

1 Predicting T cell receptor antigen specificity from
2 structural features derived from homology models of
3 receptor-peptide-major histocompatibility
4 complexes.

5 Martina Milighetti^{a, b}, John Shawe-Taylor^c and Benjamin M Chain^{a, c}

6 ^a*Division of Infection and Immunity, University College London, London, United Kingdom*

7 ^b*Cancer Institute, University College London, London, United Kingdom*

8 ^c*Department of Computer Science, University College London, London, United Kingdom*

9 1 Abstract

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

21 2 Introduction

22 T cells are key players of adaptive immunity. They are activated by the recognition of
23 a cognate peptide, a short stretch of amino acids which is displayed on a major histo-
24 compatibility complex molecule (MHC, pMHC when bound to peptide). The recognition
25 occurs via the T cell receptor (TCR), which is composed of two chains (normally an
26 α and a β), both of which are generated by a process of random recombination and
27 selection. The recombination gives rise to 3 hypervariable regions, the complementarity-
28 determining regions - CDR1, CDR2 and CDR3. Among the three regions, CDR3 is the
29 most variable as it is found at the junction of V(D)J recombination, and it can therefore
30 incorporate a number of non-template insertion and deletion events, whilst CDR1 and

31 CDR2 depend on the V gene selected in the recombination process and have therefore a
32 lower number of possible sequences.

33 A number of TCR-pMHC complexes have been crystallised and the structures solved
34 and they are collected in the Structural T-Cell Receptor Database (STCRDab, Leem et
35 al. 2018). They have given us deeper understanding of TCR-pMHC interactions and how
36 these are impacted by mutations, but also how structure and function are related. Ex-
37 amples include how cross-reactivity between bacterial and self antigens can drive disease
38 (Petersen et al. 2020), how binding mode can give different specificity profiles to TCRs
39 binding the same peptide (Coles et al. 2020), and how binding orientation is determined
40 by how the peptide is presented by the MHC (Singh et al. 2020).

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

51 The ensemble of TCRs that are present within an individual at any point in time is
52 called the TCR repertoire. Different sequences are found at widely different frequencies,
53 ranging from a few hundred copies to 10^9 copies for the larger T cell clones, which make
54 up to 1% of the total repertoire. Differences in clone size can arise both in the

55 naive repertoire, by convergent recombination (whereby an amino acid sequence is likely
56 to be produced by the process of recombination - normally with short CDR3 and few
57 nucleotide insertions, Venturi et al. 2006; Britanova et al. 2014) or because of the power-
58 law distribution of naive T cell clones produced by the thymus (Greef et al. 2020); or in
59 the memory repertoire by convergent selection, whereby similar sequences are expanded
60 because they are responding to the same antigen, Pogorelyy et al. 2018). Greef et al.
61 2020 estimates the maximum effect of generation probability to be around 10^7 , which
62 is two order of magnitudes smaller than the largest observed clone sizes, suggesting a
63 role for expansion during the immune response. By focusing solely on the CDR3, it can
64 be shown that during an immune response, expanded TCR clones are frequently part
65 of clusters of sequences that are more similar to each other than might be expected by
66 random sampling of the repertoire (Joshi et al. 2019; Pogorelyy et al. 2019; Marcou et al.
67 2018).

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

75 In the same way that conserved sequence motifs characterise TCRs recognising the
76 same antigen, we hypothesise that there will be structural features of the TCR/antigen
77 interface which are conserved in the interactions. Such conserved structural features could
78 be leveraged to gain a better understanding of the TCR-pMHC interaction and to reca-

79 pitulate and improve what has been learnt from looking purely at sequence information.

80 Our understanding of the physical interactions between TCRs and pMHC is, however,

81 limited to the set of solved and published crystal structures. The STCRDab currently re-

82 ports about 400 entries for $\alpha\beta$ TCR-pMHC complexes, and 120 different peptides, which

83 is clearly a tiny subset of all the possible TCR-pMHC interactions that can exist. To

84 solve this problem, a number of tools have been developed and subsequently optimised

85 to predict the structure of a TCR-pMHC complex based on its sequence. One of these

86 is TCRpMHCmodels (Jensen et al. 2019), which exists as a free online user interface.

87 TCRpMHCmodels leverages LYRA (Klausen et al. 2015) to model the TCR structure

88 and MODELLER (Fiser and Šali 2003) to predict the pMHC structure, to then combine

89 them together by using a third set of templates for the TCR-pMHC complex overall.

90 Tools like TCRpMHCmodels, although still limited by the amount of information that

91 has been published, allow us to delve deeper into the structural relationships between the

92 TCR and the pMHC.

93 We show here that a combination of structural and sequence features can be in-

94 corporated into a machine learning algorithm to discriminate binding and non-binding

95 TCR-pMHC pairs. The classifier presented is limited by the performance of the homology

96 modelling, but, unlike any of the previous work reviewed above, it does not rely on the

97 identification of a set of TCRs binding to a specific peptide to be able to predict whether

98 other TCRs will bind to that same peptide, but rather learns some general rules which

99 can predict TCR interaction with completely novel peptides.

100 **3 Methods**

101 **3.1 Datasets**

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

110 The 10XGenomics dataset was downloaded from the 10XGenomics website (CD8+ T
111 cells of Healthy Donor 1, *A New Way of Exploring Immunity - Linking Highly Multiplexed*
112 *Antigen Recognition to Immune Repertoire and Phenotype*). For each TCR, binding (or
113 absence of binding) to an epitope was defined as in the Application Note provided by
114 10X Genomics. Briefly, a specific binding event was defined as having UMI count higher
115 than 10 and greater than 5 times the highest negative control for that TCR clone. When
116 a TCR clone was assigned multiple barcodes, the UMI counts for each tetramer were
117 summed to determine overall binding. If these conditions were true for more than one
118 peptide, the TCR was called a binder for each of the epitopes.

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

123 The set of experimental constructs (expt) consists of a set of experimentally-validated
124 peptide-specific TCR constructs with cognate peptide, which have been characterised
125 functionally: 2 CMV-reactive TCRs (NLVPMVATV peptide), 3 influenza-reactive TCRs
126 (2 HA1-reactive - peptide VLHDDLLEA - and 1 HA2-reactive - YIGEVLVSV peptide),
127 1 EBV-reactive TCR (peptide CLGGLLTMV) from Thomas et al. 2019 and Chatterjee
128 et al. 2019; A7 TCR and 3 affinity-matured TCRs from A7 which recognise pTax as
129 well as pHud peptides (LLFGYPVYV and LGYGFVNYI, respectively) (Thomas et al.
130 2011); two TCRs identified as neoantigen-reactive in Joshi et al. 2019 and two mutated
131 versions of these, which have been shown not to bind the neoantigen (unpublished data,
132 A. Woolston, personal communication, 2020). To create the non-binders, each TCRs was
133 matched with each pMHC in the pool, as well as with peptide WT235 (control peptide
134 in Thomas et al. 2019, CMTWNQMNL) and peptide WTlung (FAFQEDDSF, wild-type
135 peptide for the neo-antigen McGranahan et al. 2016).

136 A dataset of TCR-pMHC complexes with experimentally-determined affinity was re-
137 trieval from the ATLAS (<http://atlas.wenglab.org/web/index.php>, Borrman et al.
138 2017) to evaluate the impact of affinity on the classifier performance. Any TCR-pMHC
139 pair with undetectable binding (K_d labelled as *n.d.*) was called a non-binder whilst all
140 other complexes were labelled binders regardless of the detected K_d .

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

¹⁴⁵ 3.2 Homology modelling and feature extraction

¹⁴⁶ Each structure (both binders and non-binders) in these datasets was homology-modelled
¹⁴⁷ with TCRpMHCmodels (which was kindly provided in command-line form by the authors,
¹⁴⁸ Jensen et al. 2019) in its default settings and submitted to the feature-extraction pipeline.

¹⁴⁹ To make the structures comparable, they were renumbered to the standardised IMGT
¹⁵⁰ numbering (Lefranc 1997) using ANARCI (Dunbar and Deane 2016). Moreover, the
¹⁵¹ peptide residues were renumbered to 1-20, so that the central residues would be residues
¹⁵² 10-11 in each complex.

¹⁵³ For each TCR-pMHC, 5 sets of features were extracted, namely:

¹⁵⁴ • minimum pairwise distances between each CDR residue and each peptide residue
¹⁵⁵ were calculated using BioPDB (Hamelryck and Manderick 2003). These capture
¹⁵⁶ the binding mode of the TCR-pMHC complex;

¹⁵⁷ • energetic profile of pairwise CDR-peptide residues interactions was calculated us-
¹⁵⁸ ing PyRosetta v2020.28+ (Chaudhury et al. 2010). The Rosetta energy function
¹⁵⁹ for context-independent residue-residue interactions was used to extract the fol-
¹⁶⁰ lowing terms (scorefunction: talaris2014) from a PDB file from which the MHC
¹⁶¹ complex was removed: attractive and repulsive van der Waals (atr, rep), electro-
¹⁶² static interactions (elec) and solvation energy (sol) (Alford et al. 2017). These are
¹⁶³ a representation of binding energy of the complex.

¹⁶⁴ • Atchley factors (Atchley et al. 2005) were used to encode the sequences of the
¹⁶⁵ peptide and CDRs for each TCR-pMHC pair.

¹⁶⁶ To evaluate the effect of homology modelling performance on the classifier presented,
¹⁶⁷ the structures were categorised as having or not having good homology modelling tem-

168 plates. This was defined based on the sequence homology to the most similar peptide
169 template (> 45% sequence similarity to the best pMHC model template) and complex
170 template (> 60% sequence similarity to the best complex template). These thresholds
171 were chosen based on the results presented by Jensen et al. 2019.

172 To be noted that not all structures could be successfully modelled by TCRpMHC-
173 models, and so we could not submit them to the feature extraction pipeline.

174 **3.3 Multiple kernel learning**

175 Each feature set was pre-processed separately. Missing values were imputed with the
176 median value of the feature across the train set. Each feature was then scaled to have a
177 value between 0 and 1 (sci-kit learn Minmax scaler, Pedregosa et al. 2011) and normalised.

178 To properly represent and integrate the different features extracted from the struc-
179 tures, kernels were created separately for each subset of features. Moreover, instead of
180 optimising a single kernel for each feature set, 7 Gaussian (rbf) kernels were created and
181 combined, letting the MKL algorithm decide the weights for each kernel, as in Lauriola
182 et al. 2017. The γ parameters for the 7 Gaussian kernels for each feature set were found
183 as follows:

- 183 1. calculate the distance between all positive (binding, n) and negative (non-binding,
m) examples in the train set

$$d = \sqrt{\sum_{i,j=1}^{n,m} (pos_i - neg_j)^2}$$

- 184 2. find σ values corresponding to 1st, 2nd, 5th, 50th, 55th, 98th and 99th percentile of
185 distances

3. for each σ , calculate the γ as:

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

186 The kernels generated were combined by the EasyMKL algorithm as implemented in
187 MKLPy to find an optimal combination (Aiolli and Donini 2015; Lauriola et al. 2017;
188 Lauriola and Aiolli 2020), setting sci-kit's learn SVC algorithm as a learner (Pedregosa
189 et al. 2011). The λ parameter for EasyMKL was fixed to 0 and the optimal C parameter
190 for SVC was searched in the range between 10^{-5} and 10^2 by 10-fold (internal) cross-
191 validation (CV) on the train set. This process was used both when a single feature set
192 was evaluated (by combining the 7 kernels for the set) and when combining multiple
193 feature sets (7 kernels for each set).

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

198 Out-of-sample performance was evaluated in the datasets outlined in section 3.1, by
199 training the classifier on the whole of the training set.

200 **3.4 Benchmarking against other classifiers**

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

206 //github.com/pmoris/ImReX), and the available model trained on the VDJdb was used
207 for the predictions.

208 **3.5 Data availability**

209 The complete set of sequences used, as well as prediction results are provided as supple-
210 mentary files.

211 **4 Results**

212 **4.1 Extracting physical features from available TCR-pMHC com-**
213 **plex structures allows interrogation of binding mode**

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

220 Each feature extraction process yielded a matrix with information about pairwise
221 contacts between residues in the TCR and residues in the peptide (Figure 1a). The
222 distance fingerprints are easy to compare between different structures and can give insight
223 into the binding mode for the complex: for instance, complexes 1AO7 (Garboczi et al.
224 1996) and 1MI5 (Kjer-Nielsen et al. 2003) (both MHC Class I) bind closer to the N
225 terminus of the peptide, whilst 1D9K (Reinherz et al. 1999) has the TCR bound more

226 centrally, and this is particularly evident in the α chain (Figure 1a and b).

227 We wondered whether any trends could be detected more generally and used the
228 minimum pairwise distances to identify the distribution of interactions between TCR
229 CDR residues and the peptide in class I and class II complexes (Figure 1c). While it
230 is clear that interactions between TCR chains and antigen peptide are not confined to
231 a single hotspot, some general patterns emerge. The TCR α chain, for example, tends
232 to bind the N-terminus of the peptide, whilst the β binds towards the C-terminus, as
233 has been reported previously (Garcia et al. 2009). Interestingly, while contacts were
234 dominated by the CDR3 region of the TCR, we also detected contacts between CDR1
235 and CDR2 and peptide residues in a significant proportion of complexes. Moreover, more
236 of the class I structures make contacts with the C-terminus of the peptide than class II. A
237 similar pattern is also detected when looking at the energetic interactions (Supplementary
238 Figure S1).

239 In order to look in more detail for potential conserved patterns with which to char-
240 acterise the TCR-peptide binding surface, we calculated a PCA for each of the feature
241 sets (distances and energy vectors) for all complexes (Figure 2a and Supplementary Fig-
242 ure S2a). The first dimension of the PCA of the minimum pairwise distances clearly
243 identified the few examples where the TCR is in an inverse orientation relative to the
244 peptide (stars, PDB: 4Y19 and 4Y1A Beringer et al. 2015, 5SWS and 5SWZ Gras et al.
245 2016). The second dimension of the distance PCA, on the other hand, seemed to par-
246 tially discriminate between class I and class II complexes. To gain some insight in to
247 which structural features were driving this separation, we looked at the distance vectors
248 that were used for each structure (Figure 2b, left). Both for the α and the β chains,
249 a shift towards the peptide C terminus was observed with decreasing PC2 values. Four

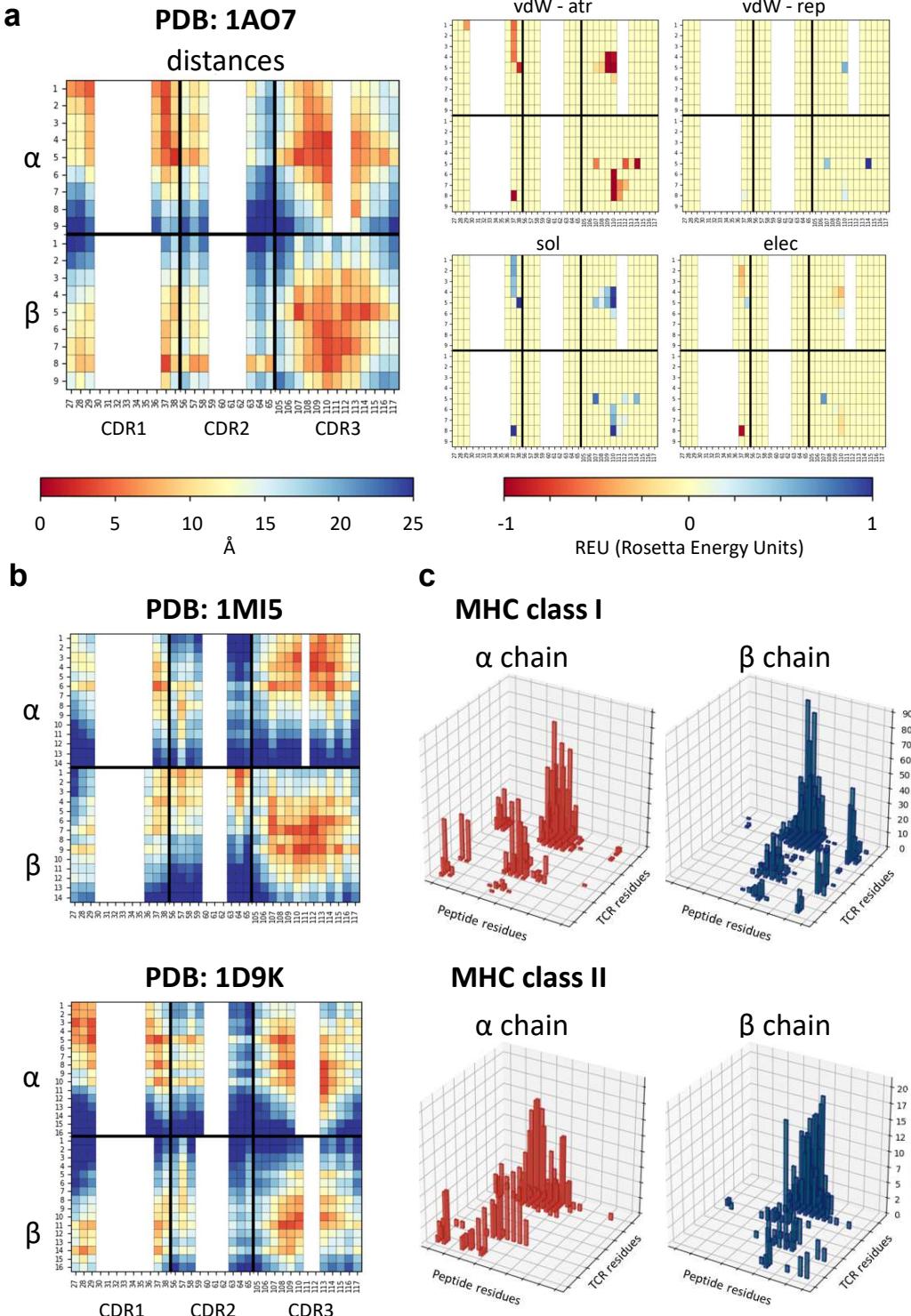


Figure 1: *Caption next page*

Figure 1: (*Previous page.*) **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 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 TCR α 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 (less than 6 \AA) 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

250 representative fingerprints from the edges of the PCA plot are also shown in which the
251 inverted orientation of 4Y19 and 5SWS as well as the shift towards the N terminus for
252 5TEZ (Yang et al. 2017) are apparent, compared to 3RGV (Yin et al. 2011). In agreement
253 with Figure 1c, class II complexes tend to have higher PC2, which is associated with a
254 shift towards binding at the N terminus of the peptide. 3RGV, which segregates with
255 the class II complexes, is actually a class I complex. Interestingly, however, the YAe62
256 TCR in the 3RGV complex is reported by the authors to bind both class I and class II
257 complexes with similar orientations, which might explain its positioning with other class
258 II complexes. Strikingly, the other class I complex found with high PC2 is 4JRY, which
259 is also reported to bind with unusual position on top of the N-terminus of the peptide,
260 rather than centrally, where the peptide bulges out (Liu et al. 2013).

261 A similar analysis was done on the solvent energy vectors (Figure 2). The PCA
262 suggested a segregation between class I and class II complexes along PC1, although
263 significant overlap was also observed. We therefore looked at what features could be
264 driving the separation along the PC1 (Supplementary Figure S2b). The only evident

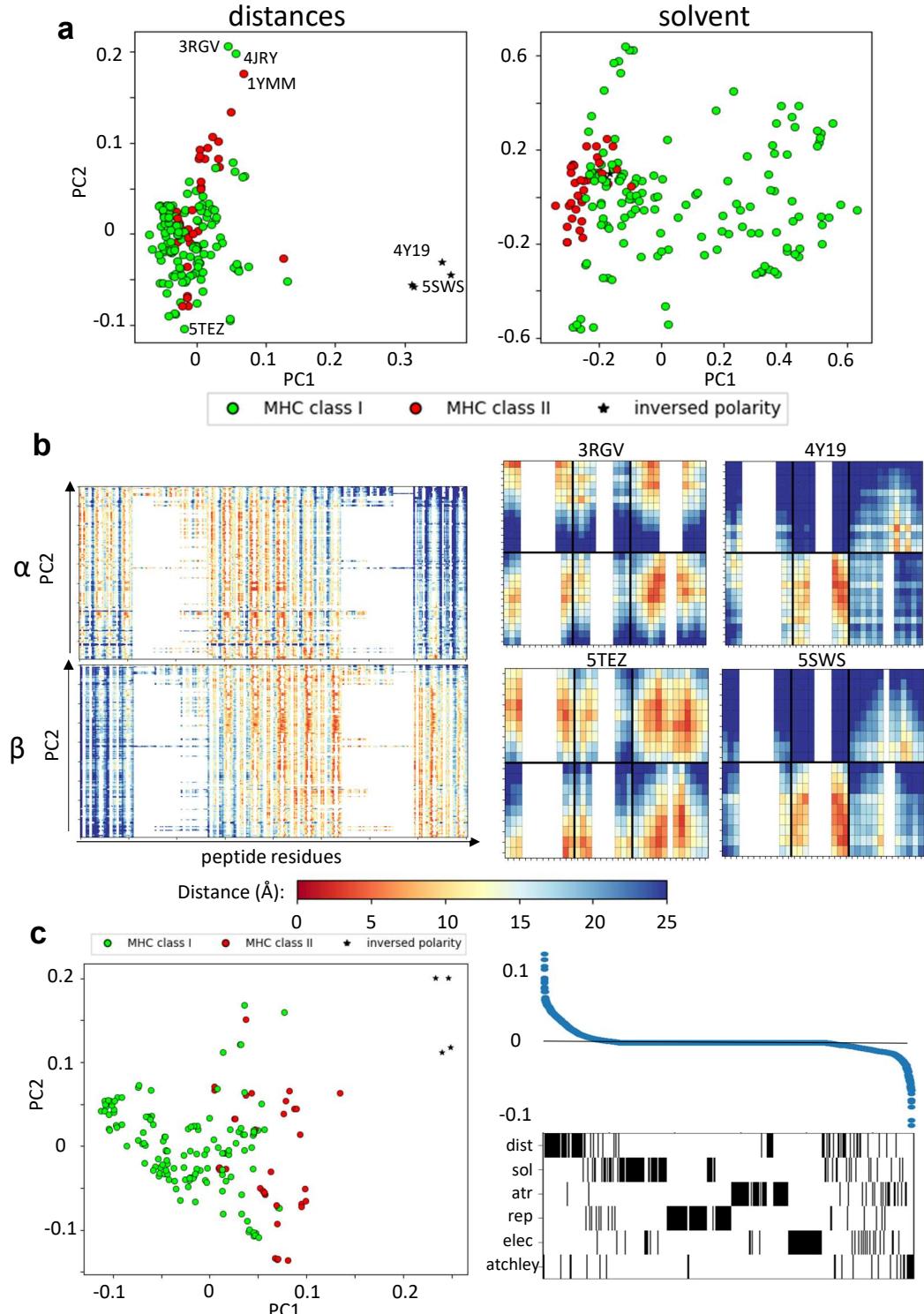


Figure 2: *Caption next page*

Figure 2: (Previous page.) **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 inverted polarity (i.e. the TCRs bind the pMHC complex at 180 degree angle). 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 4 representative structures labelled in panel **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.

265 trend was that all the complexes with high PC1 show a strong unfavourable interaction
266 between the β chain and the peptide C terminus (blue in the heatmap). As solvent energy
267 is positive (i.e. unfavourable) when a residue is not solvent-exposed, this suggests that
268 the complexes with higher PC1 make an interaction between the beta chain and the C
269 terminus of the peptide.

270 Finally, all distance and energy feature sets were combined in a single PCA plotted in
271 Figure 2c (left). Here, the structures with inverted polarity have high PC1, followed by
272 MHC class II complexes and on the left-hand side of the plot are the class I complexes.
273 The loadings of each feature in the set were calculated and the features ranked by loading
274 value (Figure 2c, right). Most of the features which had absolute values greater than 0
275 (i.e. positive or negative), belong to the distance, the solvent energy or to the Atchley
276 factors datasets, suggesting that these have the strongest discriminatory power.

277 Overall, these results gave us confidence that meaningful information about the bind-
278 ing interface could be extracted with our pipeline.

279 **4.2 Structural information from homology modelled structures**
280 **cannot distinguish binding pairs in unsupervised settings**

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

291 Because we needed to rely on a structure prediction method, we first evaluated the
292 difference between the features extracted from the original crystallographic structures
293 and from their respective modelled structures (Figure 3 and Supplementary Figure S3a).
294 Taking complex 1AO7 as an example, the fingerprints obtained from the original PDB and
295 from the predicted structures were plotted (Figure 3a). The two complexes have RMSD
296 of about 2 Å and it can be seen that the contacts seem to be slightly shifted towards the
297 N terminus of the peptide in the predicted structure compared to the crystal. However,
298 the two fingerprints did not look drastically different.

299 When combining all feature sets and looking at all structures available by PCA, no
300 systematic difference was found between modelled and original structures (Figure 3b
301 and c and Supplementary Figure S3a). There was reasonably good matching between

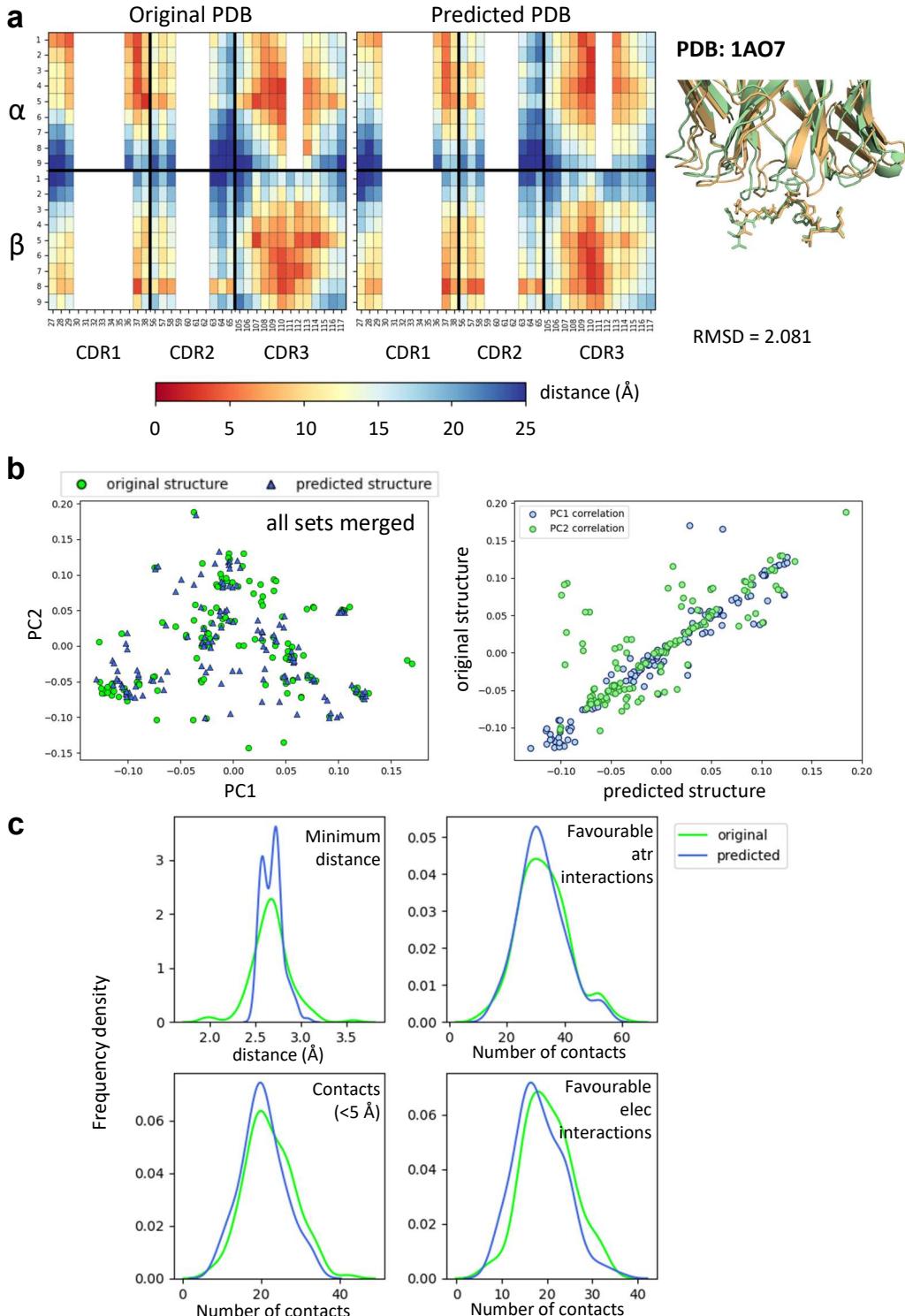


Figure 3: *Caption next page*

Figure 3: (Previous page.) 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 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 S3a. **c.** Frequency distributions of 4 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 less than 5A apart; Favourable atr/elec interactions: number of favourable (energy < 0) interactions between TCR and peptide.

302 the crystal strucutres and their homology models, although TCRpMHCmodels failed to
303 predict non-canonical binding models. We also compared the distributions of some of the
304 structural features (minimum distance between peptide and TCR, number of contacts
305 and number of favourable interactions), and in general found reasonably good agreement
306 between models and structures. As homology modelling gave us reliable predictions and
307 was necessary to create our negative examples, we decided to use modelled structures for
308 both binding and non-binding complexes, in order to avoid introducing systematic bias.

309 To create a set of non-binders, a set of shuffled TCR-pMHC complexes from the
310 STCRDab was used (Figure 4a). We then asked whether the structures predicted for
311 non-binders could be discriminated from the binders.

312 Strikingly, there was no dsicernible separation of binders and non-binders on un-
313 supervised PCAs with any of the distance or energy sets of features (Figure 4b and
314 Supplementary Figure S3b). Basic metrics such as the minimum distance between TCR
315 and peptide and the number of contacts showed similar distributions for binders and
316 non-binders (Figure 4b).

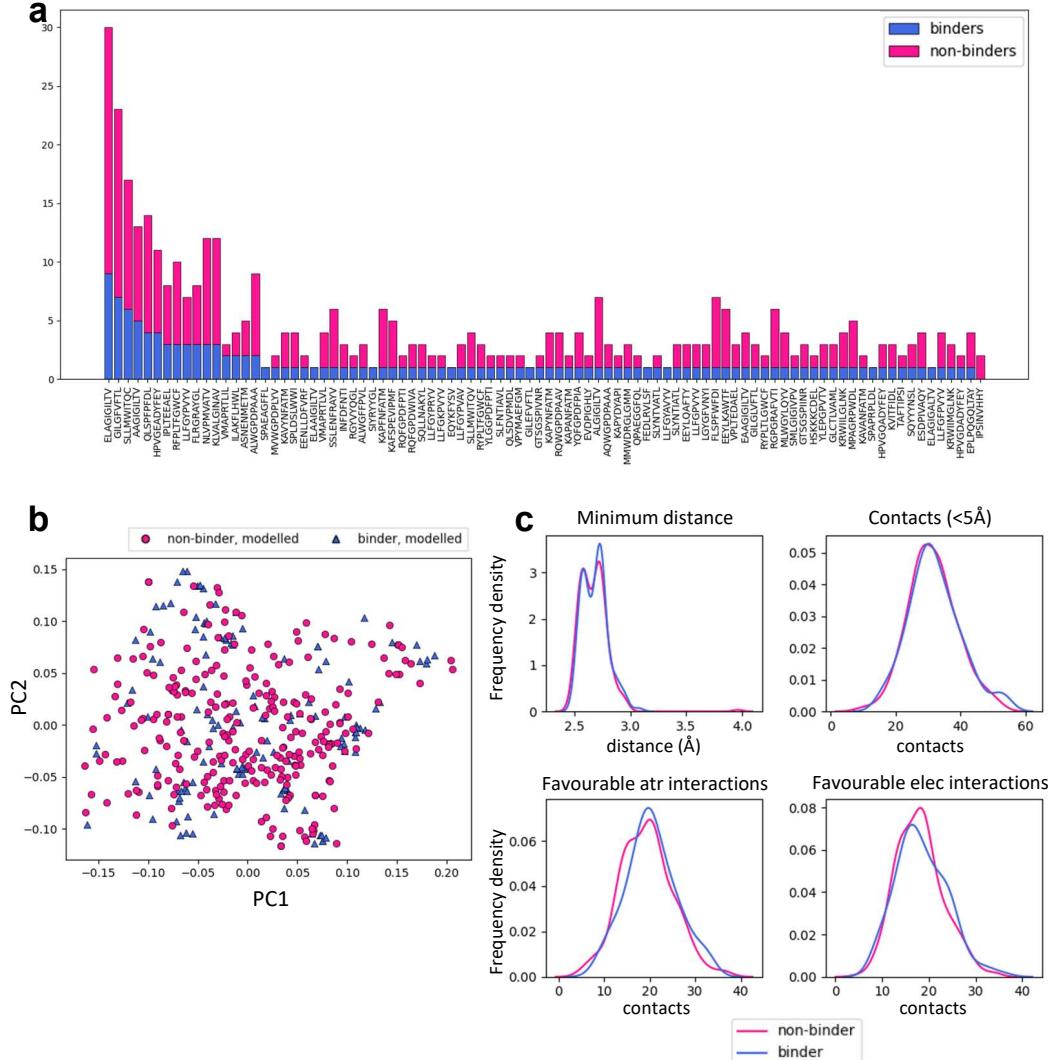


Figure 4: *Caption next page*

Figure 4: (Previous page.) 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 S3b. **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 less than 5A apart; Favourable atr/elec interactions: number of favourable (energy < 0) interactions between TCR and peptide.

317 4.3 Structural information can discriminate between binders
318 and non-binders using supervised learning

319 We turned to supervised machine learning methods to try and better discriminate be-
320 tween binding and non-binding pairs. We explored multiple kernel learning (MKL) to
321 combine information from the different feature sets extracted from the modelled interac-
322 tion surfaces using the pipeline explained above. To assess the potential of our method,
323 a model was trained and tested by cross-validation, using predicted structures derived
324 from the STCRDab, creating a dataset of positives and negatives as described in the
325 methods. Figure 5a and c show the results of 10-fold cross-validation when each different
326 feature set is used separately. Whilst Atchley factors provide the single strongest predic-
327 tive power (average ROC AUC of 0.763), similar discrimination can be obtained by using
328 distances only (ROC AUC of 0.755), followed closely by attractive van der Waals forces
329 (atr, ROC AUC of 0.74) and solvent energies (ROC AUC of 0.701). The other energetic
330 terms generally showed poorer performance and were excluded from further analysis.

331 We next combined the feature sets to create a single classifier (Figure 5b and c). Using
332 Atchley factors, distances and attractive van der Waals forces together achieved a similar

333 performance to using each set of features independently, whilst combination of Atchley
334 factors and distances only gave a slight increase in performance compared to each of the
335 two sets separately. Interestingly, although performance did not change much in this
336 more complex model, the weights assigned to the kernels constructed for each feature set
337 were similar, suggesting that no single feature set was more important than the others in
338 the overall model.

339 We then went on to validate the trained model on the other 5 datasets described in
340 the methods. Because we wanted to test how generalisable the rules that the classifier
341 had learnt were, we did not train the classifier again on the new sets, but used the model
342 trained on the STCRDab set to predict the new complexes. Results from validation
343 are presented in Figure 5d and Supplementary Figure S4 and summarised in Table 1.
344 Overall, the models with the highest ROC AUC consistently included sequence informa-
345 tion. Moreover, addition of structural features often did not improve predictive power.
346 However, structural features often allowed some level of discrimination, independently of
347 the sequence information, suggesting that the model might be learning something about
348 the binding modes of these complexes. Interestingly, the models which used structural
349 features had consistently higher recall.

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

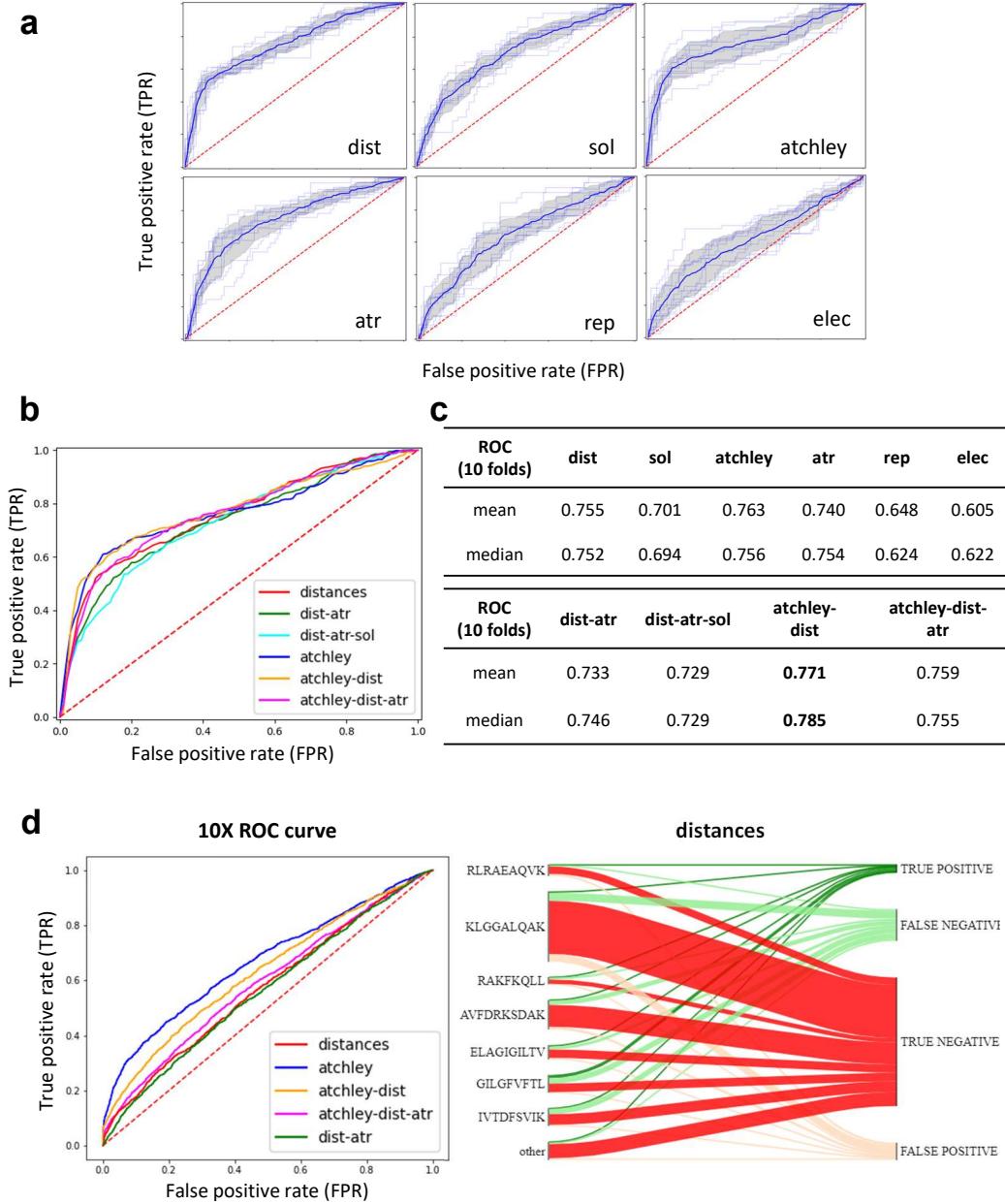


Figure 5: *Caption next page*

Figure 5: (*Previous page.*) **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 line represents 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 curves showed in a. and 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).

³⁵⁷ tetramer-sorting experiments, but rather the complexes show a range of affinities, and it
³⁵⁸ is hard to define a strict threshold to define binding.

set	% pos	combo	roc	avg precision	accuracy	precision	recall
B10x	21.17	distances	0.574	0.289	0.739	0.315	0.198
		dist-atr	0.562	0.260	0.726	0.294	<u>0.210</u>
		atchley	<u>0.668</u>	<u>0.441</u>	<u>0.805</u>	<u>0.751</u>	0.117
	7.34	atchley-dist	0.629	0.375	0.786	0.487	0.166
		atchley-dist-atr	0.590	0.317	0.766	0.382	0.173
		distances	0.591	0.114	0.757	0.116	<u>0.350</u>
Dash	12.70	dist-atr	0.645	0.123	0.802	0.139	0.326
		atchley	<u>0.700</u>	<u>0.188</u>	<u>0.905</u>	<u>0.209</u>	0.107
		atchley-dist	0.599	0.175	0.798	0.133	0.318
	89.06	atchley-dist-atr	0.645	0.146	0.824	0.153	0.309
		distances	0.727	0.326	0.714	0.262	<u>0.688</u>
		dist-atr	0.709	0.423	0.667	0.205	0.563
expt	0.72	atchley	0.816	<u>0.704</u>	<u>0.825</u>	<u>0.393</u>	<u>0.688</u>
		atchley-dist	<u>0.823</u>	0.659	0.754	0.297	<u>0.688</u>
		atchley-dist-atr	0.770	0.515	0.698	0.238	0.625
	0.72	distances	0.487	0.897	0.827	0.892	0.917
		dist-atr	0.518	0.907	0.794	<u>0.901</u>	0.863
		atchley	<u>0.632</u>	<u>0.938</u>	<u>0.891</u>	0.891	<u>1.000</u>
atlas	0.72	atchley-dist	0.551	0.918	<u>0.891</u>	0.891	<u>1.000</u>
		atchley-dist-atr	0.547	0.916	0.865	0.896	0.960
		distances	0.521	<u>0.010</u>	0.865	0.010	<u>0.186</u>
	0.72	dist-atr	0.521	0.008	0.896	<u>0.013</u>	<u>0.186</u>
		atchley	<u>0.570</u>	<u>0.010</u>	<u>0.987</u>	0.000	0.000
		atchley-dist	0.541	<u>0.010</u>	0.954	0.009	0.047
		atchley-dist-atr	0.546	0.008	0.947	0.000	0.000

Table 1: Results of out-of-sample validation. 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.

359 **4.4 Classifier performance varies between epitopes**

360 A known hard task for a classifier trained on a small subset of the epitopes that our
361 immune system is exposed to, is to generalise to epitopes not present in the training
362 set. It is apparent from the diagrams showing mis-classification in Figure 5d (right) and
363 Supplementary Figure S4b that some peptides were indeed easier to classify than others.

364 Figure 6a shows the classifier performance on 4 representative epitopes. For a perfect
365 classifier, the decision score for positive and negative samples (equivalent to the distance of
366 a point from the decision hyperplane in the case of an SVM) should have non-overlapping
367 distributions. However, for peptide antigen AVFDRKSDAK the distributions for binding
368 and non-binding TCRs almost completely overlap, suggesting that the classifier has not
369 learnt useful information from the data. For peptide LLFGYPVYV, on the other hand,
370 the separation between the two groups of TCRs is almost perfect. The classification of
371 TCRs specific for the ELAGIGILTV and ASNENMETM peptides showed an intermediate
372 pattern. Overall, the classification of TCRs for different epitopes show very significant
373 differences in performance, (Figure 6b), as has been observed previously for other models
374 (Moris et al. 2020). This also suggests that the overall performance as showed in Table
375 1 is somewhat misleading, as it will be skewed by the more abundant epitopes.

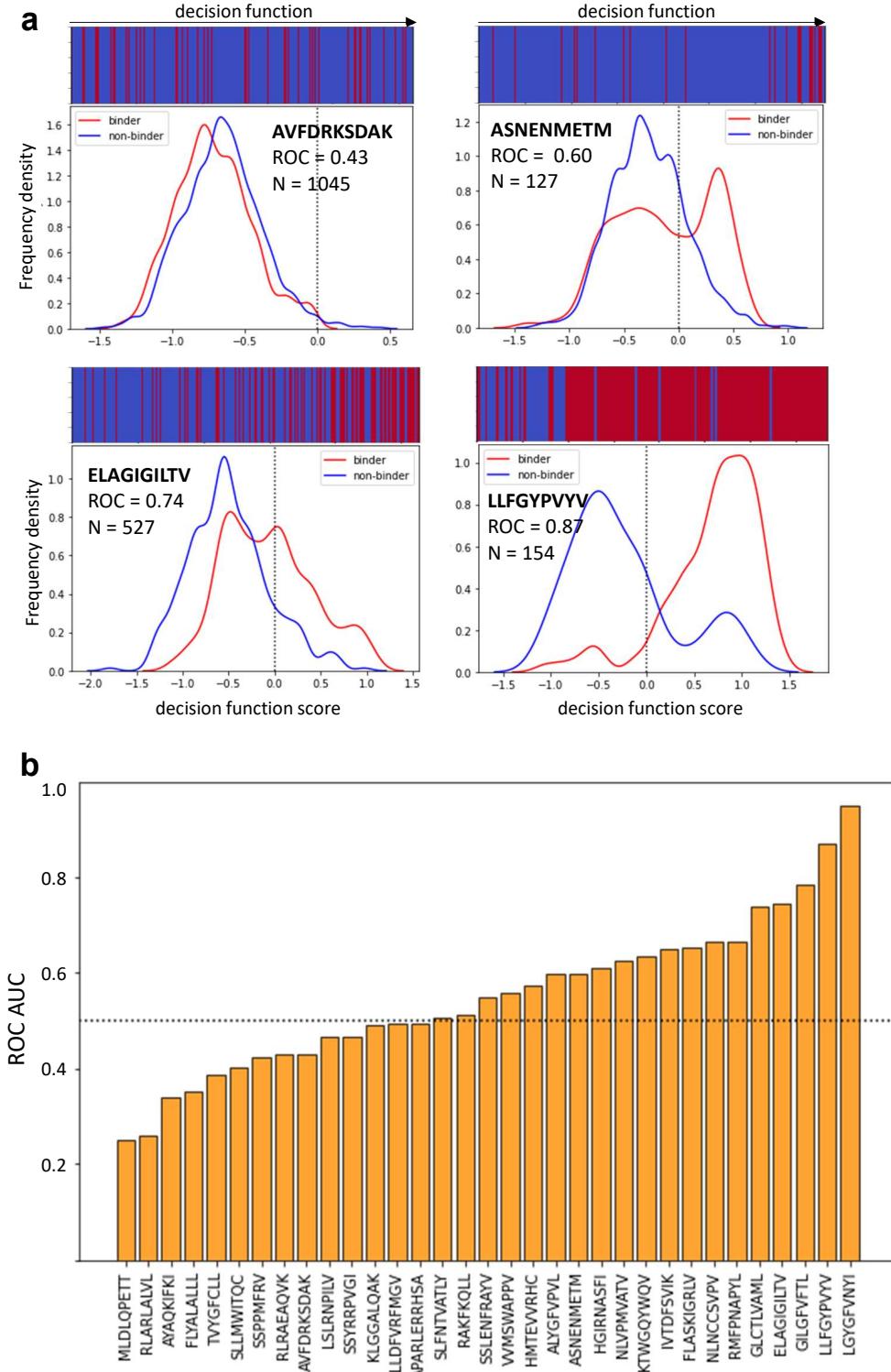


Figure 6: Caption next page

Figure 6: (Previous page.) The performance of the model is pMHC dependent. **a.** Examples of 4 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. 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.

376 **4.5 Homology modelling performance impacts classifier perfor-**
377 **mance**

378 We wondered whether the difference in performance could be due to the performance
379 of the homology modelling tool used. For each structure, we retrieved the information
380 about the sequence similarity between the structure of interest and the template used to
381 model it. We then plotted the classifier performance as a function of sequence similarity
382 (Figure 7a).

383 Overall, there was a trend for better templates (increased sequence similarity) to
384 correlate with better classifier performance (observed as an increase in performance to
385 the right of the individual panels). Interestingly, however, the same trends were observed
386 also when classification was based only on sequence information suggesting that this might
387 not be related only to the accuracy of the homology modelling. The templates for the
388 homology modelling and the training set for our classifier are overlapping sets (as both
389 are using the complexes for which a crystal structure is available) and our results might be
390 reflecting the increased density in the feature space of known complexes. To investigate
391 this, we also computed the BLOSUM scores from the train set for all the complexes
392 we predicted (Figure 7c). Indeed, a decrease in classifier performance is observed when

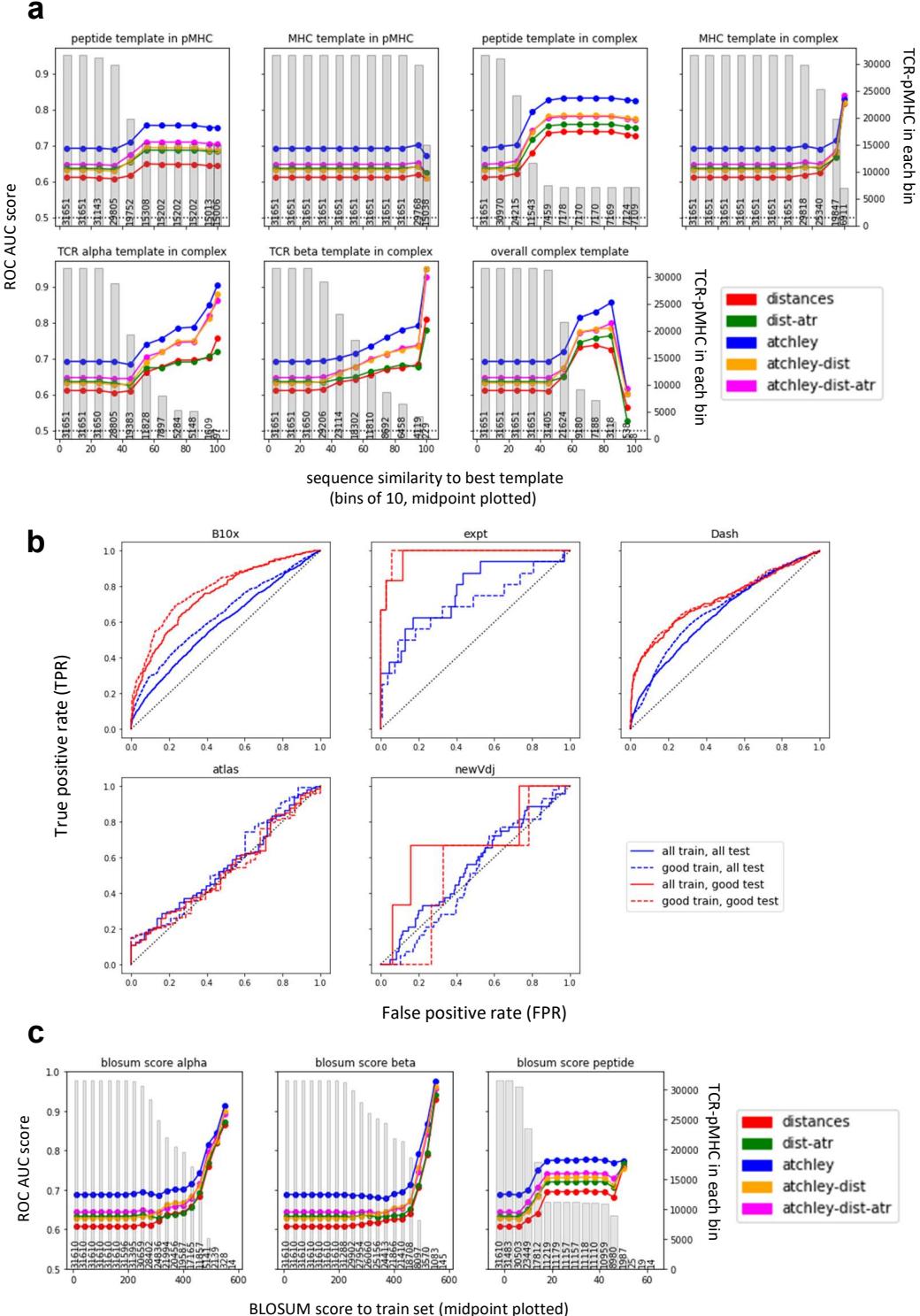


Figure 7: *Caption next page*

Figure 7: (Previous page.) 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 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 grey bars show the number of structures that contribute to the performance for each point. **b.** Performance of each of the validation set when the model is trained on the entire STCRDab set (all train) or only the STCRDab structures with good templates (as defined in methods - good train), and when predictions are made on all complexes (all test) or only complexes with good templates (good test). **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 grey bars show the number of structures in each bin.

393 the BLOSUM score decreases, i.e. when the TCR-pMHC pair that we are trying to
394 predict is less similar to the training set pairs. Interestingly, in all cases the performance
395 of the classifier is more dependent on TCR homology, than on peptide homology. It
396 is important to note that the observed relationship between classifier performance and
397 sequence homology allow us to predict *a priori* which TCR/peptide binding predictions
398 will carry greater confidence. In fact, by considering the epitope and complex homology
399 templates, we are able to select *a priori* a subset of structures on which our model will
400 perform better (Figure 7b).

401 4.6 Effect of affinity on the predictor

402 Because the classifier relies on structural information and it is trained on the set of TCR-
403 pMHC pairs that have a known crystal structure, we wondered whether the model could
404 predict binding affinity as well as a binary binding/non-binding classification or whether
405 higher decision function scores were assigned to higher-affinity complexes (i.e. whether

406 complexes which bind with high affinity are called binders with higher confidence). To
407 address this, the TCR-pMHC pairs from the ATLAS (Borrman et al. 2017) were retrieved
408 and their score predicted. The score for each complex was then correlated (Spearman) to
409 their measured affinity, removing all complexes with undetectable binding and adjusting
410 the ΔG and K_D as in the original publication (Table 2). Unexpectedly, the only significant
411 correlation was between sequence features (Atchley factors) and k_{off} . The model therefore
412 does not successfully capture the structural information which determines the affinity of
413 the complex and its performance is not biased towards detection of high-affinity pairs.

	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)	-0.076	0.188	-0.057	0.322	-0.006	0.914	-0.048	0.402	0.154	0.099
k_{on} ($M s^{-1}$)	0.126	0.177	0.153	0.101	0.084	0.371	0.173	0.063	0.050	0.592
k_{off} (s^{-1})	0.056	0.551	-0.077	0.412	0.277	0.003	0.106	0.260	-0.070	0.221
ΔG (kcal/mol)	-0.080	0.167	-0.065	0.258	-0.022	0.702	-0.060	0.338	-0.070	0.221

Table 2: Correlations of affinity metrics and decision function scores. Spearman correlation is calculated for each affinity metric for predictions made for each of the models trained.

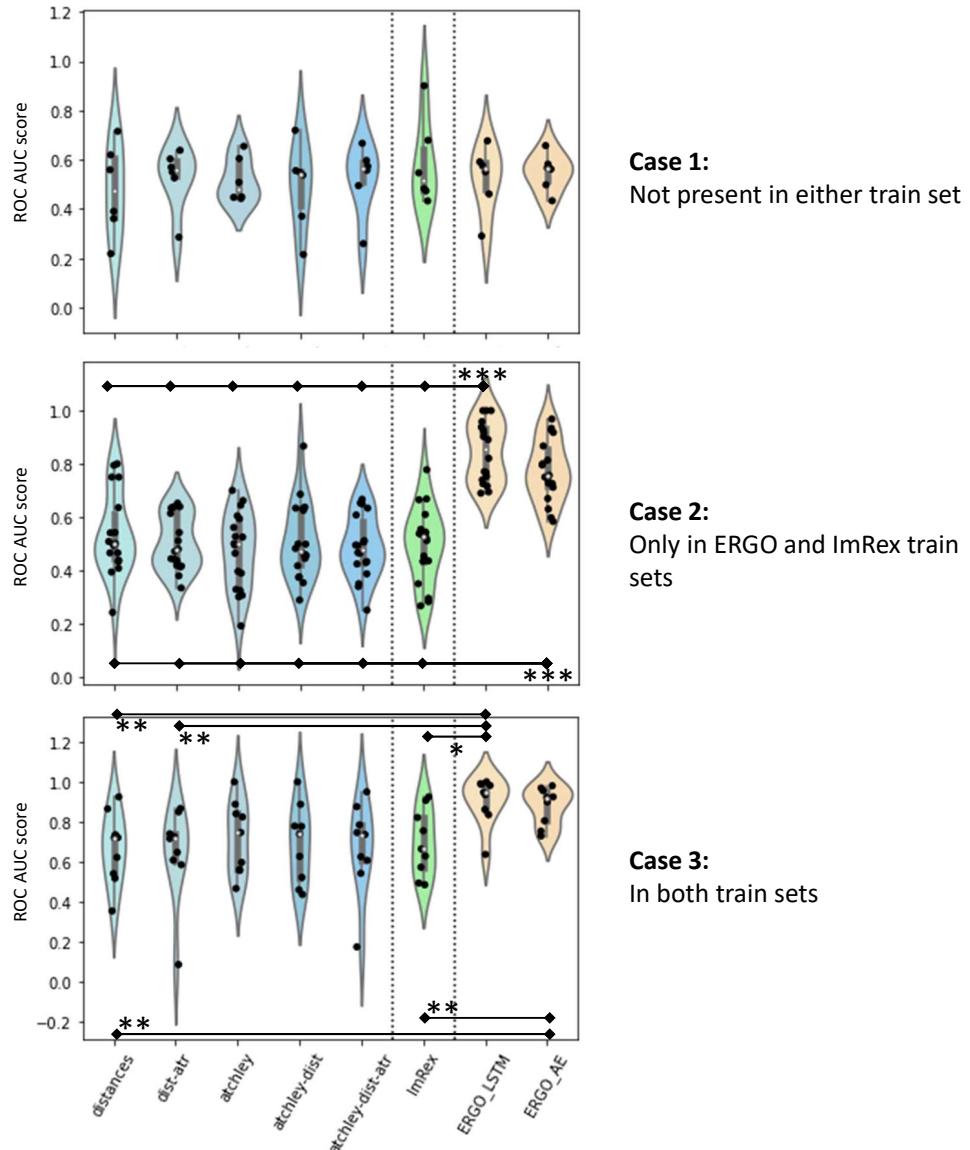
414 4.7 Benchmarking against existing tools

415 Finally, we compared the performance of our classifier against the recently published
416 ERGO (Springer et al. 2020) and ImRex (Moris et al. 2020, Table S1). Both ERGO and
417 ImRex were trained on the VDJdb set (Bagaev et al. 2020), as described in the original
418 publication, rather than the much smaller set of binder used by our algorithm. The
419 trained models are available as an online tool for ERGO (<http://tcr.cs.biu.ac.il/>)
420 and on GitHub for ImRex (<https://github.com/pmoris/ImRex>).

421 The classifiers were all tested on the same set of binder and non-binder TCR-pMHC
422 sets. Figure 8 and Supplementary Table S1 show the results divided by peptide. The
423 results are organised in 3 scenarios depending on whether the peptide is present in neither,
424 either or both of the train sets.

425 When compared on epitopes that are not present in either train set (Case 1), all the
426 models perform in a similar manner. Interestingly, none of the sequence-based classifiers
427 outperforms the structure-based classifier. When the epitopes are present in the VDJDb
428 but not in the STCRDab (PDB) set (Case 2), both ERGO models significantly outperform
429 all other models in prediction, including ImRex. Finally, when peptides are present in
430 both train sets (Case 3), ERGO outperforms all models except the ones which include
431 Atchley factors information.

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



** → <0.01

*** → <0.001

Figure 8: *Caption next page*

Figure 8: (Previous page.) 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. Significance values are shown by asterisks.

435 **5 Discussion**

436 Previous study of the binding geometry of TCRs to the pMHC complex has been largely
437 focused on measuring the diagonal angle and the orientation of the TCR with respect to
438 the MHC. In the present study, a number of different features were extracted to try and
439 recapitulate both the conformation and the energetic profile of the binding interface. A
440 survey of the crystal structures available showed that, in agreement with Glanville et al.
441 2017; Ostmeyer et al. 2019, stretches of amino acids at the centre of the CDR3 in the
442 TCR α and β chains are within contact distance of the peptide. This information was
443 also recapitulated by the energy profiles, suggesting that not only can they interact, but
444 that they make favourable interactions. Although no conserved binding hotspots were
445 detected within the CDR, we were able to identify different binding modes simply from
446 the features extracted.

447 Conserved binding geometry has been reported in TCRs that bind the same MHC
448 complex (Blevins et al. 2016) and recently Singh et al. 2020 showed that a difference
449 can be detected between pMHC class I and class II binding. Such a difference is also
450 reported in this analysis, and detected both at the conformational level (in terms of
451 pairwise distances) and at the energetic level. As reported by Singh et al. 2020, our
452 analysis also showed that TCRs binding MHC class I tend to be closer to the C-terminus
453 of the peptide, whilst TCRs binding class II complexes sit more centrally or towards the

454 N-terminus. Moreover, the energetic features suggest that a difference between class I and
455 class II complexes can also be found in the energetic profiles that drive these interactions.
456 As well as the difference between class I and class II, the spatial features extracted from
457 the structures were readily able to distinguish TCRs which bind with reversed polarity
458 to the pMHC complex, as described by Gras et al. 2016 and Beringer et al. 2015, and
459 identify class I complexes with different non-canonical binding modes to the peptide (Yin
460 et al. 2011; Liu et al. 2013). This suggests that the features extracted are informative of
461 the biology of this system.

462 The information collected from these structures was also sufficient to build a clas-
463 sifier able to discriminate between TCR-pMHC binding from non-binding pairs. The
464 generalisability of the classifier was tested on multiple independent datasets, collected
465 and analysed independently. Physical interaction features on their own proved sufficient
466 to distinguish binding and non-binding complexes to a similar degree to published tools
467 which are based on sequence information alone (Figure 8). Interestingly, merging of
468 sequence and physical features in the same model did not improve the performance in
469 terms of ROC AUC, although often improved the recall of the sequence-based model.

470 This is an important characteristic, as in real-life applications a classifier like the one
471 presented could be used to screen candidate TCRs against an epitope of interest, for
472 example with the aim of identifying tumour-infiltrating lymphocytes that can recognise
473 tumour neoantigens. In this context, *in-silico* screening would be followed by experimen-
474 tal validation. Because the events of interest are a very small number compared to the
475 total number of events (i.e. binders << non-binders), it would be more important to
476 correctly classify more of the binders than of the non-binders, i.e. a higher number of
477 false positives, which can be screened out during experimental validation, would be less

478 problematic than a higher number of false negative, which would not be experimentally
479 validated.

480 Compared to other published classifiers (Glanville et al. 2017; Dash et al. 2017; Tong
481 et al. 2020), the classifier presented here is different in that it does not need to be trained
482 on a known subset of TCRs recognising a specific peptide to be able to predict more
483 binders, but rather it can learn from any set of TCR-pMHC pairs already available
484 and generalise what it has learnt to the problem at hand. This suggests that there
485 are conserved features to the TCR-pMHC interface which can be learnt and used for
486 prediction. ERGO and ImRex (Springer et al. 2020; Moris et al. 2020) have pioneered
487 this approach, although they only focussed on information that can be extracted from
488 the sequence. ImRex is a bit more similar to the classifier presented, as it encodes the
489 binding interface using amino acid characteristics rather than pure sequence encoding.

490 Of note, all of the results that we have presented here use the model originally trained
491 on the STCRDab set, which was never re-trained on the new sets of structures. This is
492 not the case for other published tools, which achieve better discrimination but only after
493 training on a section of the validation set.

494 We extended the approach adopted by ImRex and decided to rely on the structure
495 of the whole TCR-pMHC complex. Modelling of mutations within the existing crystal
496 structures has recently proved a successful approach to ranking candidate peptide epitopes
497 from a phage screen against target TCRs (Borrman et al. 2020). Here, we see from
498 the weights assigned to each combined kernel that the physical interactions encoded
499 by the distances and the attractive van der Waals forces were equally as important as
500 the sequence information, suggesting that physical interactions can be used to predict
501 binding. Moreover, the classifier here presented is trained on about 400 binding and

502 non-binding pairs, which recognise 93 different epitopes. This is a much smaller set
503 than the VDJdb used by ERGO and ImRex (40,000 TCRs and 200 peptides in ERGO
504 and 14,000 CDR3 β and 118 peptides in ImRex), but achieves similar performances. This
505 might indicate that the information learnt from the structural information is more readily
506 generalised to an unseen case.

507 As more structures for more diverse epitopes become available, the performance of
508 the classifier may well improve. However, the complex biology of the system will always
509 be a factor limiting performance. For example, if a small proportion of TCRs bound
510 to the pMHC complex with conformations that are significantly different from canonical
511 binding, we might never be able to predict their binding with a tool that has learnt on a
512 subset of canonical TCRs. This may well be the case with other structures with reversed
513 polarity or complexes with unusual binding highlighted in Figure 2a.

514 Most of the results presented has been based on a binary classification of TCR-pMHC
515 complexes as binding or non-binding. In reality, the interaction between TCR and pMHC
516 is characterised by a graded affinity scale. This is of interest as there are multiple metrics
517 that contribute to overall affinity and are important for T cell activation dynamics - K_D ,
518 k_{on} , k_{off} , half-life - (Gálvez et al. 2019; Lever et al. 2017; Stone et al. 2009) and it is
519 not yet clear what features in the structure can drive them. No correlation between the
520 classifier score and affinity or kinetic parameters was detected for the ATLAS structures
521 (Borrman et al. 2017). However, the original ATLAS publication showed a correlation
522 between the attractive van der Waal force as calculated by Rosetta (here atr) and the
523 experimentally-measured affinity, similar to the one reported by Erijman et al. 2014 on
524 an unrelated system. Because the affinity is driven by structure, we believe the PDB
525 classifier could also be optimised for rough affinity prediction, although better methods

526 of modelling the mutations into the structures might have to be explored.

527 Finally, the major difference between this classifier and most of the work published so
528 far is that it relies on an available TCR $\alpha\beta$ pairs and cannot be used on unpaired chains.
529 This is a limitation to the direct application of the classifier as alpha/beta pairing is
530 typically not available from bulk TCRseq data. However, unpaired α and β chains only
531 contain a portion of the binding site information, and the assumption that binding of the
532 β chain only is sufficient is clearly not true in every case. Carter et al. 2019 show that
533 the information encoded in the $\alpha\beta$ pair is synergistic, i.e. that the pairing carries more
534 than the sum of the individual chain information. Moreover, their survey of the VDJdb
535 shows instances where the same α chain paired with different β chains recognise different
536 epitopes, or where CDR3 α and β annotated to bind epitopes from different species come
537 together to bind yet another peptide. Overall, we believe this to be strong motivation to
538 work on $\alpha\beta$ pairs. Future work will focus on understanding whether candidate $\alpha\beta$ pairs
539 that bind a specific antigen can be inferred from TCR clones that are expanded during
540 an immune response.

541 6 Competing interests

542 The authors declare no competing interests.

543 References

544 10XGenomics. *A New Way of Exploring Immunity - Linking Highly Multiplexed Antigen*
545 *Recognition to Immune Repertoire and Phenotype.*

546 Aiolli, Fabio and Michele Donini (Dec. 2015). “EasyMKL: A scalable multiple kernel
547 learning algorithm”. In: *Neurocomputing* 169, pp. 215–224. ISSN: 18728286. DOI: 10.
548 1016/j.neucom.2014.11.078.

549 Alford, Rebecca F., Andrew Leaver-Fay, Jeliazko R. Jeliazkov, et al. (June 2017). “The
550 Rosetta All-Atom Energy Function for Macromolecular Modeling and Design”. In:
551 *Journal of Chemical Theory and Computation* 13.6, pp. 3031–3048. ISSN: 1549-9618.
552 DOI: 10.1021/acs.jctc.7b00125.

553 Atchley, William R., Jieping Zhao, Andrew D. Fernandes, and T. Druke (May 2005).
554 “Solving the protein sequence metric problem”. In: *Proceedings of the National Academy
555 of Sciences* 102.18, pp. 6395–6400. ISSN: 0027-8424. DOI: 10.1073/pnas.0408677102.

556 Bagaev, Dmitry V., Renske M.A. Vroomans, Jerome Samir, et al. (Jan. 2020). “VDJdb
557 in 2019: Database extension, new analysis infrastructure and a T-cell receptor motif
558 compendium”. In: *Nucleic Acids Research* 48.D1, pp. D1057–D1062. ISSN: 13624962.
559 DOI: 10.1093/nar/gkz874.

560 Beringer, Dennis X., Fleur S. Kleijwegt, Florian Wiede, et al. (Oct. 2015). “T cell receptor
561 reversed polarity recognition of a self-antigen major histocompatibility complex”. In:
562 *Nature Immunology* 16.11, pp. 1153–1161. ISSN: 15292916. DOI: 10.1038/ni.3271.

563 Blevins, Sydney J., Brian G. Pierce, Nishant K. Singh, et al. (Mar. 2016). “How structural
564 adaptability exists alongside HLA-A2 bias in the human $\alpha\beta$ TCR repertoire”. In:
565 *Proceedings of the National Academy of Sciences of the United States of America*
566 113.9, E1276–E1285. ISSN: 10916490. DOI: 10.1073/pnas.1522069113.

567 Borrman, Tyler, Jennifer Cimons, Michael Cosiano, et al. (2017). “ATLAS: A database
568 linking binding affinities with structures for wild-type and mutant TCR-pMHC com-

569 plexes”. In: *Proteins: Structure, Function and Bioinformatics* 85.5, pp. 908–916. ISSN:
570 10970134. DOI: 10.1002/prot.25260.

571 Borrman, Tyler, Brian G Pierce, Thom Vreven, Brian M Baker, and Zhiping Weng (Dec.
572 2020). “High-throughput modeling and scoring of TCR-pMHC complexes to predict
573 cross-reactive peptides”. In: *Bioinformatics*. Ed. by Arne Elofsson. ISSN: 1367-4803.
574 DOI: 10.1093/bioinformatics/btaa1050.

575 Britanova, Olga V., Ekaterina V. Putintseva, Mikhail Shugay, et al. (Mar. 2014). “Age-
576 Related Decrease in TCR Repertoire Diversity Measured with Deep and Normalized
577 Sequence Profiling”. In: *The Journal of Immunology* 192.6, pp. 2689–2698. ISSN: 0022-
578 1767. DOI: 10.4049/jimmunol.1302064.

579 Carter, Jason A., Jonathan B. Preall, Kristina Grigaityte, et al. (July 2019). “Single T
580 Cell Sequencing Demonstrates the Functional Role of $\alpha\beta$ TCR Pairing in Cell Lineage
581 and Antigen Specificity”. In: *Frontiers in Immunology* 10, p. 1516. ISSN: 1664-3224.
582 DOI: 10.3389/fimmu.2019.01516.

583 Chatterjee, Bithi, Yun Deng, Angelika Holler, et al. (May 2019). “CD8+ T cells retain pro-
584 tective functions despite sustained inhibitory receptor expression during Epstein-Barr
585 virus infection in vivo”. In: *PLOS Pathogens* 15.5. Ed. by Laurent Coscoy, e1007748.
586 ISSN: 1553-7374. DOI: 10.1371/journal.ppat.1007748.

587 Chaudhury, Sidhartha, Sergey Lyskov, and Jeffrey J. Gray (Jan. 2010). *PyRosetta: A
588 script-based interface for implementing molecular modeling algorithms using Rosetta*.
589 DOI: 10.1093/bioinformatics/btq007.

590 Cinelli, Mattia, Yuxin Sun, Katharine Best, et al. (Jan. 2017). “Feature selection using
591 a one dimensional naïve Bayes’ classifier increases the accuracy of support vector

592 machine classification of CDR3 repertoires". In: *Bioinformatics* 33.7, btw771. ISSN:
593 1367-4803. DOI: 10.1093/bioinformatics/btw771.

594 Coles, Charlotte H., Rachel M Mulvaney, Sunir Malla, et al. (Apr. 2020). "TCRs with
595 Distinct Specificity Profiles Use Different Binding Modes to Engage an Identical Pep-
596 tide–HLA Complex". In: *The Journal of Immunology* 204.7, pp. 1943–1953. ISSN:
597 0022-1767. DOI: 10.4049/jimmunol.1900915.

598 Dash, Pradyot, Andrew J. Fiore-Gartland, Tomer Hertz, et al. (July 2017). "Quantifi-
599 able predictive features define epitope-specific T cell receptor repertoires". In: *Nature*
600 547.7661, pp. 89–93. ISSN: 0028-0836. DOI: 10.1038/nature22383.

601 Dunbar, James and Charlotte M. Deane (Jan. 2016). "ANARCI: Antigen receptor num-
602 bering and receptor classification". In: *Bioinformatics* 32.2, pp. 298–300. ISSN: 14602059.
603 DOI: 10.1093/bioinformatics/btv552.

604 Erijman, Ariel, Eran Rosenthal, and Julia M. Shifman (Oct. 2014). "How Structure De-
605 fines Affinity in Protein-Protein Interactions". In: *PLoS ONE* 9.10. Ed. by Bostjan
606 Kobe, e110085. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0110085.

607 Fiser, András and Andrej Šali (2003). "MODELLER: Generation and Refinement of
608 Homology-Based Protein Structure Models". In: *Methods in Enzymology* 374, pp. 461–
609 491. ISSN: 00766879. DOI: 10.1016/S0076-6879(03)74020-8.

610 Gálvez, Jesús, Juan J. Gálvez, and Pilar García-Peñaarrubia (Mar. 2019). "Is TCR/pMHC
611 Affinity a Good Estimate of the T-cell Response? An Answer Based on Predictions
612 From 12 Phenotypic Models". In: *Frontiers in Immunology* 10.MAR. ISSN: 1664-3224.
613 DOI: 10.3389/fimmu.2019.00349.

614 Garboczi, David N., Partho Ghosh, Ursula Utz, Qing R. Fan, William E. Biddison, and
615 Don C. Wiley (Nov. 1996). "Structure of the complex between human T-cell receptor,

616 viral peptide and HLA-A2". In: *Nature* 384.6605, pp. 134–141. ISSN: 00280836. DOI:
617 10.1038/384134a0.

618 Garcia, K. Christopher, Jarrett J. Adams, Dan Feng, and Lauren K. Ely (Feb. 2009).
619 "The molecular basis of TCR germline bias for MHC is surprisingly simple". In:
620 *Nature Immunology* 10.2, pp. 143–147. ISSN: 1529-2908. DOI: 10.1038/ni.f.219.

621 Glanville, Jacob, Huang Huang, Allison Nau, et al. (July 2017). "Identifying speci-
622 ficity groups in the T cell receptor repertoire". In: *Nature* 547.7661, pp. 94–98. ISSN:
623 14764687. DOI: 10.1038/nature22976.

624 Gras, Stephanie, Jesseka Chadderton, Claudia M. Del Campo, et al. (Oct. 2016). "Re-
625 versed T Cell Receptor Docking on a Major Histocompatibility Class I Complex Lim-
626 its Involvement in the Immune Response". In: *Immunity* 45.4, pp. 749–760. ISSN:
627 10974180. DOI: 10.1016/j.immuni.2016.09.007.

628 Greef, Peter C. de, Theres Oakes, Bram Gerritsen, et al. (Mar. 2020). "The naive T-cell
629 receptor repertoire has an extremely broad distribution of clone sizes". In: *eLife* 9.
630 ISSN: 2050-084X. DOI: 10.7554/eLife.49900.

631 Hamelryck, Thomas and Bernard Manderick (Nov. 2003). "PDB file parser and struc-
632 ture class implemented in Python". In: *Bioinformatics* 19.17, pp. 2308–2310. ISSN:
633 13674803. DOI: 10.1093/bioinformatics/btg299.

634 Jensen, Kamilla Kjaergaard, Vasileios Rantos, Christine Jappe, et al. (2019). "TCRpMHC-
635 models: Structural modelling of tcR-pMHC class i complexes". In: *Scientific Reports*
636 9. DOI: 10.1038/s41598-019-50932-4.

637 Joshi, Kroopa, Marc Robert de Massy, Mazlina Ismail, et al. (Oct. 2019). "Spatial hetero-
638 geneity of the T cell receptor repertoire reflects the mutational landscape in lung can-

639 cer". In: *Nature Medicine* 25.1549, p. 1559. ISSN: 1078-8956. DOI: 10.1038/s41591-
640 019-0592-2.

641 Kjer-Nielsen, Lars, Craig S. Clements, Anthony W. Purcell, et al. (Jan. 2003). "A struc-
642 tural basis for the selection of Dominant $\alpha\beta$ T cell receptors in antiviral immunity". In:
643 *Immunity* 18.1, pp. 53–64. ISSN: 10747613. DOI: 10.1016/S1074-7613(02)00513-7.

644 Klausen, Michael Schantz, Mads Valdemar Anderson, Martin Closter Jespersen, Morten
645 Nielsen, and Paolo Marcatili (2015). "LYRA, a webserver for lymphocyte receptor
646 structural modeling". In: *Nucleic Acids Research* 43, pp. 349–355. DOI: 10.1093/
647 nar/gkv535.

648 Lauriola, Ivano and Fabio Aiolli (July 2020). "MKLpy: a python-based framework for
649 Multiple Kernel Learning". In: *arXiv*.

650 Lauriola, Ivano, Michele Donini, and Fabio Aiolli (2017). "Learning dot product polyno-
651 mials for multiclass problems". In: *ESANN 2017 - Proceedings, 25th European Sympo-*
652 *sium on Artificial Neural Networks, Computational Intelligence and Machine Learning*
653 May, pp. 23–28.

654 Leem, Jinwoo, Saulo H P de Oliveira, Konrad Krawczyk, and Charlotte M Deane (Jan.
655 2018). "STCRDab: the structural T-cell receptor database". In: *Nucleic Acids Re-*
656 *search* 46.D1, pp. D406–D412. ISSN: 0305-1048. DOI: 10.1093/nar/gkx971.

657 Lefranc, Marie Paule (1997). *Unique database numbering system for immunogenetic anal-*
658 *ysis*. DOI: 10.1016/S0167-5699(97)01163-8.

659 Lever, Melissa, Hong-sheng Lim, Philipp Kruger, et al. (Jan. 2017). "Correction for Lever
660 et al., Architecture of a minimal signaling pathway explains the T-cell response to
661 a 1 million-fold variation in antigen affinity and dose". In: *Proceedings of the Na-*

662 *tional Academy of Sciences* 114.2, E267–E267. ISSN: 0027-8424. DOI: 10.1073/pnas.

663 1620047114.

664 Liu, Yu Chih, John J. Miles, Michelle A. Neller, et al. (May 2013). “Highly Divergent T-
665 cell Receptor Binding Modes Underlie Specific Recognition of a Bulged Viral Peptide
666 bound to a Human Leukocyte Antigen Class I Molecule”. In: *Journal of Biological*
667 *Chemistry* 288.22, pp. 15442–15454. ISSN: 00219258. DOI: 10.1074/jbc.M112.447185.

668 Marcou, Quentin, Thierry Mora, and Aleksandra M. Walczak (2018). “High-throughput
669 immune repertoire analysis with IGoR”. In: *Nature Communications* 9.1. ISSN: 20411723.
670 DOI: 10.1038/s41467-018-02832-w.

671 McGranahan, Nicholas, A. J. S. Furness, Rachel Rosenthal, et al. (Mar. 2016). “Clonal
672 neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint block-
673 ade”. In: *Science* 351.6280, pp. 1463–1469. ISSN: 0036-8075. DOI: 10.1126/science.
674 aaf1490.

675 Moris, Pieter, Joey De Pauw, Anna Postovskaya, et al. (Dec. 2020). “Current challenges
676 for unseen-epitope TCR interaction prediction and a new perspective derived from
677 image classification”. In: *Briefings in Bioinformatics* 2020.0, pp. 1–12. ISSN: 1467-
678 5463. DOI: 10.1093/bib/bbaa318.

679 Ostmeyer, Jared, Scott Christley, Inimary T Toby, and Lindsay G Cowell (2019). “Bio-
680 physicochemical Motifs in T-cell Receptor Sequences Distinguish Repertoires from
681 Tumor-Infiltrating Lymphocyte and Adjacent Healthy Tissue”. In: *Cancer Research*
682 79.7. DOI: 10.1158/0008-5472.CAN-18-2292.

683 Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, et al. (2011). *Scikit-learn: Ma-*
684 *chine Learning in Python*. Tech. rep. 85, pp. 2825–2830.

685 Petersen, Jan, Laura Ciacchi, Mai T. Tran, et al. (Jan. 2020). “T cell receptor cross-
686 reactivity between gliadin and bacterial peptides in celiac disease”. In: *Nature Structural & Molecular Biology* 27.1, pp. 49–61. ISSN: 1545-9993. DOI: 10.1038/s41594-
687 019-0353-4.

688

689 Pogorelyy, Mikhail V., Anastasia A. Minervina, Dmitriy M. Chudakov, et al. (Mar. 2018).
690 “Method for identification of condition-associated public antigen receptor sequences”.
691 In: *eLife* 7. ISSN: 2050084X. DOI: 10.7554/eLife.33050.

692 Pogorelyy, Mikhail V., Anastasia A. Minervina, Mikhail Shugay, et al. (June 2019). “De-
693 tecting T cell receptors involved in immune responses from single repertoire snap-
694 shots”. In: *PLOS Biology* 17.6. Ed. by Thomas C. Freeman, e3000314. ISSN: 1545-7885.
695 DOI: 10.1371/journal.pbio.3000314.

696 Reinherz, Ellis L., Kemin Tan, Lei Tang, et al. (Dec. 1999). “The crystal structure of
697 a T cell receptor in complex with peptide and MHC class II”. In: *Science* 286.5446,
698 pp. 1913–1921. ISSN: 00368075. DOI: 10.1126/science.286.5446.1913.

699 Singh, Nishant K., Esam T. Abualrous, Cory M. Ayres, et al. (Mar. 2020). “Geometrical
700 characterization of T cell receptor binding modes reveals class-specific binding to
701 maximize access to antigen”. In: *Proteins: Structure, Function, and Bioinformatics*
702 88.3, pp. 503–513. ISSN: 0887-3585. DOI: 10.1002/prot.25829.

703 Springer, Ido, Hanan Besser, Nili Tickotsky-Moskovitz, Shirit Dvorkin, and Yoram Louzoun
704 (Aug. 2020). “Prediction of Specific TCR-Peptide Binding From Large Dictionaries of
705 TCR-Peptide Pairs”. In: *Frontiers in Immunology* 11, p. 1803. ISSN: 1664-3224. DOI:
706 10.3389/fimmu.2020.01803.

707 Stone, Jennifer D., Adam S. Chervin, and David M. Kranz (Feb. 2009). “T-cell receptor
708 binding affinities and kinetics: impact on T-cell activity and specificity”. In: *Immunol-
709 ogy* 126.2, pp. 165–176. ISSN: 00192805. DOI: 10.1111/j.1365-2567.2008.03015.x.

710 Thomas, Niclas, Katharine Best, Mattia Cinelli, et al. (Feb. 2014). “Tracking global
711 changes induced in the CD4 T-cell receptor repertoire by immunization with a complex
712 antigen using short stretches of CDR3 protein sequence”. In: *Bioinformatics* 30.22,
713 pp. 3181–3188. ISSN: 14602059. DOI: 10.1093/bioinformatics/btu523.

714 Thomas, Sharyn, Fiyaz Mohammed, Rogier M Reijmers, et al. (2019). “Framework engi-
715 neering to produce dominant T cell receptors with enhanced antigen-specific function”.
716 In: *Nature Communications* 10.1. ISSN: 20411723. DOI: 10.1038/s41467-019-12441-
717 w.

718 Thomas, Sharyn, Shao-An Xue, Charles R. M. Bangham, Bent K. Jakobsen, Emma C.
719 Morris, and Hans J. Stauss (July 2011). “Human T cells expressing affinity-matured
720 TCR display accelerated responses but fail to recognize low density of MHC-peptide
721 antigen”. In: *Blood* 118.2, pp. 319–329. ISSN: 0006-4971. DOI: 10.1182/blood-2010-
722 12-326736.

723 Tong, Yao, Jiayin Wang, Tian Zheng, et al. (June 2020). “SETE: Sequence-based En-
724 semble learning approach for TCR Epitope binding prediction”. In: *Computational
725 Biology and Chemistry*, p. 107281. ISSN: 14769271. DOI: 10.1016/j.combiolchem.
726 2020.107281.

727 Venturi, Vanessa, Katherine Kedzierska, David A. Price, et al. (Dec. 2006). “Sharing of
728 T cell receptors in antigen-specific responses is driven by convergent recombination”.
729 In: *Proceedings of the National Academy of Sciences of the United States of America*
730 103.49, pp. 18691–18696. ISSN: 00278424. DOI: 10.1073/pnas.0608907103.

731 Yang, Xinbo, Guobing Chen, Nan ping Weng, and Roy A. Mariuzza (Nov. 2017). “Struc-
732 tural basis for clonal diversity of the human T-cell response to a dominant influenza
733 virus epitope”. In: *Journal of Biological Chemistry* 292.45, pp. 18618–18627. ISSN:
734 1083351X. DOI: 10.1074/jbc.M117.810382.

735 Yin, Lei, Eric Huseby, James Scott-Browne, et al. (July 2011). “A single T cell receptor
736 bound to major histocompatibility complex class I and class II glycoproteins reveals
737 switchable TCR conformers”. In: *Immunity* 35.1, pp. 23–33. ISSN: 10747613. DOI:
738 10.1016/j.immuni.2011.04.017.

739 7 Supplementary Material

740 The following are supplied as supplementary materials:

741 1. Sequences for all the datasets used, specifically:

742 • **sequences from STCRDab PDB files** - these are the sequences from the
743 PDB files used for the initial feature extraction

744 • **STCRDab set metadata** - metadata associated with the sequences from the
745 STCRDab

746 • **10XGenomics set sequences** - sequences for the structures included in the
747 10X set

748 • **experimental constructs sequences** - sequences for the structures included
749 in the expt set

750 • **Dash set** - sequences for the structures included in the Dash set

751 • **ATLAS sequences** - sequences for the structures included in the TCR AT-
752 LAS set, including the affinity information from the ATLAS

753 • **VDJDb validation sequences** - sequences for the structures included in the
754 new VDJDb set

755 2. All result files with decision function scores for each TCR-peptide pair. A README
756 file is included with filename explanations.

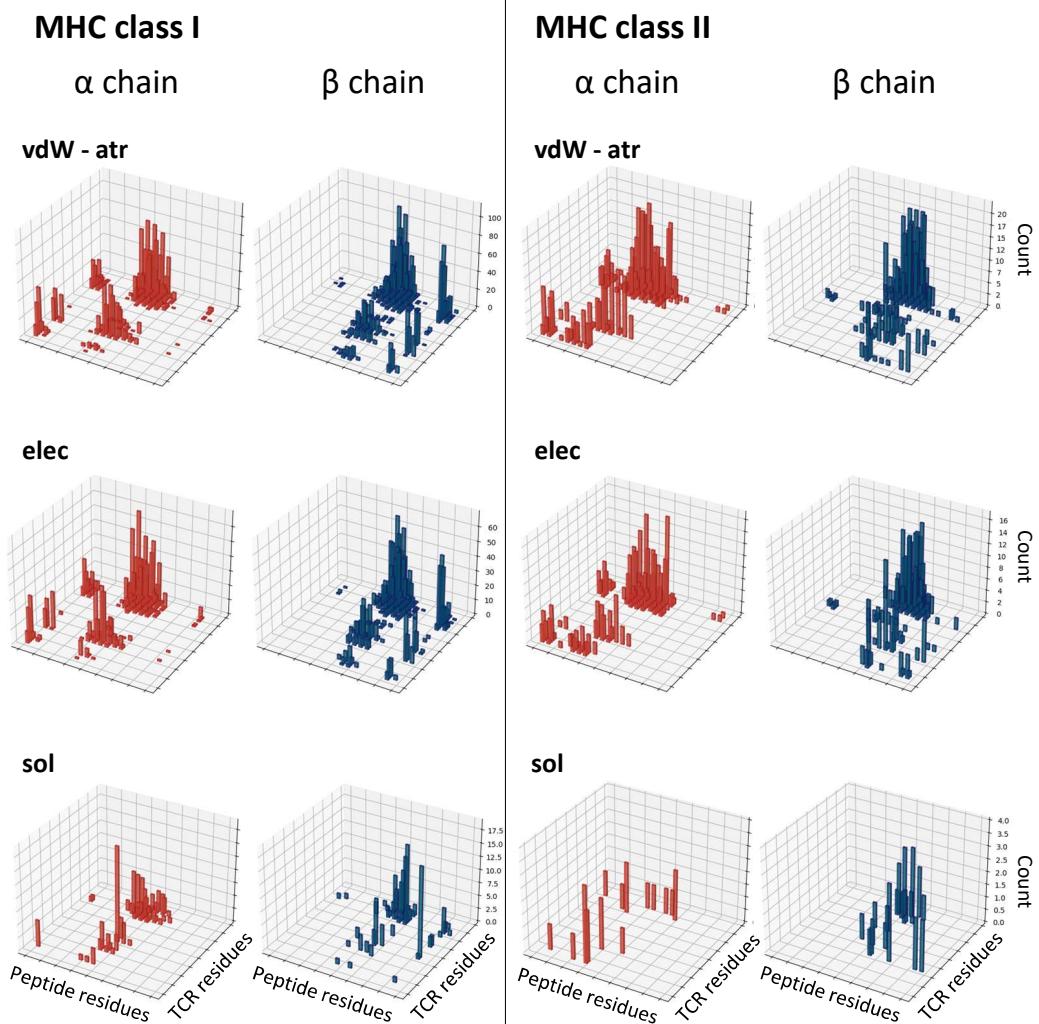


Figure S1: *Caption next page*

Figure S1: (*Previous page.*) **Energy interactions for class I and class II complexes**
Analogous to Figure 1c, but for all energy feature sets. The histograms show the number of structures that make a favourable contact (energy < 0). Repulsive vdW excluded as this component is always > 0 .

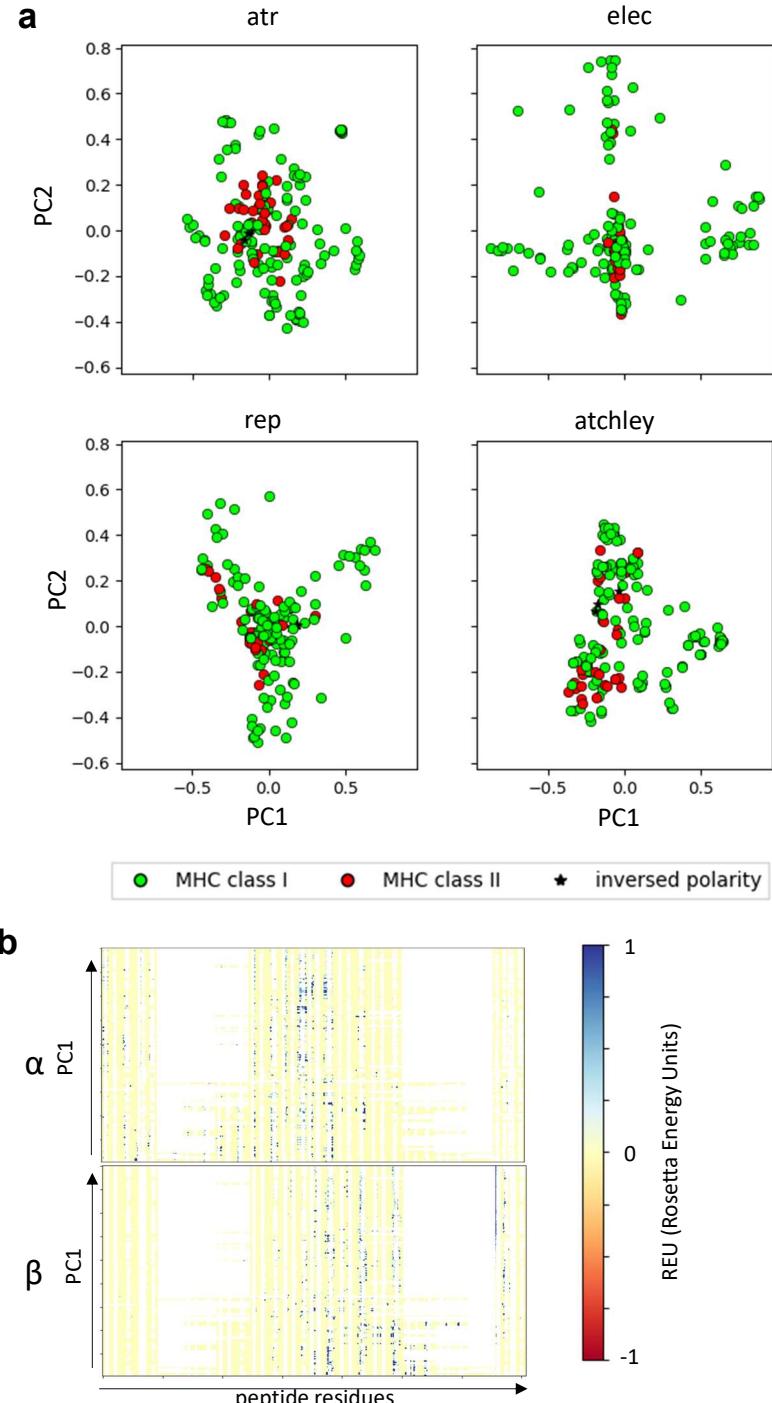


Figure S2: *Caption next page*

Figure S2: (*Previous page.*) **PCA on all extracted features.** **a.** PCA for feature sets not included in Figure 2a. Class I and class II complexes are shown in 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 degree angle). **b.** Linearised vectors used for the solvent energy PCA, ordered according to their PC1 score. On the x-axis, the calculated solvent energy between each CDR residue and each peptide residue (27-1, 28-1,...,116-1, 117-1, 27-2,...,117-20). Analogous to Figure 2b.

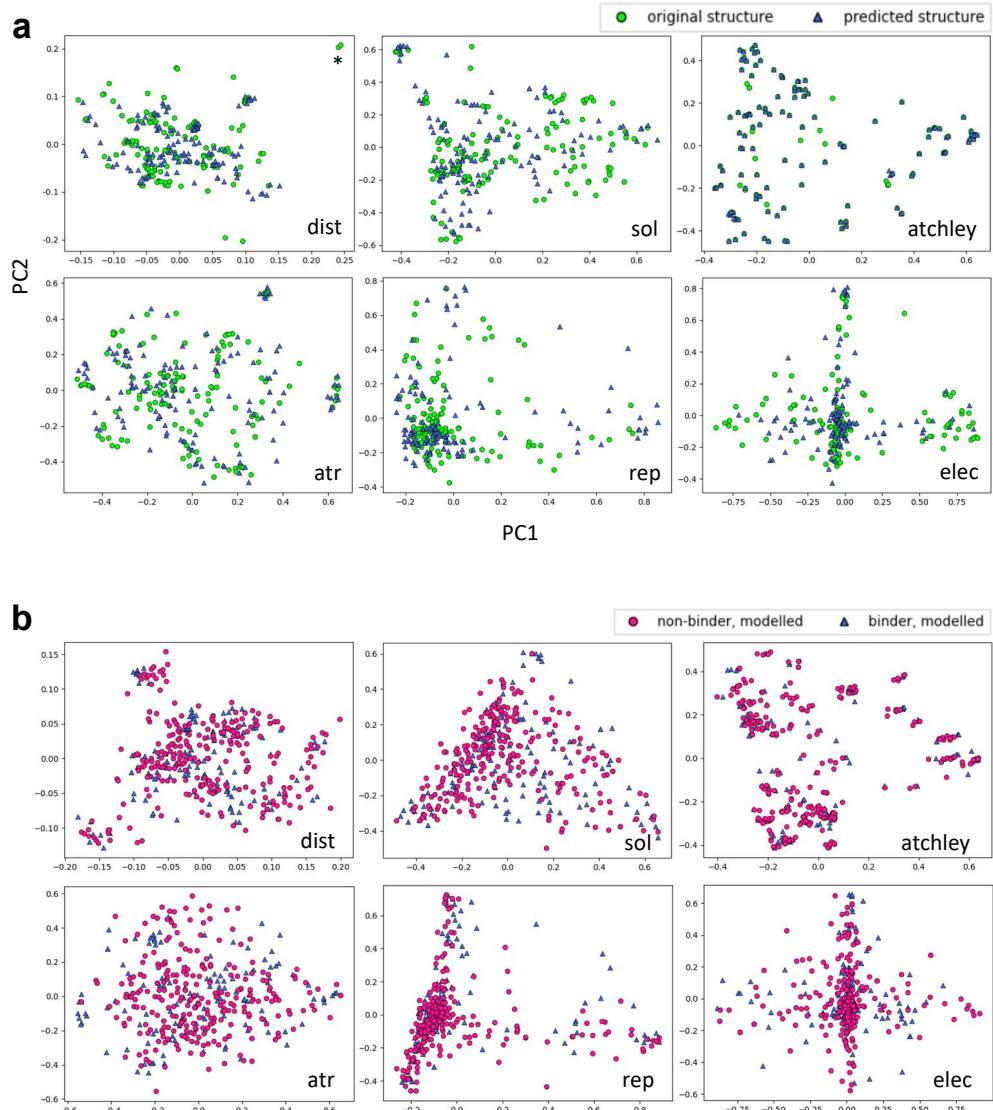


Figure S3: *Caption next page*

Figure S3: (*Previous page.*) **PCA of original vs predicted and of binding vs non-binding.** **a.** PCA for each set showing overlay between original and predicted structures. Asterisks (*) in the distance plot indicates the inversed polarity structures. **b.** PCA for each set showing overlay of binding and non-binding complexes (predicted structures, blue triangles and magenta circles, respectively).

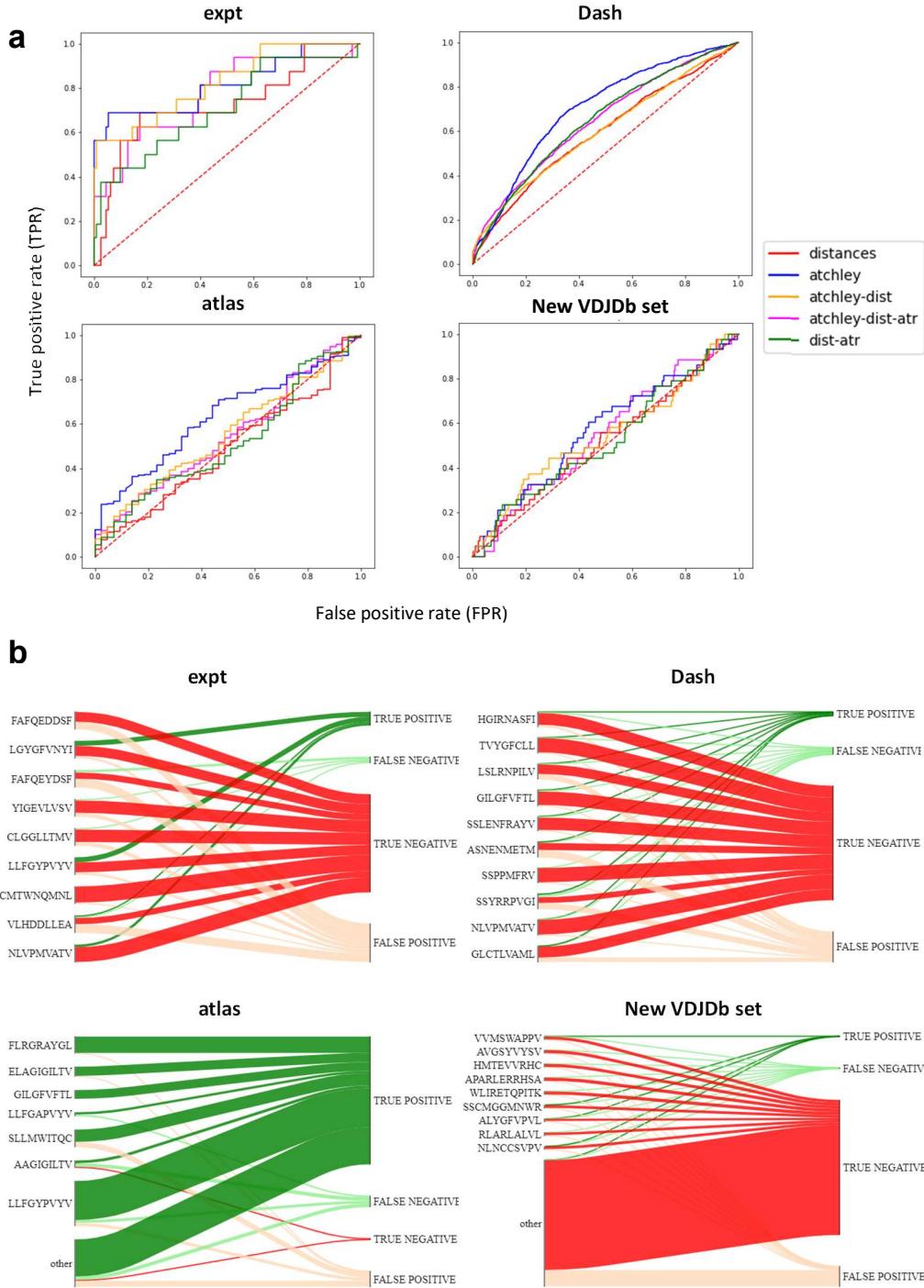


Figure S4: *Caption next page*

Figure S4: (*Previous page.*) **Results of all validation sets used.** **a.** ROC curves obtained when the model trained on the STCRDab set are used for prediction on each of the validation sets. **b.** For the model trained on STCRDab using distances 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) for each of the validation sets used.

	N pos	N neg	in_pdb	in_vdjdb	distances	dist-atr	atchley	atchley-dist	atchley-dist-atr	ImRex	ERGO LSTM	ERGO AE
VVMSWAPPV	7	120	no	no	0.361	0.526	0.605	0.370	0.557	0.482	0.461	0.433
ALYGFVPLV	5	122	no	no	0.620	0.603	0.508	0.556	0.597	0.474	0.290	0.657
HMTEVVRHC	4	123	no	no	0.390	0.551	0.654	0.549	0.573	0.679	0.551	0.498
APARLERRHSA	3	124	no	no	0.559	0.570	0.449	0.538	0.495	0.901	0.591	0.562
RLARLALVL	5	122	no	no	0.218	0.285	0.443	0.215	0.259	0.433	0.575	0.582
NLNCCSVPV	4	123	no	no	0.715	0.638	0.447	0.720	0.667	0.547	0.677	0.567
RLRAEAQVK	57	336	no	yes	0.465	0.416	0.525	0.461	0.429	0.538	0.753	0.727
SSPPMFVR	20	1795	no	yes	0.393	0.412	0.396	0.373	0.424	0.665	0.891	0.814
MLDLQPETT	6	6	no	yes	0.750	0.333	0.306	0.417	0.250	0.778	0.694	0.583
FLASKIGRLV	3	24	no	yes	0.500	0.639	0.389	0.375	0.653	0.542	1.000	0.597
TVYGFCLL	46	1839	no	yes	0.407	0.419	0.323	0.288	0.386	0.453	0.915	0.757
KTWGQYWQV	3	10	no	yes	0.800	0.633	0.700	0.867	0.633	0.433	1.000	0.933
KLGGALQAK	324	2161	no	yes	0.493	0.479	0.527	0.498	0.493	0.511	0.739	0.630
AYAQKIFKI	4	62	no	yes	0.750	0.379	0.464	0.685	0.339	0.266	0.690	0.867
LLDFVRFMGV	10	18	no	yes	0.794	0.639	0.328	0.633	0.494	0.294	0.767	0.800
HGIRNASFI	140	1674	no	yes	0.498	0.652	0.500	0.482	0.608	0.610	0.926	0.918
LSLRNPILV	64	1796	no	yes	0.437	0.443	0.644	0.456	0.465	0.520	0.902	0.745
IVTDFSVIK	207	421	no	yes	0.540	0.613	0.662	0.632	0.649	0.668	0.821	0.795
RMFPNAPYL	4	12	no	yes	0.542	0.542	0.604	0.625	0.667	0.542	0.958	0.750
SSYRRPVGI	455	1389	no	yes	0.432	0.471	0.561	0.499	0.466	0.282	0.938	0.927
AVFDRKSDAK	175	869	no	yes	0.465	0.441	0.494	0.460	0.432	0.534	0.716	0.669
SLFNTVATLY	5	34	no	yes	0.241	0.435	0.300	0.353	0.506	0.435	0.771	0.712
RAKFKQLL	77	169	no	yes	0.635	0.511	0.594	0.637	0.511	0.554	0.725	0.726
FLYALALLL	7	9	no	yes	0.508	0.635	0.190	0.444	0.349	0.349	1.000	0.968
LGYGFVNYYI	4	10	yes	yes	0.925	0.850	1.000	1.000	0.950	0.925	1.000	0.925
GLCTLVAML	98	1848	yes	yes	0.722	0.717	0.747	0.740	0.737	0.756	0.991	0.980
LLFGYPVVV	91	36	yes	yes	0.865	0.867	0.888	0.888	0.876	0.908	0.935	0.922
SLLMWITQC	33	11	yes	yes	0.355	0.088	0.598	0.438	0.176	0.665	0.638	0.806
SSLENFRAYV	147	1614	yes	yes	0.542	0.586	0.563	0.523	0.543	0.630	0.836	0.730
GILGFVFTL	534	2028	yes	yes	0.722	0.741	0.841	0.779	0.785	0.822	0.982	0.969
ELAGIGILTV	178	348	yes	yes	0.736	0.726	0.825	0.778	0.747	0.574	0.862	0.754
ASNENMETM	161	1717	yes	yes	0.518	0.609	0.468	0.461	0.608	0.486	0.948	0.900
NLVPMVATV	63	1876	yes	yes	0.623	0.648	0.558	0.628	0.626	0.495	0.987	0.956

Table S1: *Caption next page*

Table S1: (*Previous page.*) **Results of benchmarking on single epitopes.** For each epitope, the performance of each tool is calculated (ROC AUC). In each row, the best-performing tool is highlighted in bold and the best-performing model of the ones presented in this paper is boxed.