frontiers in Physiology

ORIGINAL RESEARCH
published: 08 September 2021
doi: 10.3389/fphys.2021.730908

Predicting T Cell Receptor Antigen Specificity From Structural Features Derived From Homology Models of Receptor-Peptide-Major Histocompatibility Complexes

Martina Milighetti[1,2], *John Shawe-Taylor*[3] *and Benny Chain*[1,3]*

[1] Division of Infection and Immunity, University College London, London, United Kingdom, [2] Cancer Institute, University College London, London, United Kingdom, [3] Department of Computer Science, University College London, London, United Kingdom

The physical interaction between the T cell receptor (TCR) and its cognate antigen causes T cells to activate and participate in the immune response. Understanding this physical interaction is important in predicting TCR binding to a target epitope, as well as potential cross-reactivity. Here, we propose a way of collecting informative features of the binding interface from homology models of T cell receptor-peptide-major histocompatibility complex (TCR-pMHC) complexes. The information collected from these structures is sufficient to discriminate binding from non-binding TCR-pMHC pairs in multiple independent datasets. The classifier is limited by the number of crystal structures available for the homology modelling and by the size of the training set. However, the classifier shows comparable performance to sequence-based classifiers requiring much larger training sets.

Keywords: T cell receptor, antigen prediction, TCR-pMHC binding, homology modelling, T cell

OPEN ACCESS

Edited by:
Tetsuya J. Kobayashi,
The University of Tokyo, Japan

Reviewed by:
Shunsuke Teraguchi,
Shiga University, Japan
Ryo Yokota,
National Research Institute of Police Science, Japan

***Correspondence:**
Benny Chain
b.chain@ucl.ac.uk

Specialty section:
This article was submitted to Computational Physiology and Medicine, a section of the journal Frontiers in Physiology

Received: *25 June 2021*
Accepted: *02 August 2021*
Published: *08 September 2021*

Citation:
Milighetti M, Shawe-Taylor J and Chain B (2021) Predicting T Cell Receptor Antigen Specificity From Structural Features Derived From Homology Models of Receptor-Peptide-Major Histocompatibility Complexes. Front. Physiol. 12:730908. doi: 10.3389/fphys.2021.730908

1. INTRODUCTION

T cells are key players of adaptive immunity. They are activated by the recognition of a cognate peptide, a short stretch of amino acids which is displayed on a major histocompatibility complex molecule (MHC, pMHC when bound to peptide). The recognition occurs via the T cell receptor (TCR), which is composed of two chains (normally an α and a β), both of which are generated by a process of random recombination and selection. The recombination gives rise to three hypervariable regions, the complementarity-determining regions—CDR1, CDR2, and CDR3. Among the three regions, CDR3 is the most variable as it is found at the junction of V(D)J recombination, and it can therefore incorporate a number of non-template insertion and deletion events, whilst CDR1 and CDR2 depend on the V gene selected in the recombination process and have therefore a lower number of possible sequences.

A number of TCR-pMHC complexes have been crystallised and their structures solved and they are collected in the Structural T-Cell Receptor Database (STCRDab, Leem et al., 2018). They have given us deeper understanding of TCR-pMHC interactions and how these are impacted by mutations, but also how structure and function are related. Examples include how cross-reactivity between bacterial and self antigens can drive disease (Petersen et al., 2020), how binding mode can

give different specificity profiles to TCRs binding the same peptide (Coles et al., 2020), and how binding orientation is determined by how the peptide is presented by the MHC (Singh et al., 2020).

The existing structures can also be mined for information on how the TCR interacts with the pMHC complex. By looking at the TCR residues that fall within 5 Å of the peptide in a number of published TCR-pMHC structures, both Glanville et al. (2017) and Ostmeyer et al. (2019) showed that the CDR3 is the region that makes the most extensive contacts with the peptide. These regions of contact are normally short stretches of 3 or 4 consecutive amino acids within the CDR3. Moreover, they noted that whilst the TCRβ always made contacts, there are multiple instances were the TCRα is not within contact distance of the peptide. It has also been shown that TCRs which recognise the same peptide share motifs and sequence characteristics in the CDR3 (Thomas et al., 2014; Cinelli et al., 2017; Dash et al., 2017; Glanville et al., 2017). This similarity may reflect structural similarities (Lanzarotti et al., 2019).

The ensemble of TCRs that are present within an individual at any point in time is called the TCR repertoire. Different sequences are found at widely different frequencies, ranging from a few hundred copies to 10^9 copies for the larger T cell clones, which make up to 1% of the total repertoire. Differences in clone size can arise both in the naive repertoire, by convergent recombination (whereby an amino acid sequence is likely to be produced by the process of recombination—normally with short CDR3 and few nucleotide insertions; Venturi et al., 2006; Britanova et al., 2014) or because of the power-law distribution of naive T cell clones produced by the thymus (de Greef et al., 2020); or in the memory repertoire by convergent selection, whereby similar sequences are expanded because they are responding to the same antigen, (Pogorelyy et al., 2018). de Greef et al. (2020) estimates the maximum effect of generation probability to be around 10^7, which is two orders of magnitudes smaller than the largest observed clone sizes, suggesting a role for expansion during the immune response. By focusing solely on the CDR3, it can be shown that during an immune response, expanded TCR clones are frequently part of clusters of sequences that are more similar to each other than might be expected by random sampling of the repertoire (Marcou et al., 2018; Joshi et al., 2019; Pogorelyy et al., 2019).

This observation of antigen-driven TCR sequence clustering has been used to build algorithms such as GLIPH (Glanville et al., 2017) and TCRdist (Dash et al., 2017), which can build sequence motifs starting from a cluster of TCRs known to recognise the same peptide and which are then able to find other TCRs responding to the same peptide. More recently, Tong et al. (2020) have shown that sequence information encoded in the form of overlapping amino acid quadruplets can be used to create a multi-class prediction algorithm able to correctly assign TCR-pMHC pairs.

In the same way that conserved sequence motifs characterise TCRs recognising the same antigen, we hypothesise that there will be structural features of the TCR/antigen interface which are conserved in the interactions. Such conserved structural features could be leveraged to gain a better understanding of the TCR-pMHC interaction and to recapitulate and improve what has been learnt from looking purely at sequence information. Our understanding of the physical interactions between TCRs and pMHC is, however, limited to the set of solved and published crystal structures. The STCRDab currently reports about 400 entries for $\alpha\beta$ TCR-pMHC complexes, and 120 different peptides, which is clearly a tiny subset of all the possible TCR-pMHC interactions that can exist. To solve this problem, a number of tools have been developed and subsequently optimised to predict the structure of a TCR-pMHC complex based on its sequence. One of these is TCRpMHCmodels (Jensen et al., 2019), which exists as a free online user interface. TCRpMHCmodels leverages LYRA (Klausen et al., 2015) to model the TCR structure and MODELLER (Fiser and Šali, 2003) to predict the pMHC structure, to then combine them together by using a third set of templates for the TCR-pMHC complex overall. Tools like TCRpMHCmodels, although still limited by the amount of information that has been published, allow us to delve deeper into the structural relationships between the TCR and the pMHC.

We show here that a combination of structural and sequence features can be incorporated into a machine learning algorithm to discriminate binding and non-binding TCR-pMHC pairs. The classifier presented is limited by the performance of the homology modelling, but, unlike any of the previous work reviewed above, it does not rely on the identification of a set of TCRs binding to a specific peptide to be able to predict whether other TCRs will bind to that same peptide, but rather learns some general rules which can predict TCR interaction with completely novel peptides.

2. METHODS

2.1. Datasets

The available crystal structures for TCR-pMHC complexes were retrieved from STCRDab (http://opig.stats.ox.ac.uk/webapps/stcrdab/, Leem et al., 2018). The dataset (referred to as STCRDab or PDB—Protein Data Bank—set) was refined to include only one complex per crystal, remove $\gamma\delta$ TCRs and remove non-peptide antigens. The set was then checked for repeat sequences. For the classifier step, TCRs binding MHC class II complexes were removed as these cannot be modelled by TCRpMHCmodels. To create non-binding TCR-pMHC pairs, random TCR-pMHC pairs were created from the available pool, under the condition that the pMHC from the random pairing was not the same as the original one.

The 10XGenomics dataset was downloaded from the 10XGenomics website (CD8+ T cells of Healthy Donor 1, 10XGenomics, 2020). For each TCR, binding (or absence of binding) to an epitope was defined as in the Application Note provided by 10X Genomics. Briefly, a specific binding event was defined as having UMI count higher than 10 and >5 times the highest negative control for that TCR clone. When a TCR clone was assigned multiple barcodes, the UMI counts for each tetramer were summed to determine overall binding. If these conditions were true for more than one peptide, the TCR was called a binder for each of the epitopes.

The Dash dataset (Dash et al., 2017) was obtained from the VDJDb dataset. Duplicate TCR-pMHC pairs were removed. Each unique TCR clone was paired with each pMHC in the dataset, making 1 binding and 9 non-binding complexes per TCR.

The set of experimental constructs (expt) consists of a set of experimentally-validated peptide-specific TCR constructs with cognate peptide, which have been characterised functionally: 2 CMV-reactive TCRs (NLVPMVATV peptide), 3 influenza-reactive TCRs (2 HA1-reactive—peptide VLHDDLLEA—and 1 HA2-reactive—YIGEVLVSV peptide), 1 EBV-reactive TCR (peptide CLGGLLTMV) from Thomas et al. (2019) and Chatterjee et al. (2019); A7 TCR and 3 affinity-matured TCRs from A7 which recognise pTax as well as pHud peptides (LLFGYPVYV and LGYGFVNYI, respectively; Thomas et al., 2011); two TCRs identified as neoantigen-reactive in Joshi et al. (2019) and two mutated versions of these, which have been shown not to bind the neoantigen (unpublished data, Woolston, personal communication, 2020). To create the non-binders, each TCRs was matched with each pMHC in the pool, as well as with peptide WT235 (control peptide in Thomas et al., 2019, CMTWNQMNL) and peptide WTlung (FAFQEDDSF, wild-type peptide for the neo-antigen; McGranahan et al., 2016).

A dataset of TCR-pMHC complexes with experimentally-determined affinity was retrieved from the ATLAS (http://atlas.wenglab.org/web/index.php; Borrman et al., 2017) to evaluate the impact of affinity on classifier performance. Any TCR-pMHC pair with undetectable binding (K_d labelled as *n.d.*) was removed from the set, as the binding status could not be determined. Any complex with $K_D < 200\mu M$ was called a binder and $K_D \geq 200\mu M$ called a non-binder.

Finally, a dataset of TCR-pMHC complexes with epitopes that are neither present in our training set nor in the training set of the tools we benchmarked against was downloaded from the latest version of the VDJDb (Bagaev et al., 2020). As for the PDB set, negatives were created by shuffling of TCR-pMHC pairs in the set.

The MHC alleles and number of pMHC structures for all the datasets is summarised in **Table 1**.

2.2. Homology Modelling and Feature Extraction

Each structure (both binders and non-binders) in these datasets was homology-modelled with TCRpMHCmodels (which was kindly provided in command-line form by the authors, Jensen et al., 2019) in its default settings and submitted to the feature-extraction pipeline. Structures were analysed for quality using Molprobity (data not shown; Williams et al., 2018).

To make the structures comparable, they were renumbered to the standardised IMGT numbering (Lefranc, 1997) using ANARCI (Dunbar and Deane, 2016). Moreover, the peptide residues were renumbered to 1-20, so that the central residues would be residues 10-11 in each complex.

For each TCR-pMHC, five sets of features were extracted, namely:

- minimum pairwise distances between each CDR residue and each peptide residue were calculated using BioPDB (Hamelryck and Manderick, 2003). These capture the binding mode of the TCR-pMHC complex.
- energetic profile of pairwise CDR-peptide residues interactions was calculated using PyRosetta v2020.28+ (Chaudhury et al., 2010). The Rosetta energy function for context-independent residue-residue interactions was used to extract the following terms (scorefunction: talaris2014) from a PDB file from which the MHC complex was removed: attractive and repulsive van der Waals (atr, rep), electrostatic interactions (elec) and solvation energy (sol) (Alford et al., 2017). These are a representation of binding energy of the complex.
- Atchley factors (Atchley et al., 2005) were used to encode the sequences of the peptide and CDRs for each TCR-pMHC pair.

To evaluate the effect of homology modelling performance on the classifier presented, the structures were categorised as having or not having good homology modelling templates. This was defined based on the sequence homology to the most similar peptide template ($>$ 45% sequence similarity to the best pMHC model template) and complex template ($>$ 60% sequence similarity to the best complex template). These thresholds were chosen based on the results presented in the original report (Jensen et al., 2019).

To be noted that not all structures could be successfully modelled by TCRpMHCmodels (because of lack of template, **Table 1**), and so we could not submit them to the feature extraction pipeline.

2.3. Multiple Kernel Learning

Each feature set was pre-processed separately. Missing values were imputed with the median value of the feature across the train set. Each feature was then scaled to have a value between 0 and 1 (scikit-learn Minmax scaler; Pedregosa et al., 2012) and normalised.

To properly represent and integrate the different features extracted from the structures, Gaussian (or radial-basis function, rbf) kernels were created separately for each subset of features. Since the optimum width of the Gaussian kernel (represented by the γ parameter) for each feature set was unknown, we incorporated a series of seven different kernels for each feature set each with a different Gaussian width parameter, chosen as described in the heuristic below. The seven kernels were combined during the learning step, with weights assigned to each kernel as described in Lauriola et al. (2017). This avoids having to find an optimal Gaussian parameter for each feature set. The γ parameters for the 7 Gaussian kernels for each feature set were chosen heuristically as follows:

1. the Euclidean distance, $d_{i,j}$, between each positive (binding, denoted with i) and negative (non-binding, denoted with j) examples in the train set was calculated

$$d_{i,j} = \sqrt{\sum_{w=1}^{n} (pos_{i,w} - neg_{j,w})^2}$$

for each feature w in the n-sized set.

TABLE 1 | MHC associated with peptides in each dataset.

	PDB (all)	PDB (ML)	Expt	Dash	10X	Atlas	NewVdj
A01	2	6			27 (25)		
A02	71	211	98 (98)	6,561 (5,754)	1679 (1,422)	303 (302)	5,510 (4,812)
A03					3,377 (2,922)		145 (126)
A11	2	5			1908 (1,673)		
A24	5	15			287 (239)	4 (4)	
A25							145 (126)
A68							435 (253)
B07	2				34 (29)		290 (254)
B08	4	10			273 (254)	55 (55)	
B27	2	7				1 (1)	
B35	16	33	28 (28)		2 (2)	67 (14)	145 (126)
B37	1	1					
B38							290 (127)
B44	3	11				6 (6)	145 (127)
B51	1	2					
B57	1	5					
DQ	10					34 (0)	
DR	14					9 (0)	
E	3	7				7 (7)	
H-2D	11	40		6,561 (5,434)			
H-2K	7	13		8,749 (7,423)		4 (4)	
H-2L	12	38					
IA	5					1 (0)	
IE	7					20 (0)	
Total	179	404	126 (126)	21,871 (18,611)	7,587 (6,566)	511 (393)	7,105 (5,951)

For each set analysed, the number of complexes with each MHC gene is shown. PDB (all) is used in the first section of the paper to analyse contacts, whilst PDB (ML) is the set used for the supervised learning. The numbers for PDB (ML), expt, dash, atlas, and newVdj represent the starting set of sequences that were submitted to the pipeline after removing duplicates, including both binding and non-binding complexes, as well as sequences that could not be modelled by TCRpMHCmodels. The number in brackets are the structures for which prediction was successful.

2. 7 σ values, corresponding to 1st, 2nd, 5th, 50th, 95th, 98th, and 99th percentile of distances, were retrieved
3. for each σ, a γ was calculated as:

$$\gamma = \frac{1}{2 * \sigma^2}$$

4. the calculated γ values were then used to create the 7 kernels for the specific feature set.

The family of kernels for different features were combined by the EasyMKL algorithm as implemented in MKLPy to find an optimal combination (Aiolli and Donini, 2015; Lauriola et al., 2017; Lauriola and Aiolli, 2020). The learner algorithm for MKL was set as scikit-learn's SVC (Support Vector Classification, Pedregosa et al., 2012). EasyMKL with SVC needs two hyperparameters: λ and C. λ was fixed to 0 as in Lauriola et al. (2017), and the optimal C parameter for SVC was searched in the range between 10^{-5} and 10^{2} by 10-fold (internal) cross-validation (CV) on the train set.

This same process was used both when a single feature set was evaluated (by combining the seven kernels for the set) as well as when combining multiple feature sets (seven kernels for each feature set).

To estimate performance by cross-validation, the train set was split 70-30. 70% was used to optimise the model parameters by maximising the ROC AUC score and the remaining 30% was used for prediction. The procedure was repeated 10 times with different subsets of samples.

Out-of-sample performance was evaluated in the datasets outlined in section 2.1, by training the classifier on the whole of the training set.

2.4. Impact of Homology Modelling

To evaluate the impact of homology modelling on the classifier, the structure prediction and feature extraction set were repeated by forcing TCRpMHCmodels to only use templates with sequence similarity below a set threshold.

TCRpMHCmodels uses three sets of templates (one for the TCR, one for pMHC, and one for the entire Complex), which can be manipulated independently. Here, we have set the maximum sequence similarity to be 40, 60, or 80 for each of these template sets separately and then for all three at the same time (All). From

the structures generated in this way, features were extracted and model performance was evaluated using 10-fold CV.

2.5. Benchmarking Against Other Classifiers

To evaluate the performance of the presented classifier compared to published classifiers in the field, we compared performance with ERGO (Springer et al., 2020) and ImRex (Moris et al., 2020) on the same validation sets. ERGO is available as a web tool (https://tcr.cs.biu.ac.il/), and the models trained on the VDJdb (Bagaev et al., 2020) were used for the benchmarking. ImRex is available as a GitHub repository (https://github.com/pmoris/ImRex), and the available model trained on the VDJdb was used for the predictions.

2.6. Data Availability

The complete set of sequences used, as well as prediction results are provided as **Supplementary Material**.

3. RESULTS

3.1. Extracting Physical Features From Available TCR-pMHC Complex Structures Allows Interrogation of Binding Mode

We first established a systematic pipeline to extract structural information about the TCR-peptide interface from a dataset of solved structures downloaded from the Structural T Cell Receptor Database (Leem et al., 2018). The minimum pairwise distances between TCR and peptide residues, and their corresponding attractive and repulsive van der Waals forces (atr, rep), electrostatic interactions (elec), and solvation energies (sol) were estimated for each peptide-TCR complex as described in the methods. Sequence information for each complex was also encoded in the form of Atchley factors for each TCR-pMHC pair.

Each feature extraction process yielded a matrix with information about pairwise contacts between residues in the TCR and residues in the peptide (**Figure 1A**). The distance fingerprints are easy to compare between different structures and can give insight into the binding mode for the complex: for instance, complexes 1AO7 (Garboczi et al., 1996) and 1MI5 (Kjer-Nielsen et al., 2003) (both MHC Class I) bind closer to the N terminus of the peptide, whilst 1D9K (Reinherz et al., 1999) has the TCR bound more centrally (**Figures 1A,B**).

We wondered whether any trends could be detected more generally and used the minimum pairwise distances to identify the distribution of interactions between TCR CDR residues and the peptide in class I and class II complexes (**Figure 1C**). While it is clear that interactions between TCR chains and antigen peptide are not confined to a single hotspot, some general patterns emerge. The TCRα chain, for example, tends to bind the N-terminus of the peptide, whilst the β binds toward the C-terminus, as has been reported previously (Garcia et al., 2009). Interestingly, while contacts were dominated by the CDR3 region of the TCR, we also detected contacts between CDR1 and CDR2 and peptide residues in a significant proportion of complexes. Moreover, more of the class I structures make contacts with the C-terminus of the peptide than class II. A similar pattern is also detected when looking at the energetic interactions (**Supplementary Figure 1**).

In order to look in more detail for potential conserved patterns with which to characterise the TCR-peptide binding surface, we performed a Principal Component Analysis (PCA) for each of the structural feature sets (distances and energy vectors), as well as the sequence information (in the form of Atchley factors) for all complexes (**Figure 2A** and **Supplementary Figure 2A**). The first dimension of the PCA of the minimum pairwise distances clearly identified the few examples where the TCR is in an inverse orientation relative to the peptide (stars, PDB: 4Y19 and 4Y1A, 5SWS and 5SWZ; Beringer et al., 2015; Gras et al., 2016). The second dimension of the distance PCA, on the other hand, seemed to partially discriminate between class I and class II complexes. To gain some insight in to which structural features were driving this separation, we looked at the distance vectors that were used for each structure (**Figure 2B**, left). Both for the α and the β chains, a shift toward the peptide C terminus was observed with decreasing PC2 values. Four representative fingerprints from the edges of the PCA plot are also shown in which the inverted orientation of 4Y19 and 5SWS as well as the shift toward the C terminus for 5TEZ (Yang et al., 2017) are apparent, compared to 3RGV (Yin et al., 2011). In agreement with **Figure 1C**, class II complexes tend to have higher PC2, which is associated with a shift toward binding at the N terminus of the peptide. 3RGV, which segregates with the class II complexes, is actually a class I complex. Interestingly, however, the YAe62 TCR in the 3RGV complex is reported by the authors to bind both class I and class II complexes with similar orientations, which might explain its positioning with other class II complexes. Strikingly, the other class I complex found with high PC2 is 4JRY, which is also reported to bind with unusual position on top of the N-terminus of the peptide, rather than centrally, where the peptide bulges out (Liu et al., 2013).

A similar analysis was done on the solvent energy vectors (**Figure 2A**, right). The PCA suggests a segregation between class I and class II complexes along PC1, although significant overlap was also observed. We therefore looked at what features could be driving the separation along the PC1 (**Supplementary Figure 2B**). The only evident trend was that all the complexes with high PC1 show a strong unfavourable interaction between the β chain and the peptide C terminus (blue in the heatmap). As solvent energy is positive (i.e., unfavourable) when a residue is not solvent-exposed, this suggests that the complexes with higher PC1 make an interaction between the beta chain and the C terminus of the peptide.

Finally, all distance, energy and sequence feature sets were combined in a single PCA plotted in **Figure 2C** (left). Here, the structures with inverted polarity have high PC1, followed by MHC class II complexes and on the left-hand side of the plot are the class I complexes. The loadings of each feature in the set were calculated and the features ranked by loading value (**Figure 2C**, right). Most of the features which had absolute values >0 (i.e., positive or negative), belong to the distance, the solvent energy

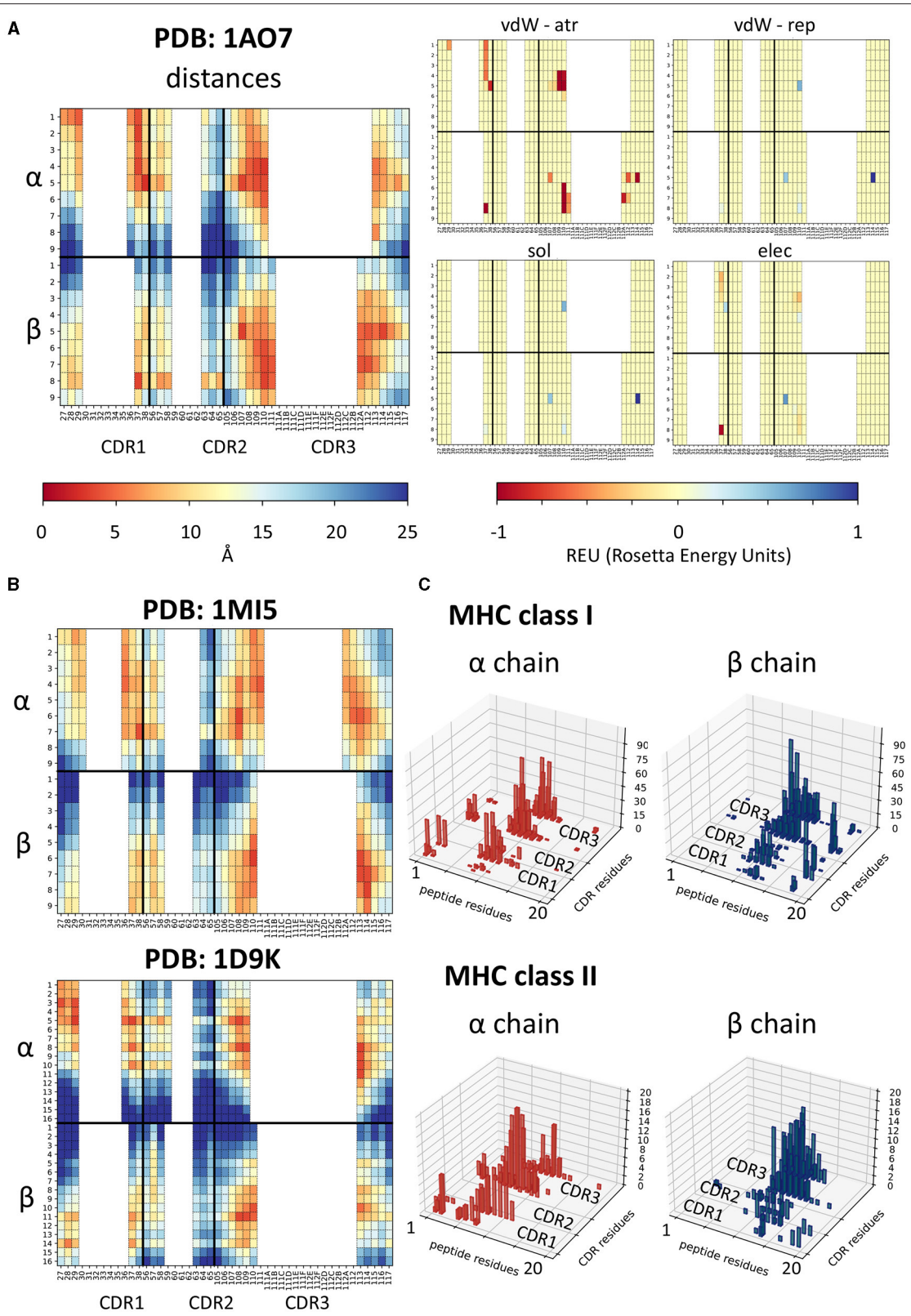


FIGURE 1 | Feature extraction from PDB structures. **(A)** Heatmaps showing the physical features extracted for structure 1AO7. In each heatmap, the top half refers to the α chain and the bottom half to the β chain. Each column is a CDR residue, each row is a peptide antigen residue and the colour of each square represents the

(Continued)

FIGURE 1 | value extracted for the CDR-peptide residue pair (i.e., top left-hand square of the distance panel is the distance between residue 1 on the peptide and residue 27 of the TCRα chain). Similar plots are shown for each energy term extracted—van der Waals attractive, van der Waals repulsive, solvent, and electrostatic. **(B)** Two other examples of distance fingerprints, a class I and a class II complex—1MI5 (class I complex, EBV peptide) and 1D9K (class II complex, conalbumin peptide)—for comparison with 1AO7. Same scale as in **(A)**. **(C)** Histograms showing the number of structures making a contact (<6Å) for each peptide residue-CDR residue pair, for alpha and beta chains separately, showed for class I and class II complexes. Peptide residues renumbered 1-20 for consistency as described in methods.

or to the Atchley factors datasets, suggesting that these have the strongest discriminatory power.

Overall, these results gave us confidence that meaningful information about the binding interface could be extracted with our pipeline.

3.2. Structural Information From Homology Modelled Structures Cannot Distinguish Binding Pairs in Unsupervised Settings

We next investigated whether given independently a TCR and a pMHC, we could determine whether we could discriminate between TCR-pMHC interactions in which the TCR binds its cognate antigen and those which do not allow effective binding. The parameters characterising non-binding interactions could obviously not be obtained directly from known structures, since by definition these TCRs would not form stable complexes with the pMHC. We therefore predicted structures for TCR-pMHC combinations by homology modeling using TCRpMHCmodels (Jensen et al., 2019). The pipeline takes a fasta file with a TCR, a peptide and a class I MHC, predicts its three dimensional structure and extracts pairwise distances and binding energies for the interface (**Figure 3**, steps 1 and 2). The actual sequences are also captured in the form of vectors of Atchley factors as described in the methods.

Because we needed to rely on a structure prediction method, we first evaluated the difference between the features extracted from the original crystallographic structures and from their respective modelled structures (**Figure 4** and **Supplementary Figure 3**). Taking complex 1AO7 as an example, the fingerprints obtained from the original PDB and from the predicted structures were plotted (**Figure 4A**). The two complexes have root mean square deviation (RMSD) of about 2 Å (calculated on all their Cα atoms) and it can be seen that the contacts seem to be slightly shifted toward the N terminus of the peptide in the predicted structure compared to the crystal. However, the two fingerprints did not look drastically different.

When combining all feature sets and looking at all structures available by PCA, no systematic difference was found between modelled and original structures (**Figures 4B,C** and **Supplementary Figure 3**). The only difference found between the two sets is in the repulsive van der Waals forces component (evident in **Supplementary Figure 3B**, where the values do not correspond for predicted and crystal structures). In fact, modelled structures are grouped on the left-hand side of the plot and are found to have higher MolProbity clashscores (data not shown).

On all other feature sets, there was reasonably good matching between the crystal structures and their homology models, although TCRpMHCmodels failed to predict non-canonical binding models. We also compared the distributions of some of the structural features (minimum distance between peptide and TCR, number of contacts and number of favourable interactions), and in general found reasonably good agreement between models and structures (**Figure 4C**). As homology modelling gave us reliable predictions and was necessary to create our negative examples, we decided to use modelled structures for both binding and non-binding complexes, in order to avoid introducing systematic bias.

To create a set of non-binders, a set of shuffled TCR-pMHC complexes from the STCRDab was used (**Figure 5A**). We then asked whether the structures predicted for non-binders could be discriminated from the binders.

Strikingly, there was no discernible separation of binders and non-binders on unsupervised PCAs with any of the distance or energy sets of features (**Figure 5B** and **Supplementary Figure 4**). Basic metrics such as the minimum distance between TCR and peptide and the number of contacts showed similar distributions for binders and non-binders (**Figure 5C**).

3.3. The Impact of Structural Information on Discrimination Between Binders and Non-binders Using Supervised Learning

We turned to supervised machine learning methods to try and better discriminate between binding and non-binding pairs (**Figure 3**, steps 3 and 4). We extracted sequence (Atchley factors) and structural features (distances, attractive and repulsive van der Waals interactions, electrostatic and solvent energies) from predicted TCR-pMHC structures from the training and test set using the pipeline described in the methods. We explored multiple kernel learning (MKL) as a tool to combine information from the different feature sets. To assess the potential of our method, a model was trained and tested by cross-validation, using predicted structures derived from the STCRDab, creating a dataset of positives and negatives as described in the methods. **Figures 6A,C** show the results of 10-fold cross-validation when each different feature set is used separately. Distances provide the single strongest predictive power (average ROC AUC of 0.791), and similar discrimination can be obtained by using Atchley factors (ROC AUC of 0.752), followed closely by attractive van der Waals forces (atr, ROC AUC of 0.746) and solvent energies (ROC AUC of 0.686). The other energetic terms generally showed poorer performance and were excluded from further analysis.

We next combined the feature sets to create a single classifier (**Figures 6B,C**). In general, combination of different feature sets did not improve performance on CV drastically compared to using the single sets. Interestingly, although performance did not change much in this more complex model, the weights assigned to the kernels constructed for each feature set were similar, suggesting that no single feature set was more important than the others in the overall model.

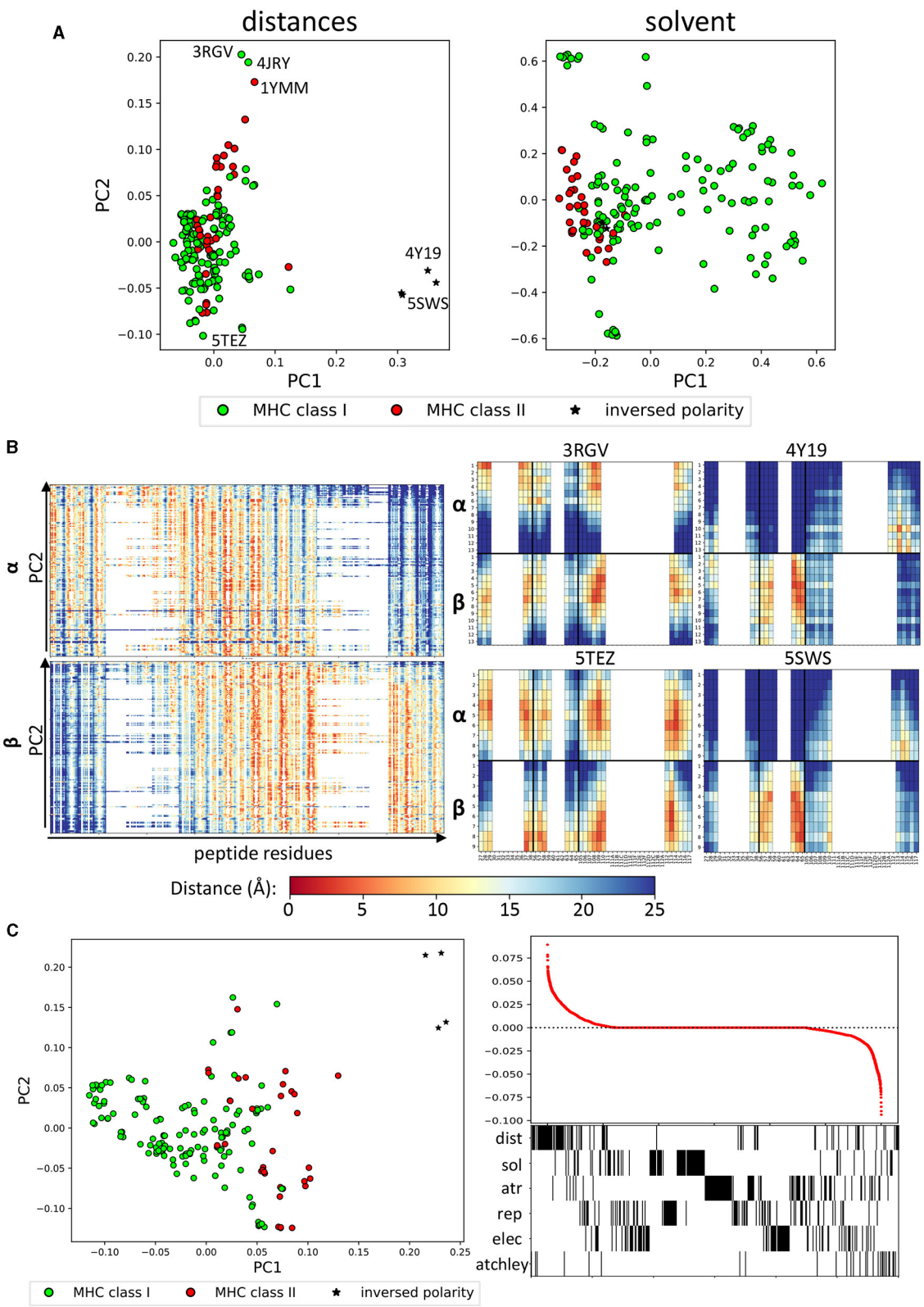


FIGURE 2 | Structural features identify different binding modes. **(A)** PCA performed on distances and on solvent energies can separate class I and class II complexes (green and red, respectively). The stars indicate the structures that have been reported to have inversed polarity (i.e., the TCRs bind the pMHC complex at 180° angle).

(Continued)

FIGURE 2 | Annotated on the distance plot, the structures at the extremes that we analyse in **(B)**. **(B)** (Left) Linearised vectors used for the distance PCA, ordered according to their PC2 score. On the x-axis, the minimum distance between each CDR residue and each peptide residue (27-1, 28-1,…,116-1, 117-1, 27-2,…,117-20). (Right) Fingerprints for four representative structures labelled in **(A)** (3RGV high PC2, 5TEZ low PC2, 5SWS, and 4Y19 high PC1). **(C)** (Left) PCA of all feature sets combined, which also shows separation along PC1. (Right) Loading coefficient of each feature on PC1 and below a barcode to show which set the feature belongs to.

We then went on to validate the trained model on the other five datasets described in the methods (**Figure 3**, steps 5 and 6). Because we wanted to test how generalisable the rules that the classifier had learnt were, we did not train the classifier again on the new sets, but used the model trained on the STCRDab set to predict the new complexes (**Figure 3**). The sets were used as retrieved, by removing duplicate sequences but without removing sequences that are similar to the ones in the training set, reflecting the way a typical user might use this tool. Results from validation are presented in **Figure 6D** and **Supplementary Figure 5** and summarised in **Table 2**. Overall, the models with the highest ROC AUC consistently included sequence information. Moreover, addition of structural features often did not improve predictive power. However, structural features often allowed some level of discrimination, independently of the sequence information, suggesting that the model might be learning something about the binding modes of these complexes. Interestingly, the models which used structural features had consistently higher recall, as measured for the optimum SVC hyperplane giving the highest AUC.

The newVdj set is interesting as it is the most unbalanced set, with $<$1% of total complexes modeling real binders. This is the closest situation to a real-life application, in which the classifier would be asked to pick out a very small number of positives in a large pool of negatives. Here, 3/5 models increase the proportion of positives in the set identified as binders, although only slightly (from 0.7% to 0.9%, 1.3% and 1.4%, in the models using distances, distances and atr or Atchley factors, **Table 2**).

The ATLAS proved to be a very hard set to predict overall. This might be due to each complex being only a few mutations away from the crystal structure deposited in the PDB, which might have on one hand made the modelling easier, but on the other hand made it harder for the classifier to tell the difference between a binding and a non-binding pair which differ at only one amino acid. Moreover, some of the included mutations occur at the MHC, which is not considered when extracting features. Finally, the ATLAS set does not have a strict definition of binding, as for the other sets which derive from tetramer-sorting experiments, but rather the complexes show a range of affinities, and it is hard to define a strict threshold to define binding.

3.4. Classifier Performance Varies Between Epitopes

A known hard task for a classifier trained on a small subset of the epitopes that our immune system is exposed to, is to generalise to epitopes not present in the training set. It is apparent from the diagrams showing mis-classification in **Figure 6D** (right) and **Supplementary Figure 5B** that some peptides were indeed easier to classify than others.

Figure 7A shows the classifier performance on 4 representative epitopes when features from the Atchley factor, distances and attractive van der Waals sets are combined. For a perfect classifier, the decision score for positive and negative samples (equivalent to the distance of a point from the decision hyperplane in the case of an SVM) should have non-overlapping distributions. However, for peptide antigen AVFDRKSDAK the distributions for binding and non-binding TCRs almost completely overlap, suggesting that the classifier has not learnt useful information from the data. For peptide LLFGYPVYV, on the other hand, the separation between the two groups of TCRs is almost perfect. The classification of TCRs specific for the ELAGIGILTV and ASNENMETM peptides showed an intermediate pattern. Overall, the classification of TCRs for different epitopes show very significant differences in performance (**Figure 7B**), as has been observed previously for other models (Moris et al., 2020). This also suggests that the overall performance as showed in **Table 2** is somewhat misleading, as it will be skewed by the more abundant epitopes.

3.5. Sequence Similarity to Train Set and Homology Modelling Template Availability Impact Classifier Performance

We wondered whether the difference in performance could be due to the performance of the homology modelling tool used. For each structure, we retrieved the information about the sequence similarity between the structure of interest and the best of the matched templates used by TCRpMHCmodels to predict the complex structure. We then plotted the classifier performance as a function of sequence similarity (**Figure 8A**).

Overall, there was a trend for better templates (increased sequence similarity) to correlate with better classifier performance (observed as an increase in performance to the right of the individual panels). The TCR homologies seem to have the largest impact on the performance, followed by peptide homology. The MHC template, on the other hand, was less strongly associated with performance. Interestingly, however, the trends were observed also when classification was based only on sequence information suggesting that this might not be related only to the accuracy of the homology modelling. The templates for the homology modelling and the training set for our classifier are overlapping sets (as both are using the complexes for which a crystal structure is available) and our results might be reflecting the increased density in the feature space of known complexes.

To further evaluate the effect of template selection on structure prediction and classification, we repeated the structure prediction step (**Figure 3**, step 2) for the STCRDab set, by

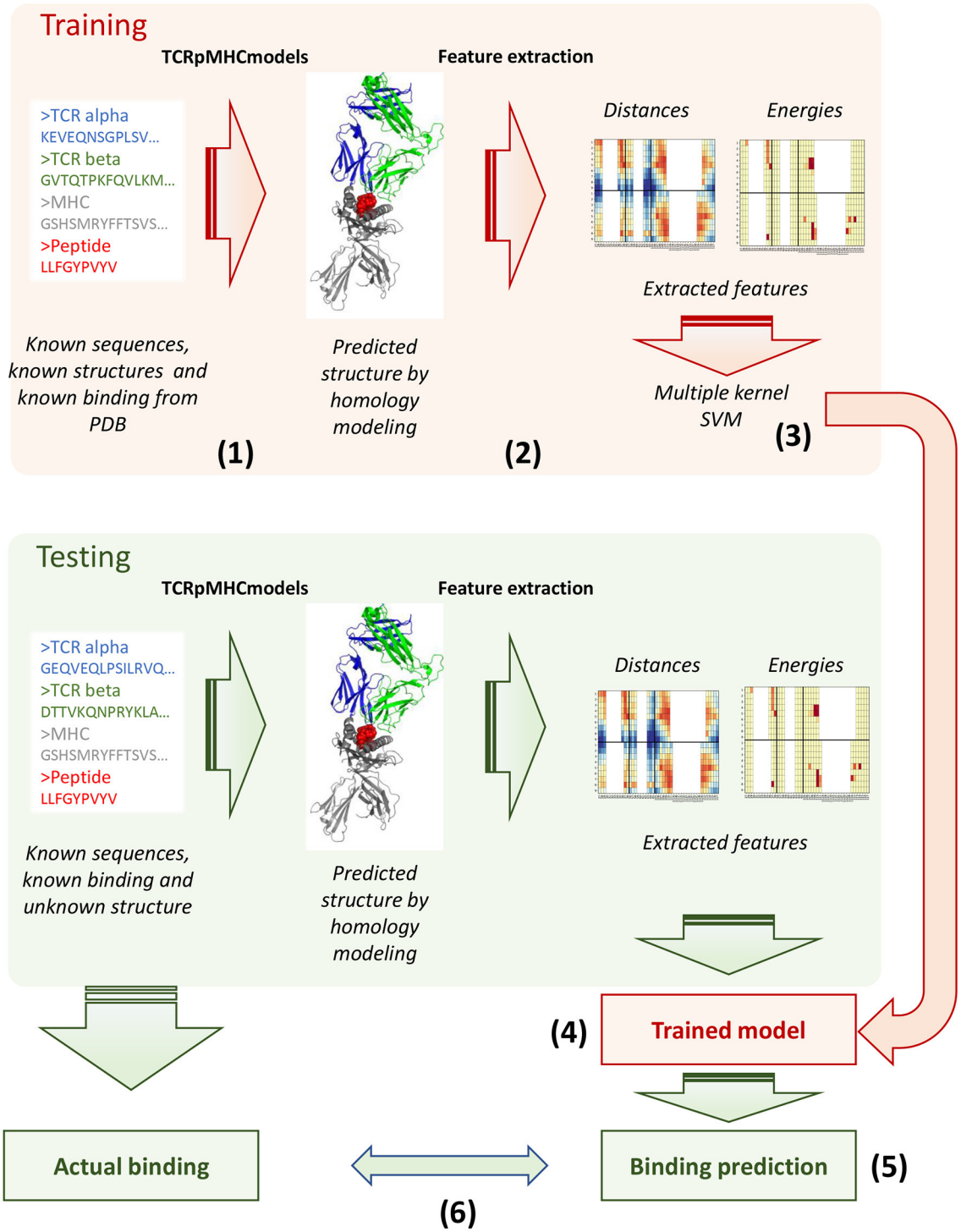


FIGURE 3 | Schematics of the modelling and prediction approach. For each sequence in the training set (the STCRDab set), a predicted structure was generated using TCRpMHCmodels (step 1) and features extracted using the pipeline described in the methods (step 2, the features extracted from the original and the predicted structures are compared to the original crystal structures in **Figure 4** and **Supplementary Figure 3**). The feature extracted were combined in multiple kernels (step 3) and multiple predictive models were generated which could distinguish binding and non-binding pairs (step 4). The model performance upon cross-validation is evaluated in **Figure 6**. The same structure prediction and feature extraction pipeline was applied to multiple test sets, for which binding is known but structure is not. The model generated in step 4 was used to make a prediction of binding for these new sequences (step 5), and the results were compared against the actual binding information to get an estimate of performance (step 6).

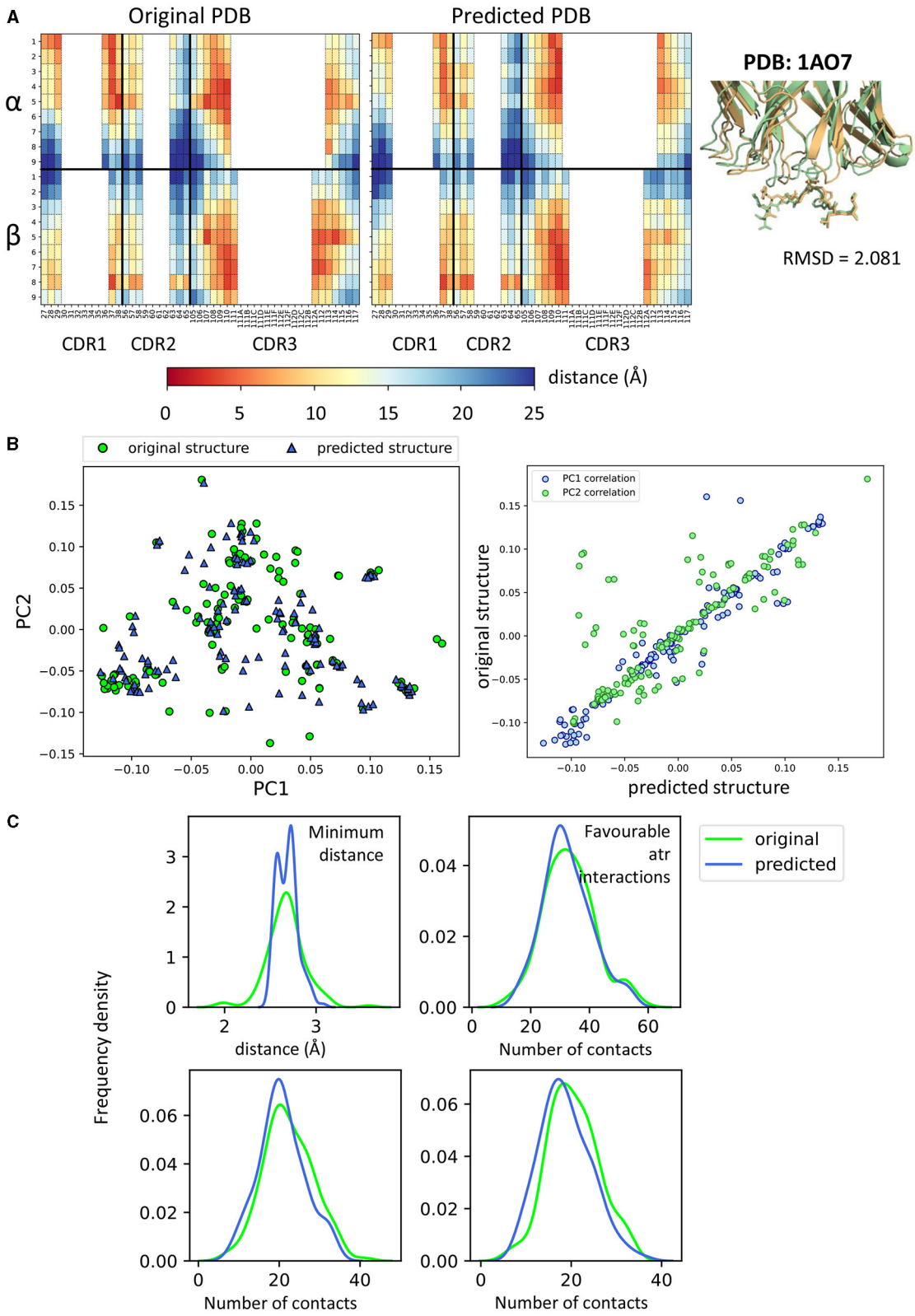


FIGURE 4 | Comparisons between crystal structures and homology predicted structures. **(A)** Comparison of fingerprint between the original 1AO7 structure and the one predicted by TCRpMHCmodels. On the right, figure showing how the two structures superimpose in cartoon form (green = original, gold = predicted). MHC not

(Continued)

FIGURE 4 | shown for clarity. **(B)** (Left) PCA on all feature sets showing the difference between crystal structures (green circles) and predicted structures (blue triangles). (Right) Correlation for PC1 and PC2 values between original and predicted structures. Each blue dot is a complex and has (x,y) coordinates that depend on PC1 values for predicted and original structure. Similarly for PC2 (green dots). PCA for other feature sets in **Supplementary Figure 3**. **(C)** Frequency distributions of four characteristics of the TCR-pMHC complexes comparing the distribution between original and predicted structures. Minimum distance: minimum distance between TCR and peptide; Contacts: number of TCR-peptide residue pairs that are <5A apart; Favourable atr/elec interactions: number of favourable (energy < 0) interactions between TCR and peptide.

setting a maximum threshold for the sequence similarity of each (TCR, pMHC, or Complex) or all templates used for the structure prediction. We then extracted features from these new structures, which we expect to be less accurately predicted, and performed 10-fold CV as in **Figure 6A**. Representative results for a single threshold (40) are shown in **Supplementary Figure 6A** and results for all thresholds are shown in **Figure 8B**. In all models but the one in which only Atchley factors are used, performance decreases when a lower threshold is set for template selection, and this is particularly evident when limiting the sequence similarity of the complex template or all templates. The pMHC template, on the other hand, seem to have little impact on performance in this setting. Note that a few TCR-pMHC pairs could not be modelled when setting a lower threshold. Details of the number of complexes represented by each point can be found in **Supplementary Table 1**.

To investigate the role of sequence similarity to the training set, we also computed the BLOSUM scores between the training set and each of the complexes we predicted (**Figure 8C**). This is a metric of similarity between the validation and the train sets. Indeed, a decrease in classifier performance is observed when the BLOSUM score decreases, i.e., when the TCR-pMHC pair that we are trying to predict is less similar to the training set pairs, it becomes harder to predict it correctly. Interestingly, in all cases the performance of the classifier is more dependent on TCR homology, than on peptide homology.

Together, these results suggest that both structure prediction accuracy and similarity to the training set impact performance of the structure classifiers.

It is important to note that the observed relationship between classifier performance and sequence homology allow us to predict *a priori* which TCR/peptide binding predictions will carry greater confidence. In fact, by considering the epitope and complex homology templates, we are able to select *a priori* a subset of structures on which our model will perform better (**Supplementary Figure 6B**).

3.6. Effect of Affinity on the Predictor

Because the classifier relies on structural information and it is trained on the set of TCR-pMHC pairs that have a known crystal structure, we wondered whether the model could predict binding affinity as well as a binary binding/non-binding classification or whether higher decision function scores were assigned to higher-affinity complexes (i.e., whether complexes which bind with high affinity are called binders with higher confidence). To address this, the TCR-pMHC pairs from the ATLAS (Borrman et al., 2017) were retrieved and their score predicted. The score for each complex was then correlated (Spearman) to their measured affinity, removing all complexes with undetectable binding and adjusting the ΔG and K_D as in the original publication (**Table 3**). Unexpectedly, the only significant correlation was between sequence features (Atchley factors) and k_{off}. The model therefore does not successfully capture the structural information which determines the affinity of the complex and its performance is not biased toward detection of high-affinity pairs.

3.7. Benchmarking Against Existing Tools

Finally, we compared the performance of our classifier against the recently published ERGO (Springer et al., 2020) and ImRex (Moris et al., 2020) (**Supplementary Table 2**). Both ERGO and ImRex were trained on the VDJdb set (Bagaev et al., 2020), as described in the original publication, rather than the much smaller set of sequences from crystal structures used by our algorithm. The trained models are available as an online tool for ERGO (https://tcr.cs.biu.ac.il/) and on GitHub for ImRex (https://github.com/pmoris/ImRex).

The classifiers were all tested on the same set of binding and non-binding TCR-pMHC sets. **Figure 9** and **Supplementary Table 2** show the results divided by peptide. The results are organised in three scenarios depending on whether the peptide is present in neither, either, or both of the train sets.

When compared on epitopes that are not present in either train set (Case 1), none of the models, whether sequence, structure or hybrid gave significant discrimination. When the epitopes are present in the VDJDb but not in the STCRDab (PDB) set (Case 2), both ERGO models significantly outperform all other models in prediction, including ImRex. Finally, when peptides are present in both train sets (Case 3), ERGO outperforms all models except the ones which include Atchley factors information.

Taken together, these results suggest that the structure-based models developed in this study perform as well as the state-of-the-art sequence-based models in predicting binding to novel pMHC, despite learning from a much smaller training set.

4. DISCUSSION

Prediction of TCR-pMHC binding is a fundamental challenge. With TCR repertoire sequencing (TCR-seq) gaining popularity, a large body of paired TCR-peptide sequences has become available in the recent years and many have attempted to leverage this vast amount of sequence information to predict TCR-antigen binding (Dash et al., 2017; Glanville et al., 2017; Huang et al., 2020; Moris et al., 2020; Springer et al., 2020; Tong et al., 2020; Sidhom et al., 2021). However, TCR-pMHC binding is at the core a biophysical problem, as it depends on the 3D structure of the two molecules and the interactions that can be formed between the two.

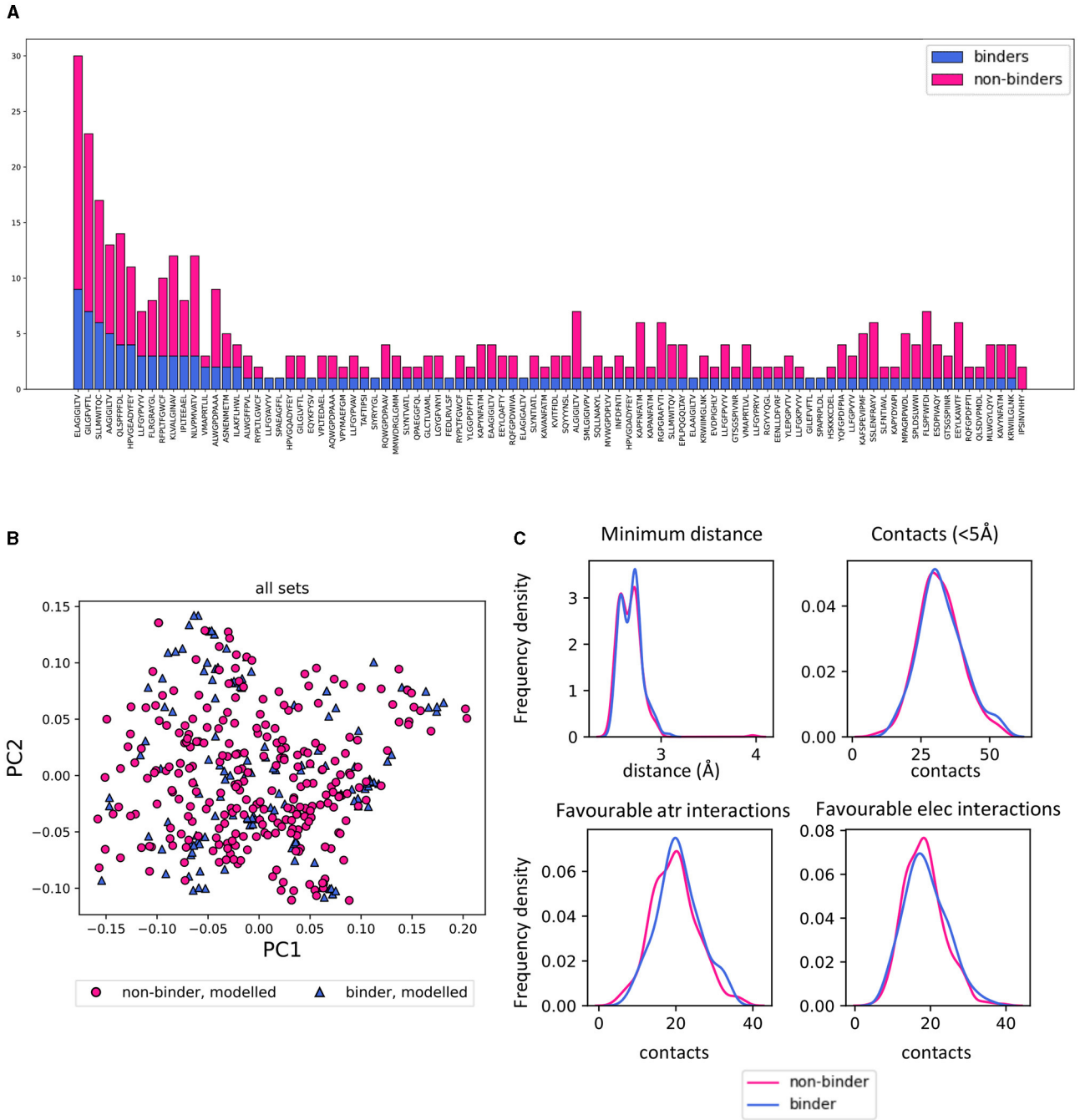
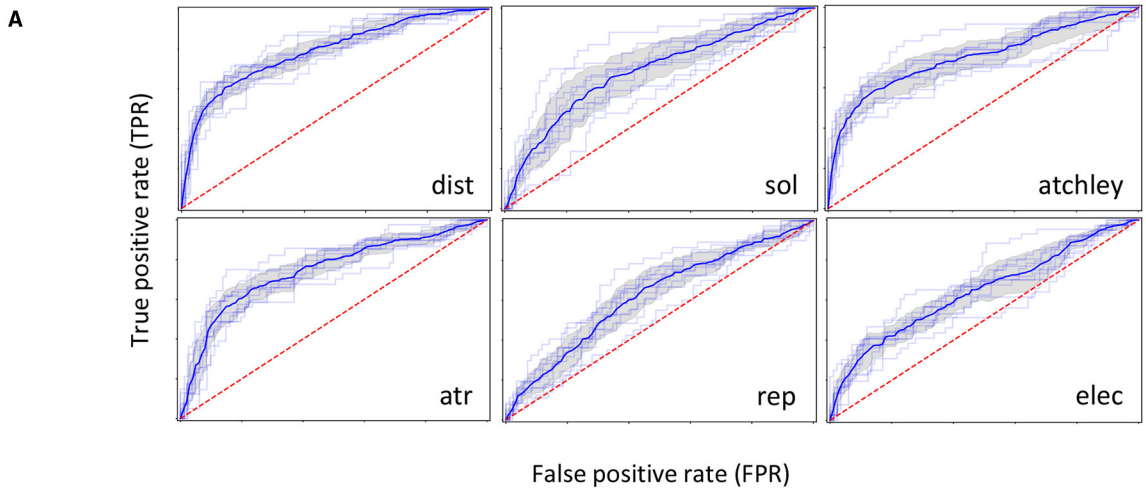
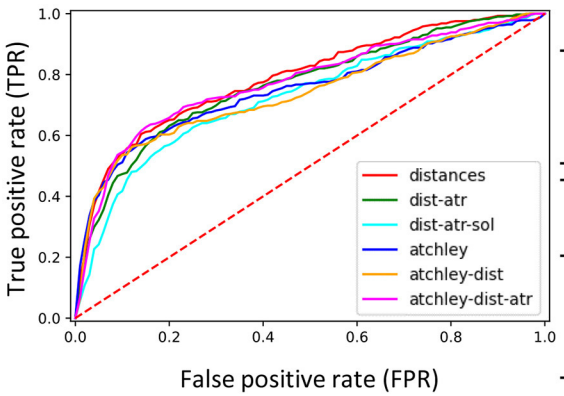


FIGURE 5 | Homology modelled binding and non-binding TCR-pMHC complexes can not be discriminated by PCA. **(A)** Summary of the number of STCRDab derived binding and non-binding structures which were modelled. For each peptide in the set, the barplot shows the number of models of binding and non-binding TCRs (blue and magenta, respectively). **(B)** PCA of all sets combined showing no separation between binding and non-binding TCR/pMHC homology models. The PCAs for each feature set separately are in **Supplementary Figure 4**. **(C)** Frequency distributions of 4 characteristics of the TCR-pMHC complexes comparing the distribution between binding and non-binding models. Minimum distance: minimum distance between TCR and peptide; Contacts: number of TCR-peptide residue pairs that are <5Å apart; Favourable atr/elec interactions: number of favourable (energy < 0) interactions between TCR and peptide.

The use of structure-derived features, energies in particular, for TCR-pMHC binding prediction had been explored before. For instance, Ogishi and Yotsuyanagi (2019) successfully modelled an energy-based sequence alignment score that can take a CDR and a peptide sequence, align them and calculate how likely they are to interact. Although sequence-based, this approach was trying to take into account biological interactions in the sequence encoding. Similarly, Lin et al. (2021) showed





C

ROC (10 folds)	dist	sol	atchley	atr	rep	elec
mean	0.791	0.686	0.752	0.746	0.622	0.642
median	0.795	0.687	0.771	0.740	0.627	0.645

ROC (10 folds)	dist-atr	dist-atr-sol	atchley-dist	atchley-dist-atr
mean	0.767	0.727	0.741	0.777
median	0.772	0.743	0.742	0.795

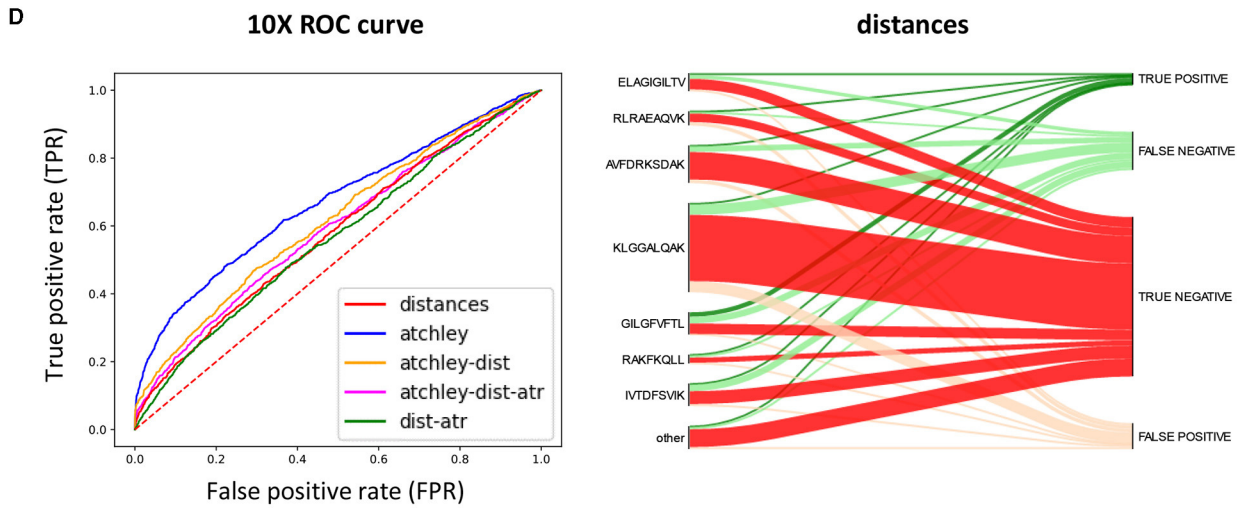


FIGURE 6 | A discriminative classification model can be trained using extracted structural features. **(A)** ROC AUC curves of 10-fold CV on the STCRDab training set with each feature set separately. The faint lines are the results for each individual fold, whilst the dark lines represent the interpolated average results, with the shaded area as the standard deviation. **(B)** Interpolated ROC AUC curves for 10-fold CV obtained when combining different feature sets for prediction. **(C)** Tabular results for

(Continued)

FIGURE 6 | curves showed in **(A,B)**. **(D)** (Left) ROC curves obtained when the model trained on the STCRDab set is used for prediction on the 10XGenomics validation set. (Right) For the model trained on STCRDab using the distance dataset only, the diagram shows which proportion of examples from each epitope are classified correctly (true positives and true negatives) or incorrectly (false positives and false negatives). This is shown in the form of a Sankey diagram, where from each epitope annotated on the left-hand side of the plot, a line is drawn for each TCR recognising that specific epitope to the final result for the complex (correctly classified as binder or non-binder, or incorrectly classified). The width of each section is proportional to the number of complexes that follow that classification.

TABLE 2 | Results of out-of-sample validation.

Set	% pos	Combo	Roc	Avg precision	Accuracy	Precision	Recall
10X	26.90%	distances	0.581	0.295	0.727	0.307	**0.230**
		Dist-atr	0.566	0.265	0.739	0.315	0.197
		Atchley	**0.669**	**0.443**	**0.807**	**0.742**	0.132
		Atchley-dist	0.616	0.359	0.782	0.459	0.163
		Atchley-dist-atr	0.592	0.322	0.773	0.406	0.159
Dash	7.33%	distances	0.605	0.108	0.740	0.111	**0.362**
		Dist-atr	0.650	0.124	0.802	0.139	0.326
		Atchley	**0.690**	**0.183**	**0.910**	**0.237**	0.104
		Atchley-dist	0.611	0.180	0.799	0.133	0.316
		Atchley-dist-atr	0.648	0.147	0.824	0.154	0.311
Expt	12.70%	distances	0.733	0.332	0.730	0.275	**0.688**
		Dist-atr	0.707	0.454	0.714	0.250	0.625
		Atchley	**0.809**	**0.698**	**0.786**	**0.333**	**0.688**
		Atchley-dist	0.807	0.667	0.738	0.270	0.625
		Atchley-dist-atr	0.768	0.532	0.722	0.256	0.625
Atlas	86.60%	distances	0.463	0.852	0.840	0.866	0.964
		Dist-atr	0.504	0.863	0.777	0.864	0.881
		Atchley	**0.570**	**0.897**	**0.866**	0.866	**1.000**
		Atchley-dist	0.471	0.867	0.863	**0.867**	0.993
		Atchley-dist-atr	0.497	0.869	0.834	0.866	0.957
NewVdj	0.70%	distances	0.528	**0.010**	0.832	0.009	**0.205**
		Dist-atr	0.535	0.009	0.908	0.013	0.159
		Atchley	0.516	0.008	**0.981**	**0.014**	0.023
		Atchley-dist	0.549	**0.010**	0.953	0.004	0.023
		Atchley-dist-atr	**0.559**	0.009	0.953	0.000	0.000

Results of predicting the validation sets with the model trained on the STCRDab set, using different subsets of features. In each section, the best-performing model is highlighted in bold and underlined. The precision and recall are measured for the optimum SVC hyperplane giving the highest AUC.

that it is possible to optimise a sequence-based energy model starting from the interactions observed in the existing crystal structure for a TCR-pMHC complex. This function can separate strong binders from weak binders to then predict whether an epitope would bind a TCR of interest in an MHC-restricted way. Lanzarotti et al. (2018) had also previously shown that some combination of energy terms derived from structural models could be used to predict TCR-antigen binding and Borrman et al. (2020) successfully ranked candidate peptide epitopes from a phage screen against target TCRs using predicted binding energies by modelling mutations within the existing crystal structures for the TCR-pMHC complexes.

Here, we too have approached the TCR-pMHC binding question from a structural perspective and wondered whether structural information could be integrated with sequence information to improve the prediction. We have here developed a method to extract features to try and recapitulate both the conformation and the energetic profile of the TCR-pMHC binding interface and integrate it with existing sequence information.

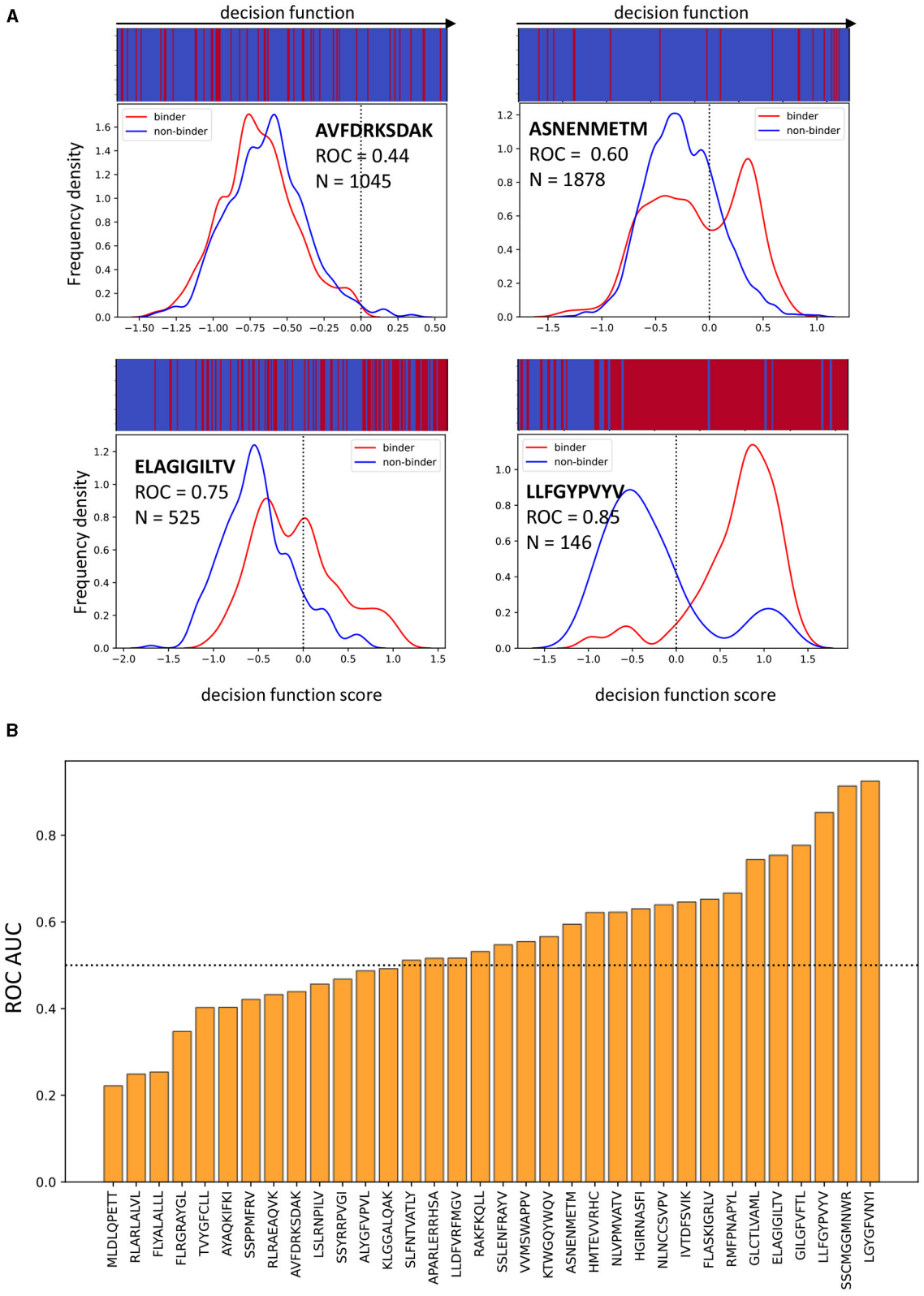


FIGURE 7 | The performance of the model is pMHC dependent. **(A)** Examples of four different epitopes. The frequency distributions of model decision function scores (for an SVM, this corresponds to the distance from the separating hyperplane, drawn as a dotted line) for binding and non-binding TCRs recognising each epitope

(Continued)

FIGURE 7 | when the model uses distance, attractive van der Waals and Atchley factors. The bar at the top shows the order in which binding and non-binding examples appear when ranked by decision function. For good classification, the bar should be mostly blue on the left and mostly red on the right. **(B)** The bar plot shows ROC AUC for all peptides which have at least 2 positive and 2 negative examples. This data comes from concatenating the predictions for all the validation sets when Atchley factors, distances and attractive van der Waals forces are used.

First, we validated the ability of our structural features to recapitulate the known biology and characteristics of the system. In a survey of the crystal structures available we showed that, in agreement with previous reports (Glanville et al., 2017; Ostmeyer et al., 2019), we can detect stretches of amino acids at the centre of the CDR3 in the TCRα and β chains that are within contact distance of the peptide. This information was also recapitulated by the energy profiles, suggesting that not only can these residues interact, but that they make favourable interactions.

Binding geometry is a known important factor in TCR-pMHC binding, which may be restricted by the way the peptide is presented by the MHC (Blevins et al., 2016). Recently, Singh et al. (2020) have shown that a difference can be detected between pMHC class I and class II binding geometry. The features we extracted also differentiate class I and class II complexes, both at the conformational level (in terms of pairwise distances) and at the energetic level. As reported by Singh et al. (2020), our analysis also showed that TCRs binding MHC class I tend to be closer to the C-terminus of the peptide, whilst TCRs binding class II complexes sit more centrally or toward the N-terminus. Moreover, our energetic features suggest that the interactions of class I and class II complexes also differ on an energetic level. As well as the difference between class I and class II, the spatial features extracted from the structures were readily able to distinguish TCRs which bind with reversed polarity to the pMHC complex, as described by Gras et al. (2016) and Beringer et al. (2015), and identify class I complexes with different, non-canonical binding modes to the peptide (Yin et al., 2011; Liu et al., 2013). This suggests that the features extracted are informative of the biology of this system.

The information collected from these structures was used to build a classifier able to discriminate between TCR-pMHC binding from non-binding pairs. Here, we see from the weights assigned to each combined kernel in the model including all features (atchley-dist-atr) that the physical interactions encoded by the distances and the attractive van der Waals forces were equally as important to the prediction as the sequence information, suggesting that physical interactions can be used to predict binding.

The generalisability of the classifier was tested on multiple datasets, collected and analysed independently. None of the available models, whether based on sequence, structure or a hybrid could successfully discriminate sets of complexes which had no overlap at all with the training set. However, for the other sets physical interaction features on their own proved sufficient to observe some discrimination between binding and non-binding complexes (**Figure 9**). Interestingly, integration of sequence and physical features in the same model did not improve the performance in terms of ROC AUC, although often improved the recall of the sequence-based model. This is an important characteristic, as in real-life applications a classifier like the one presented could be used to screen candidate TCRs against an epitope of interest, for example with the aim of identifying tumour-infiltrating lymphocytes that can recognise tumour neoantigens. In this context, *in-silico* screening would be followed by experimental validation. Because the events of interest are a very small number compared to the total number of events (i.e., binders << non-binders), it would be more important to correctly classify more of the binders than of the non-binders, i.e., a higher number of false positives, which can be screened out during experimental validation, would be less problematic than a higher number of false negatives, which would not be experimentally validated. The validity of using the method proposed to enrich for binders is explored in the newVdj set, in which <1% of all pairs are true binders. Whilst the enrichment is modest, it is a stepping stone in the right direction and shows promise for this kind of approach.

Most of the results presented have been based on a binary classification of TCR-pMHC complexes as binding or non-binding. In reality, the interaction between TCR and pMHC is characterised by a graded affinity scale. We showed that performance of the classifier is not impacted by affinity of the TCR for the pMHC (**Table 3**). However, affinity prediction would be of interest as there are multiple metrics that contribute to overall affinity and are known to be important for T cell activation dynamics—K_D, k_{on}, k_{off}, half-life—(Stone et al., 2009; Lever et al., 2016; Gálvez et al., 2019) and we are not yet able to manipulate them systematically by acting on the TCR. The original TCR ATLAS publication (Borrman et al., 2017) showed a correlation between the attractive van der Waal force as calculated by Rosetta (here atr) and the experimentally-measured affinity, similar to the one reported by Erijman et al. (2014) on an unrelated system. Because the affinity is driven by structure, we believe the PDB classifier might also be optimised for approximate affinity prediction, although better methods of modelling the mutations into the structures might have to be explored. However, this is known to be a very hard problem, which is only starting to become tractable on simpler systems (see, for instance, Leidner et al., 2019; Abbasi et al., 2020; Jiang et al., 2020).

Compared to other published classifiers (Dash et al., 2017; Glanville et al., 2017; Tong et al., 2020), the classifier presented here is different in that it does not need to be trained on a known subset of TCRs recognising a specific peptide to be able to predict more binders, but rather it can learn from any set of TCR-pMHC pairs already available and generalise what it has learnt to the problem at hand. This suggests that there are conserved features to the TCR-pMHC interface which can be learnt and used for prediction. Tools such as ERGO (Springer et al., 2020), ImRex (Moris et al., 2020), TcellMatch (Fischer et al., 2020), and NetTCR (Jurtz et al., 2018) have pioneered this approach,

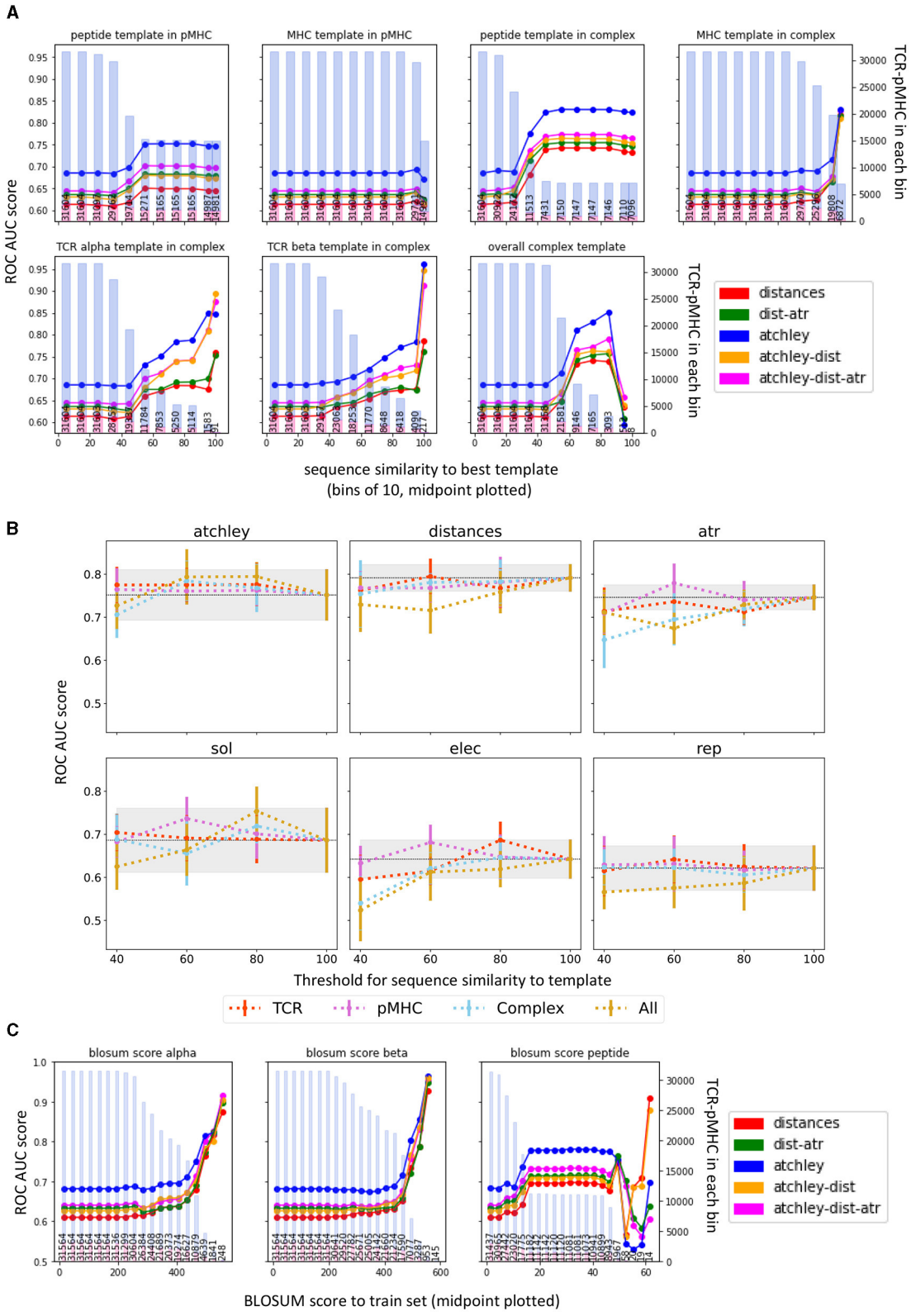


FIGURE 8 | Classifier performance is dependent on sequence homology of the target TCR-pMHC. **(A)** The performance from all validation sets were combined, and stratified by the similarity between the sequence of the target complex to be classified and the relevant homology modelling template (as outputted by

(Continued)

FIGURE 8 | TCRpMHCmodes and outlined in Jensen et al., 2019). Mean performance (ROC AUC) in each range of homology is calculated and plotted at the range midpoint. The bars show the number of structures that contribute to the performance for each point, the proportion in blue showing the number of non-binders and the proportion in red showing the number of binders. **(B)** Results of 10-fold CV when different thresholds are used for structure prediction. In each panel, the coloured lines show the average ROC AUC over 10-fold CV when setting the threshold on the TCR, pMHC, or Complex templates, or on all at the same time (All). The error bars show the standard deviation of the average ROC AUC computed over the 10-folds. The grey areas show 1 standard deviation from the performance of the model when no threshold is set (100). **(C)** Equivalent analysis to **(A)** but calculating the BLOSUM score between each example and the closest example in the train set, for each chain separately. The higher the BLOSUM score, the more similar the sequence is to one found in the training set. In each plot, the bars show the number of structures in each bin, the proportion in blue showing the number of non-binders and the proportion in red showing the number of binders.

TABLE 3 | Correlations of affinity metrics and decision function scores.

	Distances		Dist-atr		Atchley		Atchley-dist		Atchley-dist-atr	
	Spearman R	***p*-value**	**Spearman R**	***p*-value**	**Spearman R**	***p*-value**	**Spearman R**	***p*-value**	**Spearman R**	***p*-value**
K_D (μM)	−0.062	0.278	−0.075	0.195	0.006	0.911	−0.069	0.231	−0.060	0.297
k_{on} (Ms^{-1})	0.143	0.126	0.136	0.147	0.074	0.428	0.162	0.083	0.151	0.107
k_{off} (s^{-1})	0.080	0.397	−0.102	0.276	0.311	0.001	0.087	0.358	0.019	0.839
ΔG (kcal/mol)	−0.067	0.243	−0.083	0.151	−0.008	0.886	−0.074	0.198	−0.072	0.215

Spearman correlation is calculated for each affinity metric for predictions made for each of the models trained.

although they only focused on information that can be extracted from the sequence. Interestingly, Fischer et al. (2020) showed that prediction of antigens as categorical classes (where a TCR is assigned to the most likely class), is an easier problem than encoding of TCR and antigen simultaneously to predict binding, although the latter is of more practical use when new epitopes are of interest. Of note, all of the results that we have presented here use the model originally trained on the STCRDab set, which was never re-trained on a new sets of structures. Moreover, the classifier here presented is trained on about 400 binding and non-binding pairs, which recognise 93 different epitopes. This is a much smaller set than the VDJdb used by ERGO and ImRex (40,000 TCRs and 200 peptides in ERGO and 14,000 CDR3β and 118 peptides in ImRex), but achieves similar performances. This might indicate that the information learnt from the structural information is more readily generalised to an unseen case.

We have investigated the potential of structural data to predict TCR-pMHC binding, with the hope that this study can act as a stepping stone for further work which can improve the accuracy of a structure based approach. In particular, we have identified a number of limitations to this approach and highlighted avenues which show promise.

Firstly, throughout the analysis, interactions between the TCR and MHC were disregarded. Whilst this simplified the feature extraction process, it is an over-simplification of the problem, as the MHC is an active determinant of TCR specificity (Piepenbrink et al., 2013; Wang et al., 2017). For instance, Blevins et al. (2016) showed that a TCR needs to be able to complement a hot-spot region of positively-charged residues to bind peptide-HLA-A2 complex, effectively constraining binding. Addition of the MHC in the models presented might therefore significantly boost performance.

Secondly, the very large number of potential combinations of TCR-pMHC complexes that could be formed makes homology modelling a very attractive approach. However, accuracy of the modelling is severely limited by the number of available crystal structures that can be used as templates. The most interesting regions for TCR-pMHC prediction, the CDR loops, are the most variable and hardest to model, and have RMSD between predicted and solved structures which range between 0.5 and 5Å (Jensen et al., 2019), making the prediction most unreliable where the key information is encoded. In addition, because the prediction relies on templates, homology modelling may be fitting the predicted structures to look more like their templates than they should, thus reducing the variability between complexes. To be noted that TCRpMHCmodels was unable to predict all the TCR-pMHC in the set, which might have further skewed the validation sets to be more similar in sequence to the train set.

Accurately estimating the impact of homology modelling accuracy in our model has been challenging. We have here relied on the published set of crystal structures both for homology modelling and for training of the classifier. This overlap makes it hard to disentangle the effects of accuracy homology modelling from the effects of similarity to the training set (**Figure 8**). Whilst setting a threshold of sequence similarity for structure prediction has shed some light on the impact of homology modelling accuracy, use of a different training set with non-overlapping sequences with the homology modelling templates and for which estimation of homology modelling performance can be estimated might help to further clarify this point. Moreover, the set of TCR-pMHC available from the PDB is currently heavily skewed toward a few epitopes that have been studied in great detail, only a small fraction of all the possible TCR-pMHC combinations. We believe that this might be the reason why cross-validation on the classifier performed significantly better than out-of-sample validation (**Figure 6** and **Supplementary Figure 5**), as structures in the training set will be more similar to each other and therefore easier to classify.

As more structures for more diverse epitopes and TCRs become available, we expect that both homology modelling accuracy and performance of the classifier will improve.

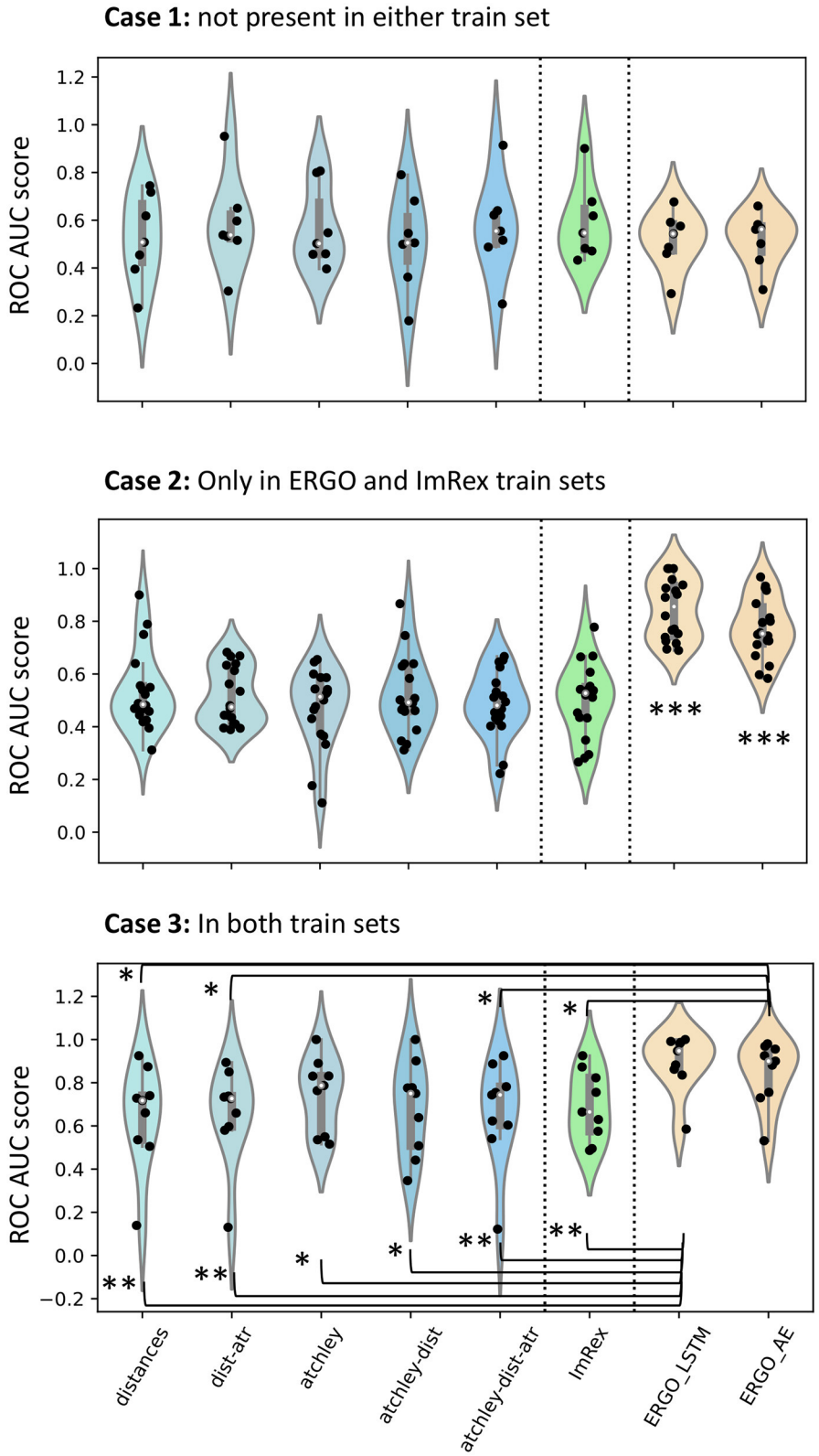


FIGURE 9 | Comparison of performance with other published tools. In each violin plot, a dot is an epitope for which performance is calculated. In Case 1, only epitopes that are not present in the PDB or in the VDJDb train sets are included. In Case 2, only epitopes that are present in the VDJDb but not in the PDB are included. In Case 3, only epitopes which are in both training sets are included. To look for differences, a Mann-Whitney *U*-test was calculated and significance values are shown: $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

Structure-prediction is now a fast-moving field and we might soon be able to apply newer and improved methods to the TCR-pMHC prediction problem, able to account for the highly diverse CDR loops (Teraguchi et al., 2020). Compared to other tools, TCRpMHCmodels is attractive in its ability to produce a quick prediction of the entire TCR-pMHC complex, without relying on docking methods, which still need to be optimised for this task (Peacock and Chain, 2021). Another interesting tool for this is ImmuneScape (Li et al., 2019), which also separately models the TCR and the pMHC, and combines them using a docking template. However, the complex biology of the system might always be a limiting factor for structure and binding prediction, from structural and/or sequence features. For example, if a small proportion of TCRs bound to the pMHC complex with conformations or sequence interactions that are significantly different from canonical binding, we might never be able to predict their binding with a tool that has learnt on a subset of canonical TCRs. This may well be the case with other structures with reversed polarity or complexes with unusual binding highlighted in **Figure 2A**.

Finally, the major difference between this classifier and most of the work published so far is that it relies on an available TCR$\alpha\beta$ pairs and cannot be used on unpaired chains. This is a limitation to the direct application of the classifier as $\alpha\beta$ pairing is typically not available from bulk TCRseq data. However, unpaired α and β chains only contain a portion of the binding site information, and the assumption that binding of the β chain only is sufficient is clearly not true in every case. Carter et al. (2019) show that the information encoded in the $\alpha\beta$ pair is synergistic, i.e., that the pairing carries more than the sum of the individual chain information. Their survey of the VDJdb also shows instances where the same α chain paired with different β chains recognise different epitopes, or where CDR3α and β annotated to bind epitopes from different species come together to bind yet another peptide. Lanzarotti et al. (2019) and Fischer et al. (2020) also showed that their prediction performance increases when information about both the α and the β TCR chains is included. Overall, we believe this to be strong motivation to work on $\alpha\beta$ pairs. Future work will focus on understanding whether candidate $\alpha\beta$ pairs that bind a specific antigen can be inferred from TCR clones that are expanded during an immune response.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

MM and BC designed the project and interpreted the data. MM performed the computational analysis. JS-T assisted with the computational models and methodology. MM wrote the manuscript with contributions from BC. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by Cancer Research UK through a studentship to MM and a grant to BC (C9200/A24314). BC was supported by The Rosetrees Foundation and The National Institute for Health Research UCL Hospitals Biomedical Research Centre.

ACKNOWLEDGMENTS

The content of this paper was previously published in a preprint (Milighetti et al., 2021).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphys.2021.730908/full#supplementary-material

REFERENCES

10XGenomics (2020). *A New Way of Exploring Immunity - Linking Highly Multiplexed Antigen Recognition to Immune Repertoire and Phenotype.* 10XGenomics.

Abbasi, W. A., Yaseen, A., Hassan, F. U., Andleeb, S., and Minhas, F. U. A. A. (2020). ISLAND: in-silico proteins binding affinity prediction using sequence information. *BioData Min.* 13, 1–13. doi: 10.1186/s13040-020-00231-w

Aiolli, F., and Donini, M. (2015). EasyMKL: A scalable multiple kernel learning algorithm. *Neurocomputing* 169, 215–224. doi: 10.1016/j.neucom.2014.11.078

Alford, R. F., Leaver-Fay, A., Jeliazkov, J. R., O'Meara, M. J., DiMaio, F. P., Park, H., et al. (2017). The rosetta all-atom energy function for macromolecular modeling and design. *J. Chem. Theory Comput.* 13, 3031–3048. doi: 10.1021/acs.jctc.7b00125

Atchley, W. R., Zhao, J., Fernandes, A. D., and Druke, T. (2005). Solving the protein sequence metric problem. *Proc. Natl. Acad. Sci. U.S.A.* 102, 6395–6400. doi: 10.1073/pnas.0408677102

Bagaev, D. V., Vroomans, R. M., Samir, J., Stervbo, U., Rius, C., Dolton, G., et al. (2020). VDJdb in 2019: Database extension, new analysis infrastructure and a T-cell receptor motif compendium. *Nucleic Acids Res.* 48, D1057–D1062. doi: 10.1093/nar/gkz874

Beringer, D. X., Kleijwegt, F. S., Wiede, F., Van Der Slik, A. R., Loh, K. L., Petersen, J., et al. (2015). T cell receptor reversed polarity recognition of a self-antigen major histocompatibility complex. *Nat. Immunol.* 16, 1153–1161. doi: 10.1038/ni.3271

Blevins, S. J., Pierce, B. G., Singh, N. K., Riley, T. P., Wang, Y., Spear, T. T., et al. (2016). How structural adaptability exists alongside HLA-A2 bias in the human $\alpha\beta$ TCR repertoire. *Proc. Natl. Acad. Sci. U.S.A.* 113, E1276–E1285. doi: 10.1073/pnas.1522069113

Borrman, T., Cimons, J., Cosiano, M., Purcaro, M., Pierce, B. G., Baker, B. M., et al. (2017). ATLAS: A database linking binding affinities with structures for wild-type and mutant TCR-pMHC complexes. *Proteins* 85, 908–916. doi: 10.1002/prot.25260

Borrman, T., Pierce, B. G., Vreven, T., Baker, B. M., and Weng, Z. (2020). High-throughput modeling and scoring of TCR-pMHC complexes to predict cross-reactive peptides. *Bioinformatics* 36, 5377–5385. doi: 10.1093/bioinformatics/btaa1050

Britanova, O. V., Putintseva, E. V., Shugay, M., Merzlyak, E. M., Turchaninova, M. A., Staroverov, D. B., et al. (2014). Age-related decrease in TCR repertoire diversity measured with deep and normalized sequence profiling. *J. Immunol.* 192, 2689–2698. doi: 10.4049/jimmunol.1302064

Carter, J. A., Preall, J. B., Grigaityte, K., Goldfless, S. J., Jeffery, E., Briggs, A. W., et al. (2019). Single T Cell sequencing demonstrates the functional role of $\alpha\beta$ TCR pairing in cell lineage and antigen specificity. *Front. Immunol.* 10:1516. doi: 10.3389/fimmu.2019.01516

Chatterjee, B., Deng, Y., Holler, A., Nunez, N., Azzi, T., Vanoaica, L. D., et al. (2019). CD8+ T cells retain protective functions despite sustained inhibitory receptor expression during Epstein-Barr virus infection *in vivo*. *PLoS Pathog.* 15:e1007748. doi: 10.1371/journal.ppat.1007748

Chaudhury, S., Lyskov, S., and Gray, J. J. (2010). PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics* 26, 689–691. doi: 10.1093/bioinformatics/btq007

Cinelli, M., Sun, Y., Best, K., Heather, J. M., Reich-Zeliger, S., Shifrut, E., et al. (2017). Feature selection using a one dimensional naïve Bayes' classifier increases the accuracy of support vector machine classification of CDR3 repertoires. *Bioinformatics* 33:btw771. doi: 10.1093/bioinformatics/btw771

Coles, C. H., Mulvaney, R. M., Malla, S., Walker, A., Smith, K. J., Lloyd, A., et al. (2020). TCRs with distinct specificity profiles use different binding modes to engage an identical peptide–HLA complex. *J. Immunol.* 204, 1943–1953. doi: 10.4049/jimmunol.1900915

Dash, P., Fiore-Gartland, A. J., Hertz, T., Wang, G. C., Sharma, S., Souquette, A., et al. (2017). Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature* 547, 89–93. doi: 10.1038/nature22383

de Greef, P. C., Oakes, T., Gerritsen, B., Ismail, M., Heather, J. M., Hermsen, R., et al. (2020). The naive T-cell receptor repertoire has an extremely broad distribution of clone sizes. *eLife* 9:49900. doi: 10.7554/eLife.49900

Dunbar, J., and Deane, C. M. (2016). ANARCI: Antigen receptor numbering and receptor classification. *Bioinformatics* 32, 298–300. doi: 10.1093/bioinformatics/btv552

Erijman, A., Rosenthal, E., and Shifman, J. M. (2014). How structure defines affinity in protein-protein interactions. *PLoS ONE* 9:e110085. doi: 10.1371/journal.pone.0110085

Fischer, D. S., Wu, Y., Schubert, B., and Theis, F. J. (2020). Predicting antigen specificity of single T cells based on TCR CDR3 regions. *Mol. Syst. Biol.* 16:e9416. doi: 10.15252/msb.20199416

Fiser, A., and Šali, A. (2003). MODELLER: generation and refinement of homology-based protein structure models. *Methods Enzymol.* 374, 461–491. doi: 10.1016/S0076-6879(03)74020-8

Gálvez, J., Gálvez, J. J., and García-Pe narrubia, P. (2019). Is TCR/pMHC affinity a good estimate of the t-cell response? An answer based on predictions from 12 phenotypic models. *Front. Immunol.* 10:349. doi: 10.3389/fimmu.2019.00349

Garboczi, D. N., Ghosh, P., Utz, U., Fan, Q. R., Biddison, W. E., and Wiley, D. C. (1996). Structure of the complex between human T-cell receptor, viral peptide and HLA-A2. *Nature* 384, 134–141. doi: 10.1038/384134a0

Garcia, K. C., Adams, J. J., Feng, D., and Ely, L. K. (2009). The molecular basis of TCR germline bias for MHC is surprisingly simple. *Nat. Immunol.* 10, 143–147. doi: 10.1038/ni.f.219

Glanville, J., Huang, H., Nau, A., Hatton, O., Wagar, L. E., Rubelt, F., et al. (2017). Identifying specificity groups in the T cell receptor repertoire. *Nature* 547, 94–98. doi: 10.1038/nature22976

Gras, S., Chadderton, J., Del Campo, C. M., Farenc, C., Wiede, F., Josephs, T. M., et al. (2016). Reversed T cell receptor docking on a major histocompatibility class i complex limits involvement in the immune response. *Immunity* 45, 749–760. doi: 10.1016/j.immuni.2016.09.007

Hamelryck, T., and Manderick, B. (2003). PDB file parser and structure class implemented in Python. *Bioinformatics* 19, 2308–2310. doi: 10.1093/bioinformatics/btg299

Huang, H., Wang, C., Rubelt, F., Scriba, T. J., and Davis, M. M. (2020). Analyzing the Mycobacterium tuberculosis immune response by T-cell receptor clustering with GLIPH2 and genome-wide antigen screening. *Nat. Biotechnol.* 38, 1194–1202. doi: 10.1038/s41587-020-0505-4

Jensen, K. K., Rantos, V., Jappe, C., Olsen, H., Closter Jespersen, M., Jurtz, V., et al. (2019). TCRpMHCmodels: structural modelling of tcR-pMHc class i complexes. *Sci. Rep.* 9:14530. doi: 10.1038/s41598-019-50932-4

Jiang, M., Li, Z., Zhang, S., Wang, S., Wang, X., Yuan, Q., et al. (2020). Drug-target affinity prediction using graph neural network and contact maps. *RSC Adv.* 10, 20701–20712. doi: 10.1039/D0RA02297G

Joshi, K., Robert de Massy, M., Ismail, M., Reading, J. L., Uddin, I., Woolston, A., et al. (2019). Spatial heterogeneity of the T cell receptor repertoire reflects the mutational landscape in lung cancer. *Nat. Med.* 25, 1549–1559. doi: 10.1038/s41591-019-0592-2

Jurtz, V. I., Jessen, L. E., Bentzen, A. K., Jespersen, M. C., Mahajan, S., Vita, R., et al. (2018). NetTCR: sequence-based prediction of TCR binding to peptide-MHC complexes using convolutional neural networks. *bioRxiv* 433706. doi: 10.1101/433706

Kjer-Nielsen, L., Clements, C. S., Purcell, A. W., Brooks, A. G., Whisstock, J. C., Burrows, S. R., et al. (2003). A structural basis for the selection of Dominant $\alpha\beta$ T cell receptors in antiviral immunity. *Immunity* 18, 53–64. doi: 10.1016/S1074-7613(02)00513-7

Klausen, M. S., Anderson, M. V., Jespersen, M. C., Nielsen, M., and Marcatili, P. (2015). LYRA, a webserver for lymphocyte receptor structural modeling. *Nucl. Acids Res.* 43, 349–355. doi: 10.1093/nar/gkv535

Lanzarotti, E., Marcatili, P., and Nielsen, M. (2018). Identification of the cognate peptide-MHC target of T cell receptors using molecular modeling and force field scoring. *Mol. Immunol.* 94, 91–97. doi: 10.1016/j.molimm.2017.12.019

Lanzarotti, E., Marcatili, P., and Nielsen, M. (2019). T-Cell receptor cognate target prediction based on paired α and β chain sequence and structural CDR loop similarities. *Front. Immunol.* 10:2080. doi: 10.3389/fimmu.2019.02080

Lauriola, I., and Aiolli, F. (2020). MKLpy: a python-based framework for Multiple Kernel Learning. *arXiv*.

Lauriola, I., Donini, M., and Aiolli, F. (2017). "Learning dot product polynomials for multiclass problems," in *ESANN 2017 - Proceedings, 25th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 23–28 *arXiv*:2007.09982.

Leem, J., de Oliveira, S. H. P., Krawczyk, K., and Deane, C. M. (2018). STCRDab: the structural T-cell receptor database. *Nucl. Acids Res.* 46, D406–D412. doi: 10.1093/nar/gkx971

Lefranc, M.-P. (1997). Unique database numberings system for immunogenetic analysis. *Immunol. Today* 18:509. doi: 10.1016/S0167-5699(97)01163-8

Leidner, F., Kurt Yilmaz, N., and Schiffer, C. A. (2019). Target-specific prediction of ligand affinity with structure-based interaction fingerprints. *J. Chem. Inform. Model.* 59, 3679–3691. doi: 10.1021/acs.jcim.9b00457

Lever, M., Lim, H.-S., Kruger, P., Nguyen, J., Trendel, N., Abu-Shah, E., et al. (2016). Architecture of a minimal signaling pathway explains the T-cell response to a 1 million-fold variation in antigen affinity and dose. *Proc. Natl. Acad. Sci. U.S.A.* 113, E6630–E6638. doi: 10.1073/pnas.1608820113

Li, S., Wilamowski, J., Teraguchi, S., van Eerden, F. J., Rozewicki, J., Davila, A., et al. (2019). "Chapter 17: Structural modeling of lymphocyte receptors and their antigens," in *In Vitro Differentiation of T-Cells. Methods in Molecular Biology, Vol 2048*, ed S. Kaneko (New York, NY: Humana), 207–229. doi: 10.1007/978-1-4939-9728-2_17

Lin, X., George, J. T., Schafer, N. P., Chau, K. N., Birnbaum, M. E., Clementi, C., et al. (2021). Rapid assessment of T-cell receptor specificity of the immune repertoire. *Nat. Comput. Sci.* 2021, 362–373. doi: 10.1038/s43588-021-00076-1

Liu, Y. C., Miles, J. J., Neller, M. A., Gostick, E., Price, D. A., Purcell, A. W., et al. (2013). Highly divergent T-cell receptor binding modes underlie specific recognition of a bulged viral peptide bound to a human leukocyte antigen class I molecule. *J. Biol. Chem.* 288, 15442–15454. doi: 10.1074/jbc.M112.447185

Marcou, Q., Mora, T., and Walczak, A. M. (2018). High-throughput immune repertoire analysis with IGoR. *Nat. Commun.* 9:561. doi: 10.1038/s41467-018-02832-w

McGranahan, N., Furness, A. J. S., Rosenthal, R., Ramskov, S., Lyngaa, R., Saini, S. K., et al. (2016). Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science* 351, 1463–1469. doi: 10.1126/science.aaf1490

Milighetti, M., Shawe-Taylor, J., and Chain, B. M. (2021). Predicting T cell receptor antigen specificity from structural features derived from homology models of receptor-peptide-major histocompatibility complexes. *bioRxiv* 2021.05.19.444843. doi: 10.1101/2021.05.19.444843

Moris, P., De Pauw, J., Postovskaya, A., Gielis, S., De Neuter, N., Bittremieux, W., et al. (2020). Current challenges for unseen-epitope TCR interaction prediction

and a new perspective derived from image classification. *Brief. Bioinformatics* 22:bbaa318. doi: 10.1101/2019.12.18.880146

Ogishi, M., and Yotsuyanagi, H. (2019). Quantitative prediction of the landscape of T cell epitope immunogenicity in sequence space. *Front. Immunol.* 10:827. doi: 10.3389/fimmu.2019.00827

Ostmeyer, J., Christley, S., Toby, I. T., and Cowell, L. G. (2019). Biophysicochemical motifs in T-cell receptor sequences distinguish repertoires from tumor-infiltrating lymphocyte and adjacent healthy tissue. *Cancer Res.* 79, 1671–1680. doi: 10.1158/0008-5472.CAN-18-2292

Peacock, T., and Chain, B. (2021). Information-driven docking for TCR-pMHC complex prediction. *Front. Immunol.* 12:1952. doi: 10.3389/fimmu.2021.686127

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2012). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.

Petersen, J., Ciacchi, L., Tran, M. T., Loh, K. L., Kooy-Winkelaar, Y., Croft, N. P., et al. (2020). T cell receptor cross-reactivity between gliadin and bacterial peptides in celiac disease. *Nat. Struct. Mol. Biol.* 27, 49–61. doi: 10.1038/s41594-019-0353-4

Piepenbrink, K. H., Blevins, S. J., Scott, D. R., and Baker, B. M. (2013). The basis for limited specificity and MHC restriction in a T cell receptor interface. *Nat. Commun.* 4:1948. doi: 10.1038/ncomms2948

Pogorelyy, M. V., Minervina, A. A., Chudakov, D. M., Mamedov, I. Z., Lebedev, Y. B., Mora, T., et al. (2018). Method for identification of condition-associated public antigen receptor sequences. *eLife* 7:e33050. doi: 10.7554/eLife.33050.015

Pogorelyy, M. V., Minervina, A. A., Shugay, M., Chudakov, D. M., Lebedev, Y. B., Mora, T., et al. (2019). Detecting T cell receptors involved in immune responses from single repertoire snapshots. *PLoS Biol.* 17:e3000314. doi: 10.1371/journal.pbio.3000314

Reinherz, E. L., Tan, K., Tang, L., Kern, P., Liu, J. H., Xiong, Y., et al. (1999). The crystal structure of a T cell receptor in complex with peptide and MHC class II. *Science* 286, 1913–1921. doi: 10.1126/science.286.5446.1913

Sidhom, J.-W., Larman, H. B., Pardoll, D. M., and Baras, A. S. (2021). DeepTCR is a deep learning framework for revealing sequence concepts within T-cell repertoires. *Nat. Commun.* 12:1605. doi: 10.1038/s41467-021-22667-2

Singh, N. K., Abualrous, E. T., Ayres, C. M., Noé, F., Gowthaman, R., Pierce, B. G., et al. (2020). Geometrical characterization of T cell receptor binding modes reveals class–specific binding to maximize access to antigen. *Proteins* 88, 503–513. doi: 10.1002/prot.25829

Springer, I., Besser, H., Tickotsky-Moskovitz, N., Dvorkin, S., and Louzoun, Y. (2020). Prediction of specific TCR-peptide binding from large dictionaries of TCR-peptide pairs. *Front. Immunol.* 11:1803. doi: 10.3389/fimmu.2020.01803

Stone, J. D., Chervin, A. S., and Kranz, D. M. (2009). T-cell receptor binding affinities and kinetics: impact on T-cell activity and specificity. *Immunology* 126, 165–176. doi: 10.1111/j.1365-2567.2008.03015.x

Teraguchi, S., Saputri, D. S., Llamas-Covarrubias, M. A., Davila, A., Diez, D., Nazlica, S. A., et al. (2020). Methods for sequence and structural analysis of B and T cell receptor repertoires. *Comput. Struct. Biotechnol. J.* 18, 2000–2011. doi: 10.1016/j.csbj.2020.07.008

Thomas, N., Best, K., Cinelli, M., Reich-Zeliger, S., Gal, H., Shifrut, E., et al. (2014). Tracking global changes induced in the CD4 T-cell receptor repertoire by immunization with a complex antigen using short stretches of CDR3 protein sequence. *Bioinformatics* 30, 3181–3188. doi: 10.1093/bioinformatics/btu523

Thomas, S., Mohammed, F., Reijmers, R. M., Woolston, A., Stauss, T., Kennedy, A., et al. (2019). Framework engineering to produce dominant T cell receptors with enhanced antigen-specific function. *Nat. Commun.* 10:4451. doi: 10.1038/s41467-019-12441-w

Thomas, S., Xue, S.-A., Bangham, C. R. M., Jakobsen, B. K., Morris, E. C., and Stauss, H. J. (2011). Human T cells expressing affinity-matured TCR display accelerated responses but fail to recognize low density of MHC-peptide antigen. *Blood* 118, 319–329. doi: 10.1182/blood-2010-12-326736

Tong, Y., Wang, J., Zheng, T., Zhang, X., Xiao, X., Zhu, X., et al. (2020). SETE: Sequence-based Ensemble learning approach for TCR Epitope binding prediction. *Comput. Biol. Chem.* 2020:107281. doi: 10.1016/j.compbiolchem.2020.107281

Venturi, V., Kedzierska, K., Price, D. A., Doherty, P. C., Douek, D. C., Turner, S. J., et al. (2006). Sharing of T cell receptors in antigen-specific responses is driven by convergent recombination. *Proc. Natl. Acad. Sci. U.S.A.* 103, 18691–18696. doi: 10.1073/pnas.0608907103

Wang, Y., Singh, N. K., Spear, T. T., Hellman, L. M., Piepenbrink, K. H., McMahan, R. H., et al. (2017). How an alloreactive T-cell receptor achieves peptide and MHC specificity. *Proc. Natl. Acad. Sci. U.S.A.* 114, E4792–E4801. doi: 10.1073/pnas.1700459114

Williams, C. J., Headd, J. J., Moriarty, N. W., Prisant, M. G., Videau, L. L., Deis, L. N., et al. (2018). MolProbity: more and better reference data for improved all-atom structure validation. *Protein Sci.* 27, 293–315. doi: 10.1002/pro.3330

Yang, X., Chen, G., Weng, N. p., and Mariuzza, R. A. (2017). Structural basis for clonal diversity of the human T-cell response to a dominant influenza virus epitope. *J. Biol. Chem.* 292, 18618–18627. doi: 10.1074/jbc.M117.810382

Yin, L., Huseby, E., Scott-Browne, J., Rubtsova, K., Pinilla, C., Crawford, F., et al. (2011). A single T cell receptor bound to major histocompatibility complex class I and class II glycoproteins reveals switchable TCR conformers. *Immunity* 35, 23–33. doi: 10.1016/j.immuni.2011.04.017

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.