte Interactions by Paratope Hotspots) [4] algorithm identifies conserved motifs within TCR sequences that determine antigen specificity. Despite progress, TCR variability and the complex nature of TCR-pMHC interactions, complicate the prediction of TCR specificity, from sequence data alone [1].

[1] suggests incorporating three-dimensional structural data could improve model predictions of epitope specificities. Addressing TCR cross-reactivity, which is the ability of a T-cell receptor to recognize and bind multiple different antigen peptides, and extending data collection are crucial for refining models accuracy. The importance of an extensive dataset is further emphasized to manage similarity within training and testing sets, crucial to improve the model's performance.

Mayer-Blackwell et al. [5] introduces meta-clonotypes, these are biochemically similar clusters of TCR sequences, to enhance biomarker identification. Using tcrdist3, to measure distances between TCR sequences based on amino acid properties, TCRs are grouped into meta-clonotypes. These are defined using centroids from antigen-associated TCR

clusters identified through methods like peptide-MHC multimer sorting. Meta-clonotypes improve the detection of antigen specific TCRs by including a radius within the TCR sequence space that optimizes sensitivity and specificity, reducing false positives across diverse patient samples.

Several challenges and suggested improvements were recommended. Variability in TCR sequencing data quality challenges consistent meta-clonotype identification. Expanding TCR data can enhance model generalizability. The tcrdist3 algorithm requires continuous refinement to manage the diversity of TCR sequences and disease variations effectively. Advanced machine learning techniques, like deep learning, could improve model accuracy and efficiency [5].

Huang et al. [6] introduces GLIPH2, an algorithm that analyses TCR response against Mycobacterium tuberculosis (Mtb). GLIPH2 uses a BLOSUM matrix for global sequence comparisons and CDRS motifs for local similarities. The study seeks to understand CD4+ T-cell responses by mapping their specificity to epitopes within the Mtb genome. GLIPH2 incorporates artificial antigen-presenting cells (aAPC) to present Mtb proteins on TCRs. This facilitates the antigen discovery process, across 95% of the Mtb proteome.

A total of 19,044 TCRβ sequences from individuals with latent Mtb infections were analysed using GLIPH2, an advanced version of GLIPH designed to process millions of sequences accurately and efficiently. The algorithm improves TCR analysis by enabling TCRs to be assigned to multiple clusters and thus mitigating the 'small world' effect, where TCRs are mistakenly clustered together by superficial similarities instead of true biological relationships. It employs Fisher's exact test for more precise statistical analysis, improving over the sampling method. This enables GLIPH2 to identify and group-specific TCR clusters by epitope specificity, including some that were previously unrecognized [6].

However, GLIPH2 faces challenges in predicting which Human Leukocyte Antigen (HLA) alleles present specific peptides via MHC molecules, due to the intrinsic complexity of TCRs and the limitations in epitope recognition. Consequently, it could miss clusters associated with HLA alleles or unique TCR sequences. Future improvements could focus on enhancing the algorithm's ability to handle diverse HLA backgrounds and optimize clustering algorithms to cover broader range of TCR specificities [6].

Zhang et al. [7], addresses the limitations of previous methods, like TCRdist and GLIPH which struggle to analyse large datasets due to high computational demands. GIANA (Geometric Isometry-based TCR AligNment Algorithm) addresses this by employing geometric isometry and using multidimensional scaling to transform sequence alignment into a high-dimensional nearest neighbour search.

The study also reveals that GIANA achieves speeds up to 600 times faster than TCRdist without sacrificing accuracy, but also processes 100,000 sequences in 23.9 seconds, while also maintaining a high clustering accuracy and purity rate of 96%, outperforming GLIPH2's 35%. GIANA's utility is further validated in real patient datasets, it accurately differentiated TCR repertoires of COVID-19 patients from cancer patients. It maintained over 99.99% specificity at sensitivities ranging from 20% to 50% for Covid-19 specific TCRs [7].

Despite its efficacy, GIANA has limitations. It requires sequences to be of equal length due to its isometric embedding framework and does not account for HLA allele differences that are crucial for TCR specificity [7].

III. Methodology

The following tasks were carried out as part of the methodology:

A. Generating TCR distances

First, tcrdist3 [5] was used to create distance matrix representations for alpha chains, beta chains, and then paired alpha and beta chains. Tcrdist3 is an open-source Python implementation adapted from the original tcr-dist package introduced by Dash et al. (2017)[8]. In total, three distance matrices were created; one matrix for alpha chains, one for beta chains, and one matrix for paired chains. The process to create the distance matrices was as follows:

The dataset consists of both TCR alpha chains and beta chains. From this dataset, a table was derived that contains only alpha chains, one that contains only beta chains, and finally, one table that contains paired alpha and beta chains combined into single rows. After this, tcrdist3's TCRrep class was used for the creation of distance matrices using each of these tables to represent the distances between all TCRs.

These matrices served as a foundation to reduce the TCRs distance matrices to two dimensions. Reducing the data into lower dimensions makes it possible to see the underlying structures within the dataset, and identify class separation, outliers, and more.

B. Dimensionality reduction and clustering

TCRDist3 uses the BLOSUM62 substitution matrix [9] to compute distances between CDR3 regions. Algorithms within the TCR domain are posed with the challenge of aligning amino acids of different lengths. Given two CDR3 chains of different length, TCRdist3 places successive gaps in the shorter CDR3 chain. These gaps are placed to minimize the penalty in the BLOSUM62 substitution matrix. This distance is calculated between every pair of amino acid sequence in the alpha, beta and alpha-beta chain tables. For each amino acid the resulting feature vector corresponds to its distance from every other TCR.

Distance-based approaches are one of several paradigms to emerge in epitope specificity prediction. This family of methods has two major limitations, the first being their awkwardness in handling amino acid sequences of different lengths. Also, such metrics rely on sequence-to-sequence comparison, which may prove ineffective in modelling complex non-linear interactions where highly similar amino acid sequences differ in epitope specificity.

This motivates an interest in the use of deep learning methods that address these weaknesses and learn a feature representation directly from the amino acid sequence data. Several deep learning algorithms [10] have demonstrated strong performance in epitope prediction tasks. The focus here is on the appliance of transformer architectures for modelling antigen specificity. Such models natively handle sequences of different length and can learn the complex rules governing the order of amino acid sequences. TCR-BERT [11] is a bidirectional transformer model which undergoes a semi-supervised pretraining phase and then task-specific training on an epitope prediction task. To obtain a feature representation, each beta chain sequence is passed through TCR-BERT and the last latent representation is extracted.

In both the TCR-Dist and deep learning approaches, the resulting feature spaces are greater than 1000 dimensions. It is appropriate to use dimensionality reduction methods to compress the feature space into a more compact representation. T-Distributed Stochastic Neighbour Embedding (T-SNE) is a non-linear reduction method which computes the similarity between data points in the high-dimensional space and optimizes the low-dimensional representation by minimizing the Kullback-Leibler divergence between the high-dimensional and low-dimensional distributions. This is an effective algorithm for revealing patterns within datasets and clustering. T-SNE's high computational cost means that it is often unfeasible for feature spaces greater than 50 dimensions. In such instances principal component analysis is used to reduce the feature space to an intermediate dimension and then t-SNE is applied on the reduced feature space.

For the alpha, beta, and alpha-beta chain's TCRDist representation, PCA is used to reduce the feature space to 50 components and t-SNE is applied to the reduced vector. The same is implemented for each beta chains' hidden representation from TCR-Bert. This facilitates qualitative and quantitative evaluation of each feature representation. For the latter, the adjusted random mutual information score is used after applying k-means clustering to the low-dimensional representations of the sequences, with the epitope labels used as ground truth. There are two reasons for choosing to visualise and cluster only the top seven most frequently-binded epitopes. Firstly, including more epitopes would make analysis of the visualisation problematic. More importantly, there are many infrequently occurring antigens in the dataset, for which it would be difficult to accurately cluster. Of the 1100 antigens in VDJDB database, approximately 100 account for 70% of TCR-epitope pairs [12].

C. Clustering

Additionally, the TCRs were clustered using GIANA 2.0. The details of how this was done are as follows:

The GIANA algorithm is a mathematical framework that transforms the CDR3 sequences in order to convert the traditional sequence alignment and clustering problem into a classic nearest neighbour search within the high-dimensional Euclidean space. Applying multidimensional scaling (MDS) to the isometric embedding of the BLOSUM62 matrix, GIANA generates a numerical vector for each amino acid sequence represented as coordinates within the high-dimensional space. The Euclidean distance between two represented string sequences is equivalent to their Smith-Waterman alignment scores.

GIANA4.1 was selected from the open-source Python package over GIANA4 and the GIANAsv variation, which employs stacked MDS vectors as the input CDR3 strings. The input data was pre-processed according to the steps detailed under *Section IV*. To remove duplicate values, the CDR3-V-J bio-identities encoding method for the alpha and beta chains was adopted from [13]; a given TCR is encoded in the format CDR3x_TRxV, where x is the specified chain segment, and TRxV is the corresponding chain segment gene for the variable region. This ensures the included TCRs contained either a TRA or TRB gene.

D. Limitations of crude one-hot approach

A consequence of the V(D)J recombination process is that lengths of CDR3 regions can vary which is important to generate the diversity necessary for effective immune response. This poses a challenge to a naive one-hot encoding approach since it is unclear how such an algorithm could effectively handle sequences of different lengths. Padding TCRs to the maximum amino acid length would be ineffective as a downstream supervised model would be unable to differentiate between padding and a meaningful data point. Using a one-hot approach each TCR is processed independently and contextual relationships between neighbouring amino acids would be lost, which are important in determining antigen specificity. Given two TCRs containing the same amino acids but in different positions, a one-hot approach treats them equivalently, overlooking important sequential information. These shortcomings make it difficult to build a robust model using this feature representation.

A solution would be to replace the one-hot encoding with a numerical representation of each TCR. To develop a feature vector for an amino acid, a comparison must be made with every other TCR in the dataset. This would result in a fixed-size representation of every amino acid and makes only local alignment of pairs of amino acid sequences necessary. This local alignment could maximize the similarity between a pair of TCRs. The next step would involve sequence-to-sequence comparison to assess the similarity of the two TCRs. This would correspond to one value in each TCR's numeric representation. A more sophisticated version of this approach forms the basis of the TCRDist algorithm. Deep learning models, such as recurrent neural networks (RNNs), which handle sequences of arbitrary length and learn end-to-end represent a separate methodology for overcoming these issues and more effectively learn the sequential information within a TCR.

E. *Prediction algorithm*

1) *Multi-classification using K-nearest neighbour (KNN)*

The final task in the report involves building a model to predict a TCR's antigen specificity. It is decided to build a classifier for only the top seven most occurring epitopes in the VDJdb database. As discussed, the VDJdb database is highly imbalanced towards a small number of epitopes. This increases the difficulty of building a classifier for all epitopes. A characteristic of TCRs is that they can bind to multiple antigens. However, it is decided to not account for this in the predictive algorithm since this is occurs for less than 5% of the filtered alpha, beta and alpha-beta chain datasets. After pre-processing, a k-nearest neighbour classifier was built and trained for the epitope prediction task. Note for a given data point that its distance is computed only between itself and the other points in the training set to prevent data leakage. The parameters of the algorithm were optimized using grid search. The performance of the algorithm was evaluated using the F1 score at an epitope-specific level and for the entire test set.

This is a commonly implemented methodology in epitope prediction tasks. This serves as a baseline model within the report.

2) *KNN reduced feature space*

The nearest neighbour algorithm is a non-parametric supervised model and thus the number of parameters grows with the size of the training set. This coupled with the high dimensionality of the feature space above makes it desirable to use a dimensionality reduction method to reduce its complexity whilst preserving information. It is chosen to use PCA for this task and the optimal number of components is chosen by inspecting when the explained variance is greater than 95%.

IV. Data Description / Preparation

VDJdb was used as the data source for the project. VDJdb is a database of T-cell receptor (TCR) sequences, each mapped with an antigen epitope. The database contains a table with columns that include the following:

TABLE I. COLUMNS USED IN VDJDB DATABASE

Column	***Description***
Gene	Denotes where this is a TCR alpha or beta chain (e.g. TCRB represents a beta chain for the CDR3 sequence).
CDR3	Represents the CDR3 sequence.
V, J	Represent the V and J segments of the TCR.
Species	Represents the source species of the TCR, such as human or mouse.
Epitope	The amino acid sequence of the antigen's epitope.
Score	Confidence score. The higher the score, the more certainty that the epitope matched the relevant TCR.
MHC Class	Major Histocompatibility Complex- a class of the epitope-responsible for presenting epitopes to TCRs.

The data size to use from this database depends on how it is filtered, but there are as many as 92771 total rows in the table. The majority of TCRs in the table belong to the human species. The database contains TCRs mapped to many epitope species, such as SARS-CoV-2, Cancer, Flu, HIV, and more. There are many epitopes which are categorized into each of these species.

The following preprocessing was applied to the data to prepare for later tasks:

1. Extract the following columns: Gene, CDR3, V, J, MHCA, MHCB, MHC Class, Epitope, Epitope gene, Epitope species, Complex ID, Score (minimum 1 or more.)
2. Limit data to: Human species, TRA, TRB, and Paired (alpha and beta chains).
3. Remove any rows that have empty values and remove duplicate rows.

The decision to include or exclude columns was based on studies like [12]. Furthermore, the dataset was limited to records with at least 1 or higher confidence score. This choice was made based on work from [1] and others.

V. Results and Discussion

A. *Dimensionality reduction results*

Figures [1-4] shows the low dimensional representations of the TCR distance vectors for the amino acids in the alpha, beta, and alpha-beta datasets in addition to the beta chain's last hidden state using TCRBert.

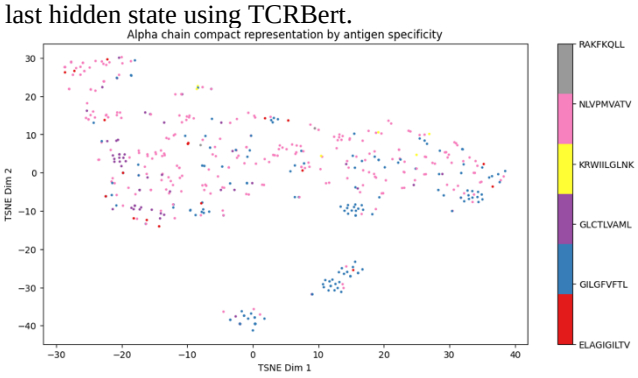

FIGURE 1. ALPHA CHAINS BY ANTIGEN SPECIFICITY AFTER DIMENSIONALITY REDUCTION (PCA FOLLOWED BY TSNE) – TCRDIST REPRESENTATION

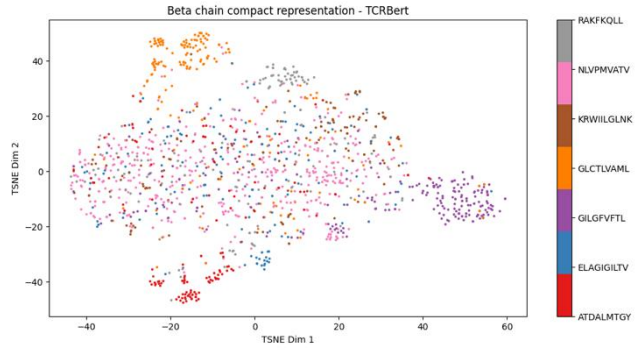

FIGURE 2. BETA CHAINS BY ANTIGEN SPECIFICITY AFTER DIMENSIONALITY REDUCTION (PCA FOLLOWED BY TSNE) -TCRBERT

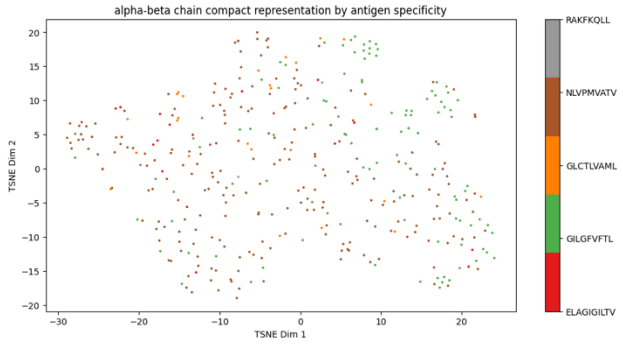

FIGURE 3. ALPH-BETA CHAINS BY ANTIGEN SPECIFICITY AFTER DIMENSIONALITY REDUCTION (PCA FOLLOWED BY TSNE) – TCRDIST REPRESENTATION

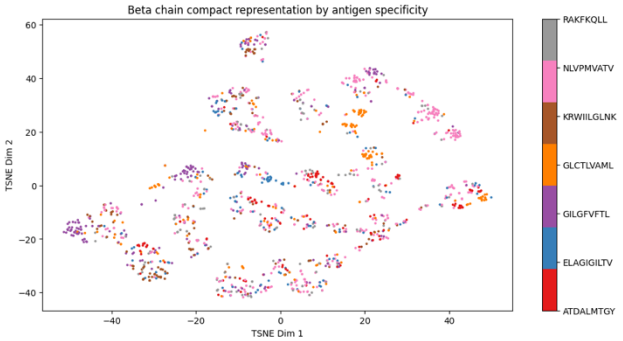

FIGURE 4. BETA CHAINS BY ANTIGEN SPECIFICITY AFTER DIMENSIONALITY REDUCTION (PCA FOLLOWED BY TSNE) – TCRDIST REPRESENTATION

TABLE II. APPLIED CLUSTERING METHODS AND RESULTS

Method	**Chain**	**Adjusted Mutual= information Score**
TCRDist	α	0.0674
TCRDist	β	0.0600
TCRDist	α-β	0.0575
TCRBert	β	0.1621

Table II displays the results of the clustering applied to each compact representation and shows that the TCRBert embedding most effectively separates TCRs into epitope-specific clusters. The adjusted mutual information score of 0.1627 indicates the clustering is materially above what would be expected if clustering was performed randomly, without closely matching the ground truth clusters. It is well established that the multi-head attention mechanism allows the transformer to learn highly contextualised representations from input sequences, but it is necessary to suggest one of numerous reasons why this contextualised learning yields performance improvement from the TCRDist representations. There is a paradox in TCR specificity where similar amino acids may bind to different epitopes. This has been attributed to 'hotspot residues' in the CDR3 loops determining the binding to an antigen [14] outside of which more sequence diversity is permitted. This means specific parts of the input sequence are more important in determining antigen specificity. This is a relationship that is likely captured over the training of a transformer but is more difficult using sequence-to-sequence comparison. TCR-Bert has some limitations including its lack of interpretability. Another restriction is that transformers require larger amounts of data to learn thus the size of the filtered VDJdb data is likely insufficient to effectively train a transformer algorithm. Building a deep learning classifier is beyond the scope of this paper, rather TCR-BERT serves as a different paradigm for epitope prediction which addresses some of the shortcomings the distanced-based model poses.

Table II shows a marginal improvement in clustering in the TCRDist's beta chain compared to the alpha chain. The beta chain of a TCR is widely regarded as more predictive of epitope binding than the alpha chain, owing to its greater diversity and its exposure on the surface of the TCR. For the TCRDist feature space, the three clusterings slightly improve on random clustering, with the highest mutual information (AMI) only 0.067. However, in Figure 4, localised clusters of epitope-specific TCRs can be observed. This is explained by TCRs binding to the same antigen using alternative docking topologies [15]. This suggests that the several of the small clusters in Figure 4 may be binding to the same epitope in different ways. Thus, these groups should be considered distinct, suggesting AMI may be a misleading metric. Alternative docking topologies prove a challenge to clustering using TCRDist or TCRBert.

Recent work [7] has underlined the performance benefits of using paired alpha-beta sequences in epitope specificity tasks. However, Table II shows the clustering performance of the TCRDist beta chains surpasses the alpha-beta sequences. This contradiction is explained by the limitations of the VDJdb dataset. Bulk sequencing methods have been applied to TCRs one chain at a time or just to the beta chain. This means that paired chains are insufficiently represented in VDJdb and thus the dataset used for alpha-beta was significantly smaller than either of the single-chain datasets. More recently single-cell technology has been used to generate paired data which mitigates this issue.

B. Clustering results

TABLE III. CLUSTERING BASED ON SPECIFICITY USING GIANA

Chain	***Adjusted Mutual Information score***	***Cluster Purity score***
α	0.7241	0.788

β	0.8401	0.861

Table III shows the results of clustering using the GIANA algorithm. The chosen metrics are AMI scores, implemented using scikit-learn, and cluster purity, defined as the percentage of TCRs specific to the most common epitope within a given cluster. The beta chain clustering produces a higher AMI score than that of the alpha chain, perhaps because the CDR3 beta region has more contact with a given epitope than the CDR3 alpha region [1]. Another possible explanation could be an unintended bias in the way the default hyperparameters of GIANA and other clustering models are configured. [17] shows they tend to be optimised for beta chain data, as beta chain data makes up most of the published TCR-epitope pair data.

Nevertheless, with AMI scores of 0.7241 and 0.8401 for the alpha and beta chains respectively, the GIANA algorithm is shown to be able to cluster TCRs well based on specificity. Table IV shows that for both chains, every cluster produced by GIANA is considered small in that each cluster contains less than or equal to 10 TCRs, showing that by avoiding producing larger clusters, GIANA can largely avoid impure clustering. This is reflected in the high cluster purity scores for both chains.

TABLE IV. GIANA PRODUCED CLUSTER PROPERTIES

Chain	***Number of sequences***	***Number of clusters produced***	***Number of small clusters***
α	404	54	54
β	634	77	77

GIANA's performance can be explained by the underlying approach to convert the clustering problem into a nearest neighbour search. Following the MDS scaling on the CDR3 sequences, GIANA applies the Faiss library to assist in identifying neighbouring vectors. Faiss enables efficient similarity search by indexing a given set of vectors, upon which the most similar set of vectors are identified [16]. GIANA will then run a K-mer guided alignment upon these vectors to produce the final TCR clusters.

C. KNN results

1) Multi-classification using KNN.

Table V shows the macro F1 scores for the six classifiers built, by chain used and the size of the feature space. For the reduced feature representations, it is decided to compress the feature spaces to 30 components, which retains approximately 95% of the variance from the original feature spaces. This is chosen from analysis of the variance contributed by each additional component in the three distance matrices. The variance from each added component for the beta chain is shown in the appendix section. For each of the distance matrices and reduced feature spaces a nearest neighbour classifier was built, the results are as follows:

TABLE V. SINGLE-LABEL KNN BENCHMARK RESULTS: A

Chain	**Full Feature Space - Macro F1 Score**	**30 PCA Components – Macro F1 Score**
α	0.591	0.591
β	**0.611**	**0.609**
Paired	0.457	0.455

Table V shows that for the full feature representation, the highest macro F1 score is achieved using the beta chain distance matrix. The beta chain is commonly regarded as the most predictive of TCR-epitope binding thus it makes sense that this exhibits the best performance. However as established in Section V Part A, research shows the benefits of using paired sequences. For similar reasons to those outlined in Section V Part A, the alpha-beta chain matrix performs poorly on the dataset. An example of the dataset imbalances is illustrated by the fact that for three of the seven epitopes in the alpha-beta matrix, there are less than 5 TCR observations in the test set. The alpha-beta representation has the strongest performance for epitopes where there is an adequate amount of paired data within the database. For example, for the 'GILGFVFTL' epitope, the alpha-beta chain representation has a 0.04 higher F1 score than when using only the beta representation.

Table VI shows that the reduced feature representation achieves similar performance to the full distance matrices for the alpha, beta and alpha-beta chain, which was the desired result. This shows the feature space's complexity can be significantly reduced without affecting performance.

TABLE VI BETA-CHAIN PERFORMANCE ACROSS EPITOPES – REDUCED DIMENSION (NUMBER OF NEIGHBORS = 20)

Chain	**F1 Score**	**Average distance from 20 nearest neighbours**
NLVPMVATV	0.75	111.44
GILGFVFTL	**0.81**	**76.81**
GLCTLVAML	0.74	101.72
RAKFKQLL	0.33	117.68
ELAGIGILTV	0.42	115.66
KRWIILGLNK	0.46	112.26
ATDALMTGY	0.70	110.60

Table VI shows the F1 scores of the individual epitopes for the reduced representation of the beta chain after applying KNN. It demonstrates that performance varies significantly across the different epitopes. There is a 0.48 difference between the F1 scores for the best and worst-performing epitopes. Further analysis into the relationships between TCRs in the test set and the neighbours used to classify the TCR were conducted. Evaluation of average distance column shows that TCR-epitope pairings with the highest average distance to their nearest neighbors had the worst F1 score whilst the TCR-epitope pairings with the smallest average distance had the best F1 score. This demonstrates a fundamental characteristic of feature spaces derived from TCRDist. TCRDist is based on predefined metrics which is

effective for TCRs which follow a structure that aligns well with TCRDist's established rules but is ineffective for TCRs with less common motifs that occupy sparser regions in the feature space such as those TCRs with less common neighbours.

VI. Further Work and Improvement

Predicting T-Cell specificity will inevitably improve as dataset limitations in the domain are addressed. Single cell technology will allow more paired chain data to be collected and will address the epitope imbalances within the dataset. Section V part C shows the positive effect of using paired chains on predictive performance. Table VI suggests that the sparsity of data in some areas of the feature space resulted in weaker performance for some epitopes and thus dataset advances would directly improve performance of the predictive model. Even if model design was to stay fixed, improved datasets would significantly improve TCR-epitope binding specificity.

Within Section V, clustering using different feature representations and techniques has been performed. An expansion of this analysis would be to provide the TCR subsequence patterns characterizing these clusters. For example, a naïve approach could simply obtain the centers of each cluster and return the TCR sequence of this center.

The predictive model built in this paper utilized predefined rules to develop feature representations for TCRs. Moreover, TCRDist3 and other distance-based measures must contend with the awkward problem of aligning TCRs of different length. An obvious step would be to build on the TCRBert work in Section V Part A and build a deep learning model which can learn directly from the TCRs and natively handles different length sequences. To train a transformer would require a self-supervised pretraining step which would require increasing the size of the dataset for the transformer to learn effectively. Among other design considerations would be the inputs to the model as studies [18] have shown incorporating other TCR information beyond the beta chain has yielded effective results.

Following on from Zhang et al [7], future improvements suggested including the integration of HLA typing (identifying HLA genes) and adapting GIANA algorithm to accommodate variable-length sequences. This would enhance GIANA's applicability across different types of TCR data and improve its use in clinical applications.

VII. Conclusions

In this work, distance matrices were first generated for the alpha, beta and paired alpha-beta chains for TCRs in the VDJdb. The dataset was dimensionally reduced using t-SNE and PCA. Clustering was also performed using GIANA. A prediction algorithm was explored to enable the classification of T-cell receptor bindings to epitopes.

The GIANA algorithm is shown to be an effective method for clustering TCRs based on antigen specificity. Results show a greater success in creating small, pure clusters for the beta chain selection than the alpha chain selection, which could be due to the CDR3 region on beta chain having more contact than that of the alpha chain with an antigen epitope, therefore producing more diverse TCRs. Another possible explanation is an unseen bias in the configuration of model hyperparameters that has them optimised for working with beta chain data.

Turning our attention to the prediction algorithm, the results suggest that for the full feature representation the highest macro F1 score is achieved using the beta chain distance matrix. The F1 scores achieved for the individual epitopes based on the reduced representation of the beta chain after applying KNN, demonstrates that the model's performance varies significantly across the different epitopes.

From rules-based systems to end-to-end learning, this paper uses various approaches to address the TCR-epitope prediction problem. All these approaches display moderate effectiveness on the various tasks, but there is clearly much progress to be made in this domain to build a system capable of predicting TCR specificity for more than a small number of epitopes. Currently, deep learning techniques [11], [18] are the pre-eminent model class in the field, and the dimensionality reduction methods presented in Section V demonstrate the promise of such approaches. Dataset improvements will lead to better performance using the TCRDist representations implemented in this paper, which offer a greater degree of interpretability than deep neural networks.

References

[1] M. Vujovic *et al.*, "T cell receptor sequence clustering and antigen specificity," *Computational and Structural Biotechnology Journal*, vol. 18, pp. 2166–2173, 2020, doi: https://doi.org/10.1016/j.csbj.2020.06.041.

[2] Koshlan Mayer-Blackwell, Stefan Schattgen, Liel Cohen-Lavi, Jeremy Chase Crawford, Aisha Souquette, Jessica A. Gaevert, Tomer Hertz, Paul G. Thomas, Philip Bradley, Andrew Fiore-Gartland (2020). TCR meta-clonotypes for biomarker discovery with tcrdist3: identification of public, HLA-restricted SARS-CoV-2 associated TCR features. bioRxiv. doi: https://doi.org/10.1101/2020.12.24.424260

[3] N. Thakkar, C. Bailey-Kellogg, "Balancing sensitivity and specificity in distinguishing TCR groups by CDR sequence similarity," BMC Bioinformatics, vol. 20, p. 241, 2019. [Online]. Available: https://doi.org/10.1186/s12859-019-2864-8

[4] Glanville J, Huang H, Nau A, Hatton O, Wagar LE, Rubelt F, Ji X, Han A, Krams SM, Pettus C, Haas N, Arlehamn CSL, Sette A, Boyd SD, Scriba TJ, Martinez OM, Davis MM. Identifying specificity groups in the T cell receptor repertoire. Nature. 2017 Jul 6;547(7661):94-98. doi: 10.1038/nature22976. Epub 2017 Jun 21. PMID: 28636589; PMCID: PMC5794212.

[5] K. Mayer-Blackwell *et al.*, "TCR meta-clonotypes for biomarker discovery with tcrdist3 enabled identification of public, HLA-restricted clusters of SARS-CoV-2 TCRs," *eLife*, vol. 10, p. e68605, Nov. 2021, doi: https://doi.org/10.7554/eLife.68605.

[6] H. Huang, C. Wang, F. Rubelt, T. J. Scriba, and M. M. Davis, "Analyzing the M. tuberculosis immune response by T cell receptor clustering with GLIPH2 and genome-wide antigen screening," *Nature biotechnology*, vol. 38, no. 10, p. 1194, Oct. 2020, doi: https://doi.org/10.1038/s41587-020-0505-4.

[7] H. Zhang, X. Zhan, and B. Li, "Publisher Correction: GIANA allows computationally-efficient TCR clustering and multi-disease repertoire classification by isometric transformation," *Nature Communications*, vol. 12, no. 1, Sep. 2021, doi: https://doi.org/10.1038/s41467-021-25693-2

[8] Dash P, Fiore-Gartland AJ, Hertz T, Wang GC, Sharma S, Souquette A, Crawford JC, Clemens EB, Nguyen THO, Kedzierska K, La Gruta NL, Bradley P, Thomas PG. (2017). Quantifiable predictive features define epitope-specific T cell receptor repertoires. Nature, 547(7661), 89-93. https://doi.org/10.1038/nature22383

[9] Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A. 1992 Nov 15;89(22):10915-9. doi: 10.1073/pnas.89.22.10915. PMID: 1438297; PMCID: PMC50453.

[10] Weber A., Born J., Rodriguez Martínez M. (2021). TITAN: T-Cell receptor specificity prediction with bimodal attention networks. *Bioinformatics* 37, i237–i244. 10.1093/bioinformatics/btab294

[11] Wu K, Yost KE, Belk JA, Xia Y, Egawa T, Sathpathy AT, Chang HY, Zou J. TCR-BERT: learning the grammar of T-cell receptors for flexible antigen-xbinding analyses. bioRxiv. 2021. doi: 10.1101/2021.11.18.469186

[12] Hudson, D., Fernandes, R. A., Basham, M., Ogg, G., & Hashem, K. (2023). Can we predict T-cell specificity with digital biology and machine learning? Nature Reviews Immunology, 23(8), 511–521. https://doi.org/10.1038/s41577-023-00835-3

[13] Hudson, D., Lubbock, A., Basham, M., & Koohy, H. (2023). A comparison of clustering models for inference of T cell receptor antigen specificity. *bioRxiv*, 2023.08.04.551940. https://doi.org/10.1101/2023.08.04.551940

[14] Singh NK, Riley TP, Baker SCB, Borrman T, Weng Z, Baker BM. Emerging Concepts in TCR Specificity: Rationalizing and (Maybe) Predicting Outcomes. J Immunol. 2017 Oct 1;199(7):2203-2213. doi: 10.4049/jimmunol.1700744. PMID: 28923982; PMCID: PMC5679125.

[15] Leem, J., de Oliveira, S. H. P., Krawczyk, K. & Deane, C. M. STCRDab: the structural T-cell receptor database. *Nucleic Acids Res.* **46**, D406–D412 (2018)

[16] M. Douze et al., “The Faiss library,” arXiv.org, Jan. 16, 2024. https://arxiv.org/abs/2401.08281

[17] P. Moris, J. De Pauw, A. Postovskaya, S. Gielis, N. De Neuter, W. Bittremieux, B. Ogunjimi, K. Laukens, and P. Meysman, "Current challenges for unseen-epitope TCR interaction prediction and a new perspective derived from image classification," *Brief Bioinform.*, vol. 22, no. 4, pp. bbaa318, Jul. 2021. [Online]. Available: https://doi.org/10.1093/bib/bbaa318

[18] Croce, G., Bobisse, S., Moreno, D.L. *et al.* Deep learning predictions of TCR-epitope interactions reveal epitope-specific chains in dual alpha T cells. *Nat Commun* **15**, 3211 (2024). https://doi.org/10.1038/s41467-024-47461-8

APPENDIX

Link to GitHub project code base: https://github.com/UoB-DSMP-2023-24/dsmp-2024-groupol1