nature machine intelligence

Article

https://doi.org/10.1038/s42256-023-00634-4

Characterizing the interaction conformation between T-cell receptors and epitopes with deep learning

Received: 29 March 2022

Accepted: 27 February 2023

Published online: 27 March 2023

Check for updates

Xingang Peng[1,2,6], Yipin Lei[3,6], Peiyuan Feng[3], Lemei Jia[4], Jianzhu Ma[5], Dan Zhao[3] ✉ & Jianyang Zeng[3] ✉

Computational modelling of the interactions between T-cell receptors (TCRs) and epitopes is of great importance for immunotherapy and antigen discovery. However, current TCR–epitope interaction prediction tools are still in a relatively primitive stage and have limited capacity in deciphering the underlying binding mechanisms, for example, characterizing the pairwise residue interactions between TCRs and epitopes. Here we designed a new deep-learning-based framework for modelling TCR–epitope interactions, called TCR–Epitope Interaction Modelling at Residue Level (TEIM-Res), which took the sequences of TCRs and epitopes as input and predicted both pairwise residue distances and contact sites involved in the interactions. To tackle the current bottleneck of data deficiency, we applied a few-shot learning strategy by incorporating sequence-level binding information into residue-level interaction prediction. The validation experiments and analyses indicated its good prediction performance and the effectiveness of its design. We demonstrated three potential applications: revealing the subtle conformation changes of mutant TCR–epitope pairs, uncovering the key contacts based on epitope-specific TCR pools, and mining the intrinsic binding rules and patterns. In summary, our model can serve as a useful tool for comprehensively characterizing TCR–epitope interactions and understanding the molecular basis of binding mechanisms.

The recognition of epitopes by T cells plays a vital role in adaptive immune response. Pathogenic peptides presented by major histocompatibility complex (MHC) molecules are identified by T-cell receptors (TCRs) to stimulate cell-mediated immunity[1–3], thus eliminating the infected cells and activating the corresponding immune cells. Therefore, understanding the binding mechanisms of TCR–peptide–MHC complexes (TCR-pMHCs) is of great significance for cancer immunology, autoimmunity antigen discovery and vaccine design[2,4,5]. However, due to the intrinsic complexity of such recognition mechanisms, experimental detection and determination of TCR-pMHC interactions are often time-consuming and expensive[6]. To alleviate these problems, numerous computational methods have been developed to model the TCR-pMHC interactions.

For adaptive immune recognition, peptides (also referred to as epitopes in this context) are presented by MHC molecules on cell surfaces and then recognized by TCRs. Although a TCR binds to an epitope and the corresponding MHC molecule partner simultaneously, the core binding regions of the complex are between the complementarity

[1]School of Intelligence Science and Technology, Peking University, Beijing, China. [2]Institute for Artificial Intelligence, Peking University, Beijing, China. [3]Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China. [4]Institute for Immunology and School of Medicine, Tsinghua University, Beijing, China. [5]Institute for AI Industry Research, Tsinghua University, Beijing, China. [6]These authors contributed equally: Xingang Peng and Yipin Lei. ✉e-mail: zhaodan2018@tsinghua.edu.cn; zengjy321@tsinghua.edu.cn

https://doi.org/10.1038/s42256-023-00634-4

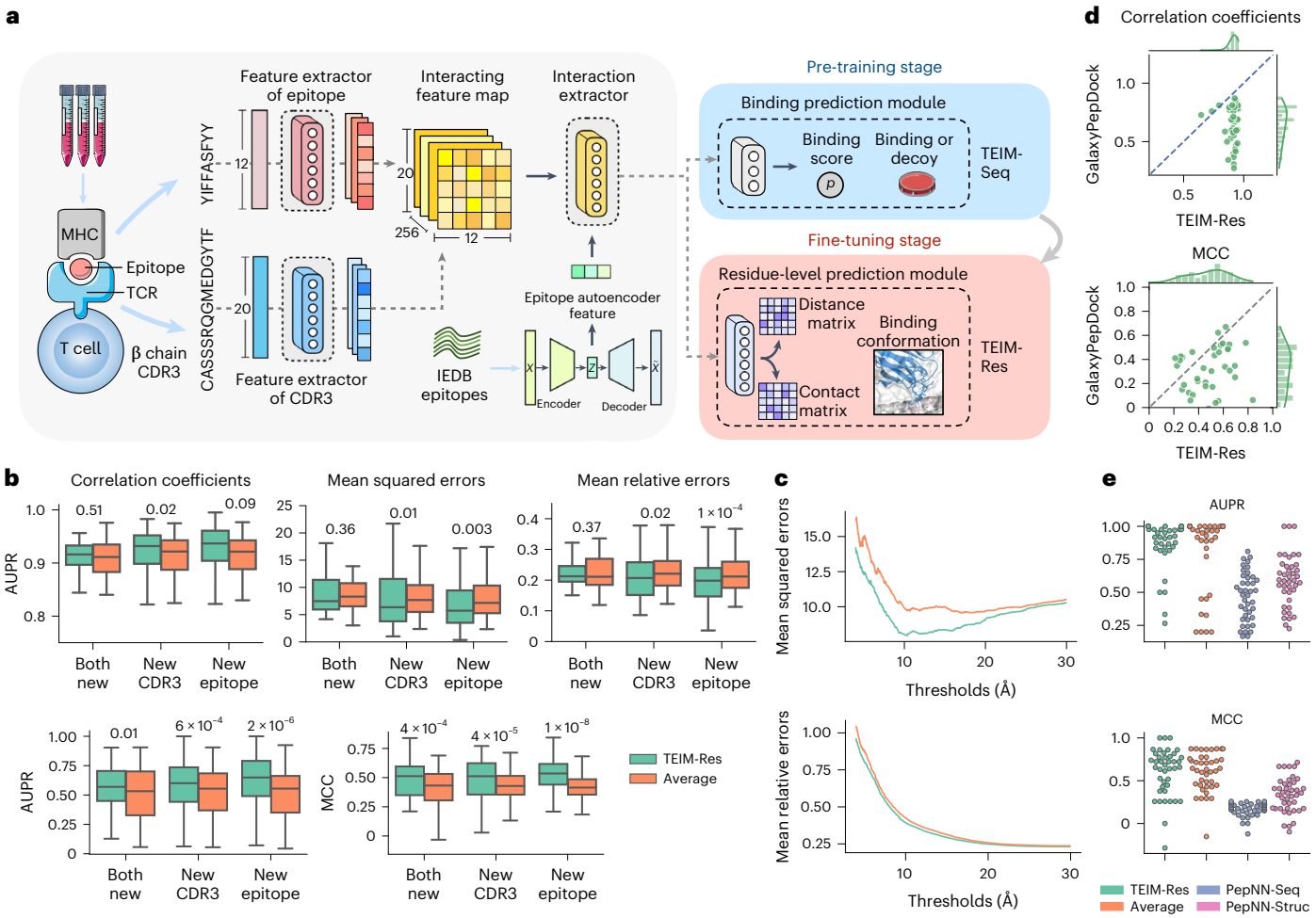


Fig. 1 | Model architectures and performance evaluation. a, The architectures of TEIM-Seq and TEIM-Res and the training pipeline. Both TEIM-Seq and TEIM-Res models share similar architectures except for the last modules. They both have feature extractors to learn sequence features of CDR3s and epitopes separately and then extend them to different dimensions to form an interaction feature map. Next, an interaction extractor, mainly consisting of 2D CNNs, is used to extract the pairwise residue interaction information. Alongside, an epitope feature vector generated by an autoencoder is fed into an interaction extractor for global epitope information. Finally, TEIM-Seq uses a binding prediction module to aggregate all pairwise interactions to predict the binding score (that is, binding probability) while TEIM-Res uses a residue-level prediction module consisting of 2D CNN layers to predict the distance matrix and the contact matrix. The training pipeline includes two stages: pre-training TEIM-Seq on the sequence-level binding data and then fine-tuning TEIM-Res on the residue-level binding data. **b**, Performance of TEIM-Res and the average baseline for residue-level interaction prediction under three different data splitting settings. One-sided paired *t*-tests were conducted and the resulting *P* values are annotated at the top of the corresponding metrics. The sample sizes for the three splittings were 42, 122 and 122, respectively. The box plots show the medians as centre lines, interquartiles as hinges and 1.5 times the interquartile ranges as whiskers (outliers are not shown). **c**, The mean squared/relative errors of residue pairs within different distance thresholds for TEIM-Res and the average baseline. **d**, Performance comparison between GalaxyPepDock and TEIM-Res under the both-new splitting setting (the mean squared/relative errors are shown in Extended Data Fig. 1, and the comparison between GalaxyPepDock and the average baseline is also shown there). **e**, Performance comparison among TEIM-Res, average baseline and PepNN for contact prediction under the both-new splitting setting.

determining region 3 of the TCR β chain (CDR3β) and the epitope[1,6,7]. There exist many computational tools to predict the binding of a CDR3β–epitope pair, for example, pMTnet[8], TCRex[9], NetTCR[10], ERGO[11] and ImRex[12]. These tools exploit different machine learning methods to predict whether a given CDR3–epitope sequence pair can bind to each other, which we refer to as the sequence-level binding classification task. However, they cannot reveal the detailed underlying mechanisms of interactions between TCRs and epitopes. Predicting the binding distances between residues and the contact residue pairs from TCRs and epitopes, which we refer to as the residue-level binding prediction task, has not been fully explored yet.

There exist several related works that predict the structure of a TCR–epitope–MHC complex and thus can be used to analyse their residue-level interactions[13,14]. Unfortunately, most of them require the availability of homologous sequences for accurate homology modelling, which is generally hard for CDRs, the most important regions of TCRs[15,16]. Although an energy-based model called RACER that can learn contact maps from TCR–epitope binding data has recently been designed[17], it requires numerous binding epitopes with the querying TCRs for training and cannot be directly used for arbitrary TCRs.

Because TCR–epitope interactions can be regarded as a special case of protein–peptide interactions, in principle it is possible to utilize existing tools designed for modelling protein–peptide interactions to characterize residue-level TCR–epitope interactions. These include protein–peptide docking tools[18–21] and machine learning-based models for protein–peptide contact prediction[22–25]. However, their performance on TCR–epitope prediction has never been evaluated. Additionally, TCR–epitope interactions possess several unique and distinct binding patterns compared with general protein–peptide interactions and hence require specific modelling methods.

https://doi.org/10.1038/s42256-023-00634-4

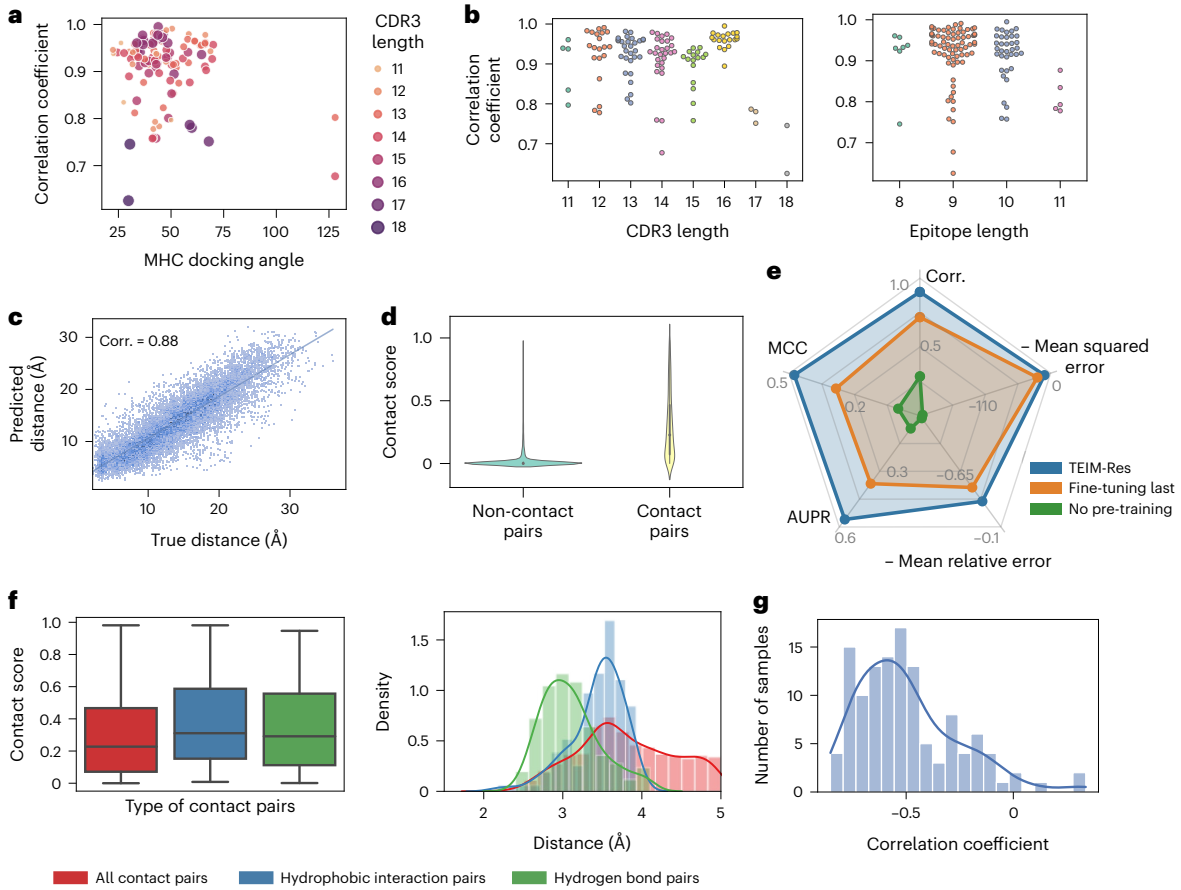


Fig. 2 | Detailed analyses of model performance. a, The relationship between the distance prediction accuracy and properties of samples such as MHC docking angles and CDR3β lengths, where the accuracy was represented by the correlation coefficient between true and predicted distances. **b**, The distributions of the correlation coefficients for different sequence lengths. **c**, The correlations between true and predicted distances of all residue pairs. The overall correlation coefficient was 0.88. **d**, The distributions of predicted contact scores for true contact pairs (n = 1,098) or non-contact pairs (n = 14,700). The contact pairs and non-contact pairs had totally different contact score distributions. The boxes within the violin plots show the medians as centres, interquartiles as hinges and 1.5 times the interquartile ranges as whiskers. **e**, Comparison of the original TEIM-Res training pipeline, the 'fine-tuning last' pipeline, in which the pre-training stage remained unchanged but only the last module of TEIM-Res (that is, the residue-level prediction module) was fine-tuned, and 'no pre-training' pipeline, in which we skipped the pre-training stage, on five metrics (that is, correlation coefficients, negative mean squared errors and negative mean relative errors for distance prediction, and AUPR and MCC for contact prediction). The radar plot shows the average values of different pipelines and metrics. The original TEIM-Res model outperformed the other two on all metrics, and fine-tuning last outperformed no pre-training on all metrics. **f**, The distributions of predicted contact scores and distances for residue pairs involved in different types of non-covalent bonds. The pairs of hydrophobic interactions and hydrogen bonds had relatively higher contact scores and closer pairwise distances than the other contact pairs. The sample sizes were 1098, 114 and 146, respectively. The box plots show the medians as centre lines, interquartiles as hinges and 1.5 times the interquartile ranges as whiskers. **g**, The distribution of correlation coefficients between class activation map values that were calculated from the pre-trained model and the true residue distances.

To better understand TCR–epitope interactions, we proposed a deep learning-based framework, called TCR–Epitope Interaction Modelling at Residue Level (TEIM-Res), to characterize the interaction conformation between $CD8^+$ (positive for cluster of differentiation 8) TCRs and MHC-I presented epitopes. More specifically, based on the sequences of a binding TCR–epitope pair, we predicted the distances and contact probabilities of all pairwise residues between the CDR3β and the epitope. Due to the scarcity of the corresponding structural data, we employed a few-shot learning strategy to exploit prior knowledge[26,27]. More specifically, before training a model on residue-level binding data, we pre-trained it on the rich high-throughput sequence-level binary binding data to equip our model with relevant knowledge of TCR–epitope interactions, which was inspired by the fact that the overall bindings between TCRs and epitopes are actually determined by the interactions of all residue pairs between them. With such sequence-level binding data, our model is able to implicitly capture residue-level information during the pre-training stage. Comprehensive validation tests and analyses showed that our model outperformed other existing methods and further demonstrated its effectiveness. We also showed that our model can be successfully applied to mutation analysis, epitope-specific TCR analysis, and binding pattern discovery, indicating its effective usage as a valuable tool for TCR–epitope interaction-related analyses. Finally, we revealed that the by-product model obtained during our pre-training stage, called TEIM-Seq (TEIM at Sequence Level), also displayed great application potential for sequence-level binding analyses.

Results

TEIM-Res successfully recognizes residue-level interactions

We proposed a deep learning-based model, called TEIM-Res, which takes the sequences of a CDR3β and an epitope as input and outputs a distance matrix and a contact matrix that represent the corresponding distances and contact probabilities of all residue pairs from the CDR3β and the epitope. Due to the limited availability of TCR–epitope data with residue-level interaction labels, it is hard to directly predict residue-level TCR–epitope bindings through deep learning models. To

https://doi.org/10.1038/s42256-023-00634-4

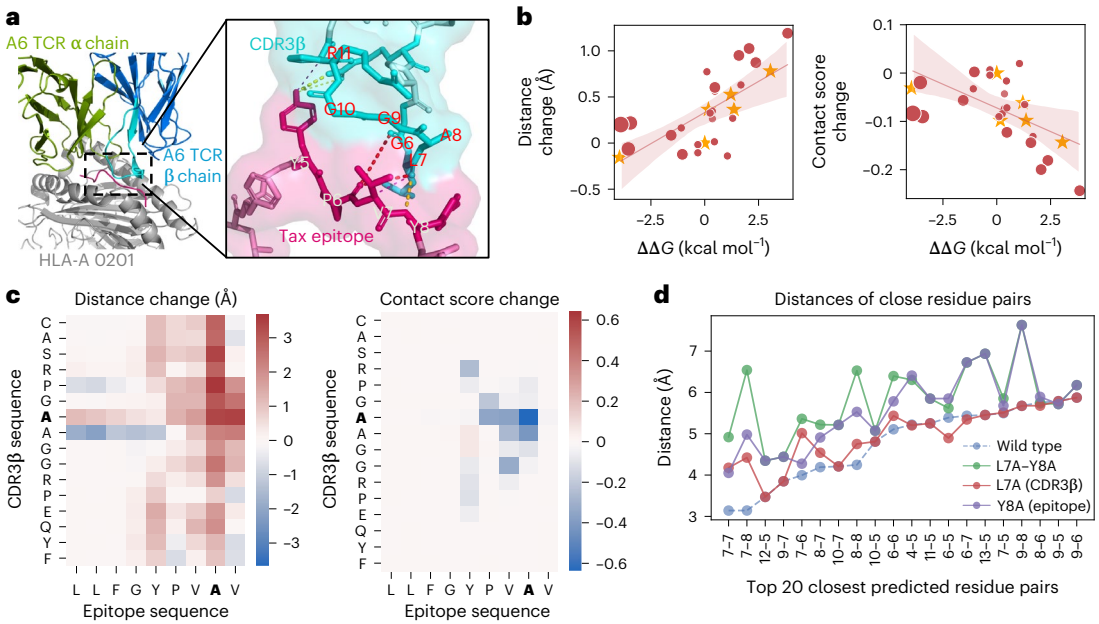


Fig. 3 | The performance of TEIM-Res on the mutation analysis on the interactions between A6 TCR and Tax epitope. a, The experimental complex structure of A6 TCR and Tax epitope presented by HLA-A:0201 and the detailed residue interactions between the CDR3β and the epitope (PDB: 1AO7). The α and β chains in the overall structure of A6 TCR are coloured in olive and dark blue, respectively. The CDR3β and the epitope in the detailed interaction conformation are coloured in cyan and pink, respectively. The MHC (HLA-A:0201) is coloured in grey and the non-covalent interactions between residues are plotted by PyMOL as dotted lines. **b**, The correlations between the average distance/contact changes of residue pairs predicted by TEIM-Res and the experimentally determined affinity changes (represented by $\Delta\Delta G$) of mutant samples. The bands represent the 95% confidence intervals for the linear fitting. The samples existing in the training set are marked as orange stars, and the remaining ones are shown as red circles. The sizes of circles represent the total edit distances between the mutant sequences of the TCR–epitope pair and the corresponding wild-type one. For the samples that were not in the training set, the correlations of distance changes and contact score changes with the affinity changes were −0.6513 and −0.5471, respectively. **c**, The distance and contact score changes of residue pairs of the mutant sample A6_L7A-Tax_Y8A predicted by TEIM-Res. The mutant residues are shown in bold. The majority of the mutant pairs showed positive distance changes and negative contact score changes, indicating that most residue pairs became more distant and had lower contact scores after mutation. **d**, The distances of the residue pairs with the top 20 closest distances in the wild-type sample before and after mutations. The *x* axis represents the position indices of the CDR3β–epitope residue pairs.

tackle such a data deficiency problem, we utilized a few-shot learning technology. In particular, before training TEIM-Res, we first build a sequence-level binary binding (that is, whether the input pair of CDR3β and epitope can bind to each other) prediction model, named TEIM-Seq, which leverages the large-scale sequence-level binding data to implicitly learn the residue-level information of TCR–epitope interactions. Then we fine-tune TEIM-Res on the limited residue-level binding data. TEIM-Seq and TEIM-Res frameworks share almost the same model structure except for the last modules (Fig. 1a, Methods). TEIM-Seq contains a binding prediction module that aggregates information to predict the probability that the TCR–epitope pair can bind to each other, while TEIM-Res has a residue-level prediction module to predict the distances and contact probabilities of residue pairs.

The core of our model is the interaction extractor, which utilizes two-dimensional convolutional neural networks (2D CNNs) to learn local interaction information between pairs of CDR3βs and epitopes. The biological intuition of the architecture is that the binding between a TCR and epitope pair is actually determined by the non-covalent interactions from the corresponding residue pairs. During the pre-training stage, although the dataset contains no residue-level information, the particularly designed interaction extractor can capture the underlying residue-level interactions from the sequence-level binding data. Therefore, the pre-training and fine-tuning pipeline enables the model to utilize the relatively abundant sequence-level interaction information to boost residue-level interaction prediction.

We applied a cross-validation strategy to evaluate the model performance under three different data splitting settings to avoid data redundancy: both-new splitting, new-CDR3 splitting and new-epitope splitting, which represented that the CDR3β–epitope pairs, the CDR3βs, and the epitopes in the validation set were not in the training set, respectively. We used the Pearson correlation coefficient per sample, the mean squared error per sample and the mean relative error per sample to evaluate the pairwise residue distance prediction results, and used the area under the precision-recall curve (AUPR) per sample and Matthews correlation coefficient (MCC) per sample to assess the contact site prediction results.

As the residue distances are probably biased by the residue positions along the sequences, we provided an average baseline that calculated the average of the matrices of training samples as the prediction for any validating sample. As shown in Fig. 1b, for distance prediction, TEIM-Res outperformed the average baseline for most metrics and splitting settings. For most samples, TEIM-Res achieved correlation coefficients greater than 0.9, mean squared error less than 10 and mean relative error less than 0.2. Regarding the average baseline, it was not surprising that it achieved similar performance on some metrics for the distance prediction because most residue pairs are far away from each other and these pairs actually have relatively weak interactions. Therefore, their distances are more easily speculated from their positions and less relevant to other features. Additionally, distant residue pairs are the majority and thus it was possible that the average baseline sometimes achieved similar values of these metrics with TEIM-Res. However, this did not mean that the baseline was as good as TEIM-Res. As shown in Fig. 1c, TEIM-Res showed smaller errors for the residue pairs within close distance thresholds. In real applications, we care more about the close residue pairs than those distant ones and therefore the prediction of the average baseline can provide less useful

https://doi.org/10.1038/s42256-023-00634-4

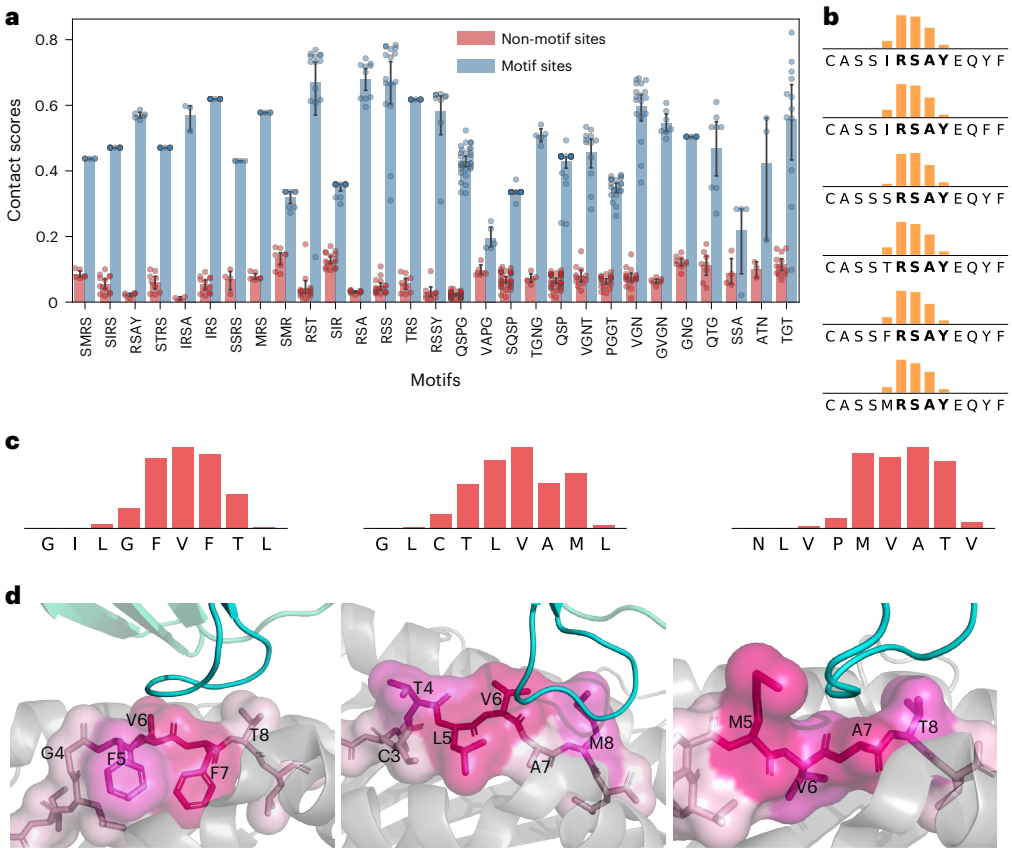


Fig. 4 | The analyses of three epitope-specific TCR pools. a, The contact scores of motif sites and 'non-motif' regions (shown as mean values with 95% confidence intervals, the sample sizes are provided in Supplementary Table 2). The motifs were identified using GLIPH[7]. TEIM-Res generated much higher contact scores for the motif sites than the non-motif regions, which validated that the motifs were highly correlated with the contact sites. **b**, The six CDR3β sequences that contained the motif RSAY in the dataset and the predicted contact scores for individual residues. The motif RSAY sites had relatively high contact scores on all these CDR3s. **c**, The average contact scores over all interacting CDR3βs in the TCR pools of the three epitopes. Higher bars indicate larger contact scores. **d**, The crystal structures of the three epitopes interacting with CDR3βs (PDB: 2VLJ, 3O4L and 3GSN). The epitope chains are coloured in dark pink; the CDR3β chains are in cyan; the neighbouring regions of the MHC are in grey. Other views of the binding conformations are shown in Extended Data Fig. 4. The prediction perfectly matched the true structures for the three epitopes with distinct binding conformations. Specifically, the first epitope GILGFVFTL presented a relatively flat surface at the interface, where the crucial contact sites of the epitope were V6 and F7 and the contact scores of the remaining residues decreased as their positions were far away from the crucial sites. The second epitope GLCTLVAML had two crucial contact sites, that is, L5 and V6, and the remaining residues had generally lower contact scores as their positions were far from the crucial sites. However, we noticed that the A7 site had relatively lower scores, which can be explained by the fact that the A7 residue has a much smaller side-chain group than its neighbour residues and is buried under the surface. The third epitope NLVPMVATV formed a groove at the interface to hold the CDR3β loop and thus the residues outside the groove had much lower contact scores than those forming the groove.

information while TEIM-Res still provides more accurate prediction for these important residue pairs. In the contact prediction task, TEIM-Res obviously outperformed the average baseline. For contact prediction, models have to focus on the close residue pairs to figure out whether they can interact and contact, in which the positional information is not enough and more features should be learned by the model. Therefore, TEIM-Res can make more biologically meaningful predictions for the residue-level prediction than the average baseline that only uses the positional information.

Next, we benchmarked TEIM-Res against several alternative previous methods. Firstly, we compared TEIM-Res with a template-based TCR–epitope molecular modelling tool called TCRpMHCmodels[13], which searches homologous sequences as structure templates and then uses molecule energy functions to construct the complex structures. We found that TEIM-Res achieved similar performance with TCRpMHCmodels (correlation coefficient, mean squared error and mean relative error were 0.902, 10.209 and 0.234 for TEIM-Res, respectively; and 0.909, 11.112 and 0.242 for TCRpMHCmodels, respectively). However, a number of samples in the validation set cannot be predicted by TCRpMHCmodels because TCRpMHCmodels relies on template structures to model the complexes and refuses to make predictions for those samples without templates in the training set. TEIM-Res does not suffer from this limitation and still achieved good performance for the samples that TCRpMHCmodels failed to predict, with average correlation coefficient of 0.896, mean squared error of 11.711 and mean relative error of 0.210. Therefore, TEIM-Res has a wider range of applications than such a template-based modelling method. Furthermore, TEIM-Res only takes around 0.002 s (without graphics processing units) per sample, which is almost 20,000 times faster than TCRpMHCmodels.

We further compared with approaches originally designed for protein–peptide interaction prediction, including a docking tool GalaxyPepDock[18] and the deep learning-based approaches PepNN-Seq and PepNN-Struc[22]. GalaxyPepDock can beat neither TEIM-Res nor the average baseline in any metric (Fig. 1d and Extended Data Fig. 1). PepNN-Seq and PepNN-Struc also achieved worse performance for contact site prediction than TEIM-Res and the average baseline, in terms of both AUPR and MCC scores (Fig. 1e). The reasons why these protein–peptide interaction models failed may be that they only roughly recognize the

https://doi.org/10.1038/s42256-023-00634-4

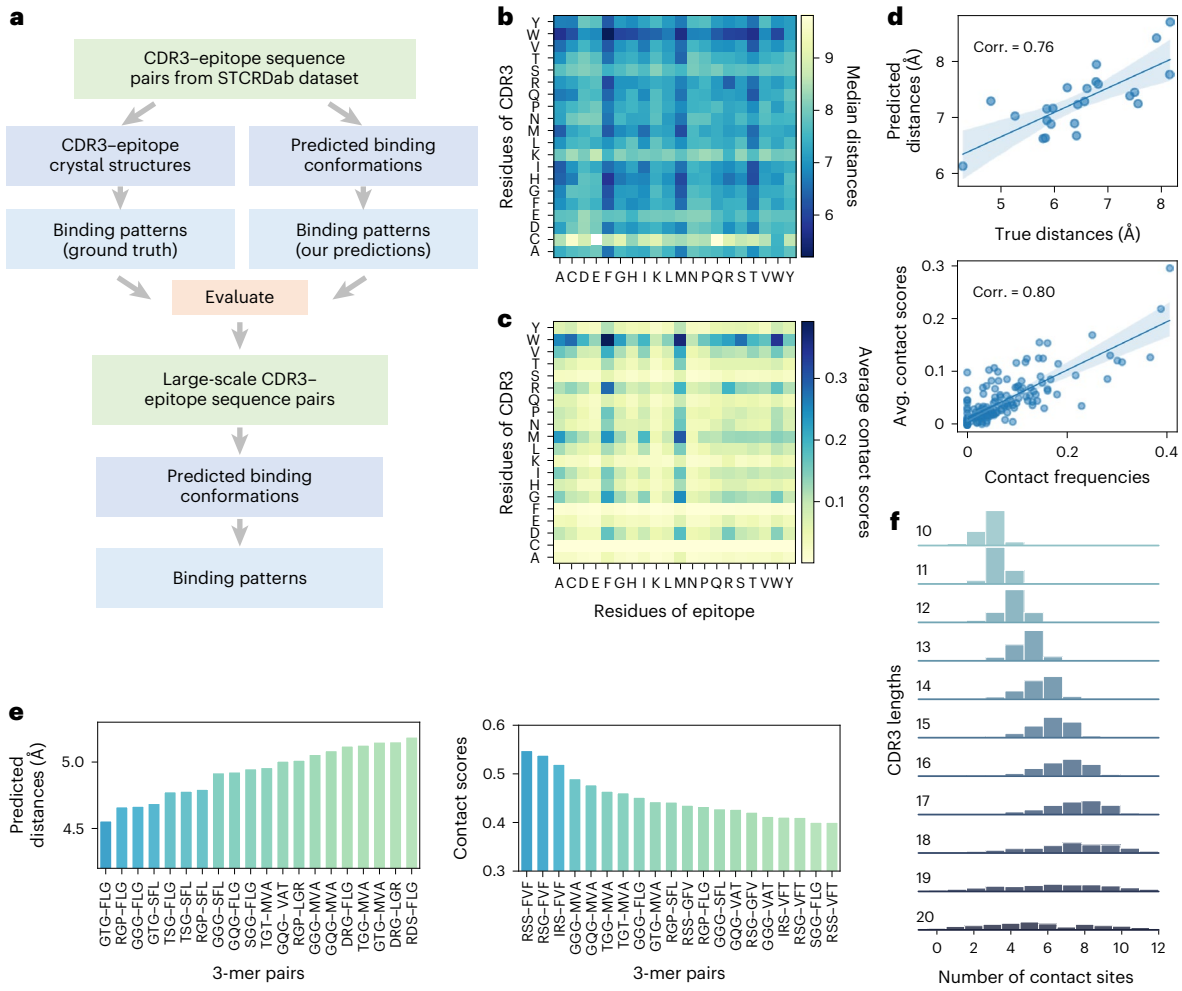


Fig. 5 | TEIM-Res can be used to discover the residue-level binding patterns of TCR–epitope interactions. a, The procedure of discovering the binding patterns. Firstly, from the STCRDab dataset which was relatively small, we calculated the binding patterns based on the true complex structures and predicted binding conformations separately, and then validated whether they were consistent. After validation, we applied TEIM-Res on a large TCR–epitope sample dataset with only sequence information and then obtained more binding patterns from this augmented dataset. **b**, The median distances of different types of residue pairs in the augmented dataset. Only the residue pairs with distances less than 10 Å were considered. **c**, The average contact scores of different contact pairs in the augmented dataset. **d**, The true and predicted average distances/contacts of different types of residue pairs in the STCRDab dataset. Each point in the scatter plots represents a type of residue pairs, and the bands represent the 95% confidence intervals for the linear fitting. The true and predicted distances were calculated as the average distances of residue pairs belonging to the same types and only those residue pairs closer than 10 Å were considered. The contact frequencies were calculated as the frequencies of the residue pairs forming contact, that is, the counts of contact pairs divided by the counts of all pairs for individual residue pair types. The average contact scores were calculated as the average predicted contact scores of residue pairs belonging to the same types. To avoid data bias, only those residue pairs that had over 30 counts in the dataset were considered. **e**, The average distances/contact scores of 3-mer pairs in the augmented dataset. Only the top 20 pairs with the closest distances or highest contact scores are shown. To avoid data bias, only those residue pairs with over 100 counts in the dataset were considered. **f**, The numbers of contact sites for CDR3s with different lengths in the augmented dataset.

CDR3β region with high contact probabilities but cannot precisely distinguish the contact sites within the CDR3β regions (Supplementary Fig. 1). Another reason is that, although TCR–epitope is a special case of protein–peptide, they may have unique binding patterns. Therefore, the models designed for protein–peptide interactions are not qualified for TCR–epitope interaction and thus demonstrate the necessity of developing specialized models like TEIM-Res.

Detailed analyses of model performance

We investigated how different properties of the samples influenced prediction accuracy. TEIM-Res showed satisfying accuracy on most samples (Extended Data Fig. 2) but failed for samples that were related to high MHC docking angles (the angles at which the TCR interacts with MHC[18]) or long sequences (Fig. 2a,b). The samples with high MHC docking angles (Protein Data Bank (PDB): 5SWS and 5SWZ) were actually outliers of canonical TCR-pMHC docking conformations[28]. The poor performance for long sequences was probably caused by insufficient data size (Supplementary Fig. 2).

Next, we evaluated the performance of distance prediction and contact prediction of all residue pairs (Supplementary Fig. 3). The predicted and true distances of all residue pairs showed a high correlation (Fig. 2c) and TEIM-Res can also distinguish contact residue pairs from those non-contact ones (Fig. 2d). The residue pairs of different types of amino acids showed different accuracy (Supplementary Fig. 4), but these differences were not caused by the different numbers of pairs in the dataset (Supplementary Fig. 5), suggesting that our model did not merely memorize the dataset but learned the underlying patterns. The residue pairs that were involved in the hydrophobic interactions or hydrogen bonds (annotated by PLIP[29]) showed relatively higher predicted contact scores (Fig. 2f), which can be explained by the fact

https://doi.org/10.1038/s42256-023-00634-4

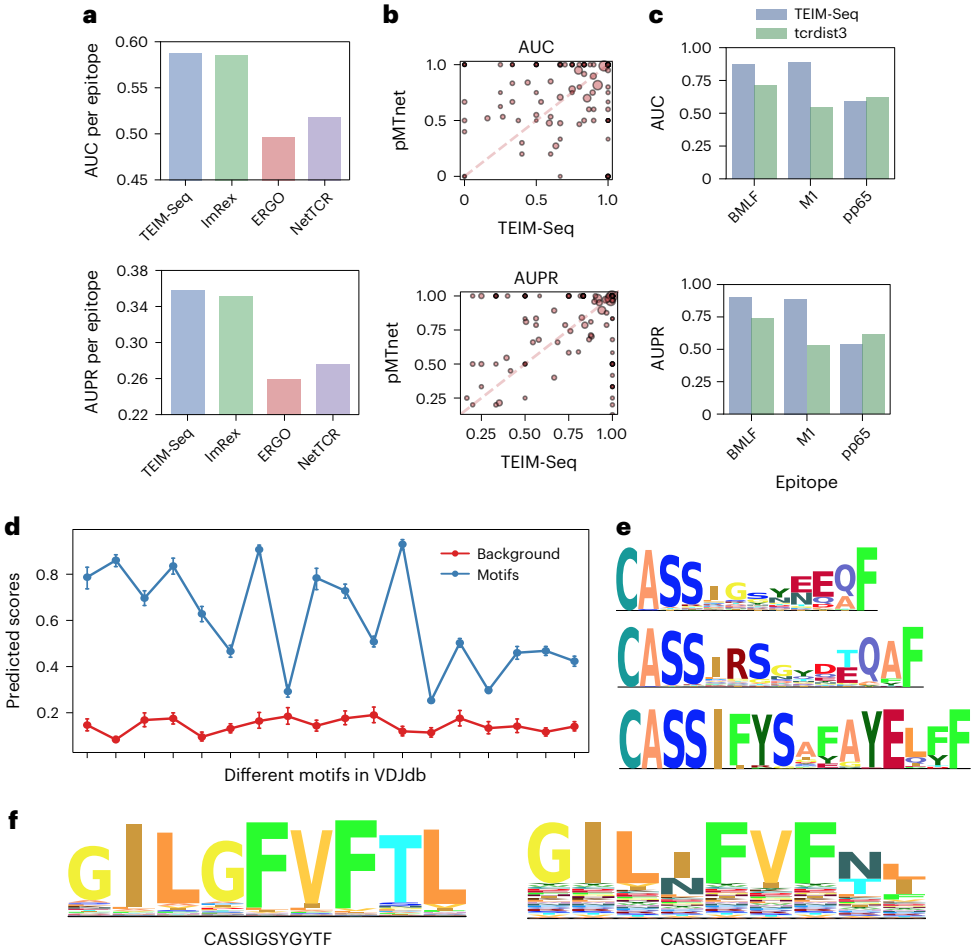


Fig. 6 | Validation and application of TEIM-Seq in sequence-level binding prediction. a, Comparisons between TEIM-Seq and the baselines ImRex, ERGO and NetTCR on sequence-level binding prediction in terms of AUC per epitope and AUPR per epitope. TEIM-Seq outperformed ERGO and NetTCR by a large margin and slightly exceeded ImRex. **b**, The scatter plots of AUC and AUPR scores for individual epitopes between TEIM-Seq and pMTnet. The average AUC scores of TEIM-Seq and pMTnet were 0.777 and 0.797, respectively, and the average AUPR scores of TEIM-Seq and pMTnet were 0.821 and 0.843, respectively. Thus, TEIM-Seq achieved slightly lower but still comparable performance to pMTnet. Furthermore, if we only considered the most frequent ten epitopes in the test set, the average AUC and AUPR scores were 0.798 and 0.873 for TEIM-Seq, respectively, which were higher than those of pMTnet (0.734 and 0.848 for AUC and AUPR, respectively). **c**, Comparisons of AUC and AUPR scores between TEIM-Seq and tcrdist3 for the three epitopes. TEIM-Seq greatly outperformed tcrdist3 on two epitope datasets and slightly fell behind for one epitope. **d**, The predicted binding scores of motifs for the epitope GILGFVFTL and background in the CDR3βs (shown as mean values with 95% confidence intervals). All motifs had higher predicted scores than their background CDR3β counterparts. **e**, The new motifs generated by TEIM-Seq. Here, we only show three example motifs of lengths 12, 14 and 16, respectively, and the cases of other lengths are shown in Supplementary Fig. 13. **f**, The cross-reactivities of two CDR3βs CASSIGSYGYTF and CASSIGTGEAFF identified by TEIM-Seq. Both sequences were the consensus sequences of the binding motifs of the epitope GILGFVFTL. The latter had higher cross-reactivity than the former, in that the latter recognized different residues at the fourth, eighth and ninth positions while the former only recognized unique residues at individual positions of the epitope.

that these residue pairs generally had closer distances than the others and hence were more easily predicted as contact pairs. The analyses about the effects of other factors on the performance, including the use of the autoencoder, the MHC features, the hyperparameters, the epitopes used for pre-training the autoencoder, the size of training set and the distance thresholds are provided in Supplementary Section 3.

We then analysed the contributions of the pre-training stage by comparing our original pipeline with two modified ones: the 'fine-tuning last' and 'no pre-training' pipelines. TEIM-Res achieved the best performance among the three methods (Fig. 2e), thus indicating the effectiveness of our training pipeline. Even the model that only fine-tuned the output modules (fine-tuning last) still achieved much better performance than the one without pre-training (no pre-training), demonstrating that the pre-training stage significantly contributed to the final residue-level prediction.

Furthermore, to investigate how much residue-level information was learned during the pre-training stage, we calculated the class activation map (CAM) of TEIM-Seq (Methods). Here, CAM can be interpreted as how much each residue pair contributed to the binary binding prediction. As expected, most samples showed a highly negative correlation between the CAMs derived by TEIM-Seq and the true distances between residues (Fig. 2g), confirming that the pre-trained model had already learned residue-level information. Therefore, although the pre-training stage did not engage any residue-level binding data, our model effectively captures the importance of different residue pairs contributing to the binding prediction by learning the latent residue-level interaction information from the sequence-level binding data.

TEIM-Res successfully captures subtle mutation effects

To better understand the detailed interactions between TCR–epitope pairs, we investigated whether TEIM-Res can reveal how the mutations influence the binding activity and the conformation through virtually mutating the CDR3β or epitope sequences. We tested our model on the

available mutation data of a well-studied TCR–epitope sample, A6 TCR and Tax epitope presented by HLA-A:0201 (named as A6-Tax)[30]. For the wild-type A6-Tax complex (in which the sequences of the CDR3β and the epitope are CASRPGLAGGRPEQYF and LLFGYPVYV, respectively), TEIM-Res can accurately capture the two contact regions, corresponding to the interactions of residue Y5 and residues P6V7Y8 of the epitope with residues of CDR3β (Fig. 3a and Extended Data Fig. 3). Next, we collected 32 mutant sequences of the A6-Tax complex from previous studies[31–40] (Supplementary Table 1) and then applied TEIM-Res to predict their binding conformations. The predicted distance/contact changes of the mutant samples were highly correlated with the affinity changes $\Delta\Delta G$ reported by the previous studies (Fig. 3b), indicating the ability of TEIM-Res in detecting subtle mutation effects on TCR–epitope interactions and revealing the corresponding binding affinity changes.

Among the mutant samples, we examined the conformation change of the one with the most affinity decrease, named A6_L7A-Tax_Y8A. This sample had only two mutant sites: L7A of CDR3β and Y8A of the epitope[39]. According to the prediction, most residue pairs around the mutant positions moved further away and displayed lower contact scores after mutation (Fig. 3c and Supplementary Fig. 20). We then examined the distances of the top 20 closest pairs of the wild-type sample for A6_L7A-Tax_Y8A and those with individual mutations (Fig. 3d), and found that both mutant sites contributed to the affinity decrease of the A6_L7A-Tax_Y8A sample (Supplementary Section 5). Such information captured by TEIM-Res can provide useful biological insights for applications such as optimizing TCR sequences for immunotherapy[41–44].

TEIM-Res detects key contact sites from TCR pools

For a large set of TCRs targeting a specific epitope, TEIM-Res can predict the binding conformations between all TCRs with the epitope and thus reveal the key contact sites of TCRs and epitopes. Glanville et al. proposed GLIPH to identify the enriched CDR3β motifs from TCR repertoires[7]. We therefore validated our model by comparing the predicted contact scores of the CDR3β motifs revealed by GLIPH with those non-motifs sites. As shown in Fig. 4a, the predicted contact scores of the motifs were significantly higher than those of the non-motif regions. Fig. 4b showed six representative cases containing the motif RSAY, in which the corresponding predicted contact scores were much higher than those of the non-motif regions for all these six CDR3β sequences.

Furthermore, for individual epitopes, we checked the crucial contact sites by averaging the predicted contact scores over all CDR3βs in the epitope-specific TCR pools. We analysed the predicted contact scores for three epitopes (Fig. 4c) and compared them to their corresponding crystal structure complexes binding with the partner TCRs. As shown in Fig. 4d and Extended Data Fig. 4, the predicted contact scores of all these three epitopes precisely matched their corresponding complex structures, suggesting that the predictions can assist in better understanding the underlying binding mechanisms of TCR–epitope interactions.

TEIM-Res discovers the patterns of TCR–epitope interactions

To further mine more knowledge about TCR–epitope interactions, intuitively we could utilize statistical methods to analyse the TCR–epitope complex structure dataset STCRDab[45] to obtain fine-grained patterns, such as the distance distributions and contact propensities of different residue pairs and different positions. However, due to the limited available structural data, these estimated patterns calculated through counting the frequencies from the dataset may suffer from bias and thus deviate from the true cases. Now, with TEIM-Res, we can construct a large-scale predicted binding conformation dataset from sequence data and use this augmented dataset for binding pattern discovery (Fig. 5a).

We first validated that the patterns derived from our predictions were consistent with those calculated from the ground truth conformations in the STCRDab dataset (Fig. 5d, Supplementary Figs. 6 and 7). Next, we derived an augmented large-scale conformation dataset predicted by TEIM-Res from sequence data to analyse residue binding patterns (Supplementary Fig. 8). We found that different types of residue pairs displayed totally different distance distributions, contact propensities and contact positions (Fig. 5b,c, Supplementary Fig. 9). We further split the CDR3β and epitope sequences into 3-mers and found that several 3-mers were more inclined to form close distances or contacts with others (Fig. 5e). The numbers of contact sites became more divergent as the lengths of sequences increased (Fig. 5f and Supplementary Figs. 10 and 11). These interesting results can provide useful hints about understanding the molecule-level mechanisms of TCR–epitope interactions.

TEIM-Seq performs well for sequence-level binding prediction

TEIM-Seq, the by-product obtained during our pre-training stage, was originally utilized for TEIM-Res initialization. Now we compared TEIM-Seq with other sequence-level TCR–epitope interaction prediction models to have a better sense of its performance. We compared TEIM-Seq with ImRex[12], ERGO[11], NetTCR[10], pMTnet[8] and tcrdist3[46] (Methods). TEIM-Seq achieved better or comparable performance with these models (Fig. 6a–c), which was satisfying considering that the architecture of TEIM-Seq was not originally designed and optimized for this sequence-level binding prediction task. More analyses of TEIM-Seq such as the rare epitopes, the influence of V/J genes and the influence of different datasets are provided in Supplementary Section 4.

Next, we introduced three potential applications of TEIM-Seq. Firstly, TEIM-Seq can achieve improved performance for a new epitope after being fine-tuned on extra data (Supplementary Fig. 12). Secondly, TEIM-Seq can discover TCR motifs of given epitopes. As an example, we first validated that TEIM-Seq can distinguish motifs[47] from the background (Fig. 6d) and then applied TEIM-Seq to search new motifs for the epitope of interest (Fig. 6e and Supplementary Fig. S13). Thirdly, we employed TEIM-Seq to detect the cross-reactivity of given TCRs[48]. We first validated TEIM-Seq on the DMF4 TCR with known cross-reactivity[48] (Supplementary Fig. 14) and then calculated the cross-reactivities for two TCR CDR3βs (Fig. 6f). In conclusion, TEIM-Seq can also provide decent performance for sequence-level binding prediction and may offer certain application potential in modelling TCR–epitope interactions.

Discussion

The interacting mechanisms between TCRs and epitopes are crucial for understanding TCR recognition. However, the number of solved TCR–epitope complex structures is limited, making it difficult to accurately characterize the interaction patterns. We proposed a specially designed model with a few-shot learning strategy to accurately predict the residue-level interactions of TCR–epitope pairs. We demonstrated its applications in mutation effect prediction, epitope-specific TCR analyses and interaction pattern discovery. In addition, the pre-trained model TEIM-Seq also achieved decent performance on sequence-level binding prediction.

Our pre-training stage can assist the model to learn abundant information from the sequence-level binding data and thus provide appropriate model initialization for residue-level interaction prediction. We speculate that this strategy can be used for other molecule–molecule interaction modelling tasks, in which a model is first pre-trained to predict the coarse-grained molecule-level interactions and then fine-tuned to further predict the fine-grained interactions.

MHC molecules are also involved in the TCR–epitope interactions and it is necessary to simultaneously characterize the interactions among TCRs, epitopes and MHCs in the future. Moreover, it is also important to generalize the model for interactions between $CD4^+$ TCRs and MHC-II presented epitopes in future work.

Methods

Data processing

We obtained TCR–epitope pairs with complex structures from STCRDab[45]. We limited the lengths of TCR CDR3βs between 10 and 20, and the lengths of epitopes between 8 and 12. We dropped the obvious noise data (PDB ID: 6UZI) and duplicate samples. With these filtering criteria, we built a CDR3–epitope complex structure dataset with 122 samples. Then we derived the pairwise residue distance matrices for individual structures in which each element represented the closest distance between the heavy atoms of two residues. We then defined the binding sites as those residue pairs within 5 Å and derived the contact matrices from the above-defined distance matrices. After that, we obtained the residue-level interaction dataset with binding distance and binding site labels. We mainly used packages Biopython 1.78, Numpy 1.19.1 and Pandas 1.1.3 to process data.

Next, we retrieved the sequence-level binding data from three databases: VDJdb[47], McPAS-TCR[49] and ImmuneCODE[50,51]. We only kept those pairs derived from human MHC class I, which were the majority of the datasets. We also limited the lengths of CDR3β sequences between 10 and 20, and the lengths of epitope sequences between 8 and 12. Furthermore, we excluded the 10x Genomics data in VDJdb mainly due to the controversial post-processing cut-offs[12]. After these processing procedures, there remained 14,933 positive samples from VDJdb, 4,759 from McPAS-TCR and 25,789 from ImmuneCODE. By combining these three datasets and removing duplicated pairs, we obtained a dataset with 45,481 positive samples covering 355 unique epitopes. We then generated negative pairs through random shuffling. We finally built a binding dataset containing 272,886 samples with a positive-to-negative ratio of 1:5.

Since our binding dataset only contained 355 unique epitopes, we also incorporated an autoencoder to extract informative features from the large-scale unlabelled epitope sequence dataset. In particular, the epitope sequences were downloaded from the Immune Epitope Database (IEDB)[52] with three filters: 'Epitope Structure: Linear Sequence', 'No B cell assays' and 'MHC Restriction Type I'. We dropped those epitopes with residue modifications and restricted the lengths of epitope sequences between 8 and 12. Finally, we constructed an epitope sequence dataset with 450,395 unique epitope sequences. We only kept sequences containing 20 standard amino acids. All epitopes were centre aligned and padded to length 12. All CDR3βs were aligned with IMGT numbering[53] using the ANARCI tool (v. 2021.02.04)[54] and padded to length 20.

Model architecture

The sequence feature extractors of TEIM consist of an embedding layer and a 1D CNN module, which is made up of a 1D CNN layer with kernel size 3, a batch normalization layer and a rectified linear unit (ReLU) activation function with hidden dimension 256. The interaction extractor is made up of two 2D CNN modules, each of which consists of a 2D CNN layer with kernel size 3 × 3 and hidden dimension 256, a batch normalization layer and a ReLU activation function. The autoencoder separately encodes the epitope sequence into a feature vector of dimension 32, which is then directly fed to an interaction module. The binding prediction module of TEIM-Seq is made up of a max-pooling layer and a fully connected layer with a sigmoid activation function. The residue-level prediction module consists of a 2D CNN module with two output channels. The first channel uses a ReLU activation function to predict residue distances while the second one uses a sigmoid activation to predict the probabilities of being binding sites. The autoencoder for encoding the epitope features is made up of two 1D CNN modules, both of which have 32 output dimensions, followed by a flatten layer and a fully connected layer with 32 output dimensions.

In the training process, an L2 regularization with a coefficient of 0.005 is utilized for all weights. All dropout rates are set to 0.2. TEIM-Seq was optimized using an Adam optimizer[55] with a learning rate of 0.0002. The losses of distance prediction and contact site prediction for TEIM-Res were summed up and optimized using an Adam optimizer with a learning rate of 0.001. Our model was implemented using Python 3.8.5 mainly with Pytorch 1.6.0 and Pytorch Lightning 1.0.3.

Validation settings

We mainly used a k-fold cross-validation for model evaluation on our collected datasets. More specifically, we randomly partitioned the whole dataset into k subsets and validated on one subset when trained on the other $k-1$ subsets. The average prediction over all k folds was used to evaluate the model performance. We conducted several computational experiments to evaluate the performance of TEIM-Res with different hyperparameters and the results indicated that our models maintained relatively robust performance across different hyperparameters (Supplementary Section 4). Since our models were not sensitive to the variation of hyperparameters, we did not explicitly search for hyperparameters.

During cross-validation, the data redundancy problem caused by similar TCR or epitope sequences may result in 'easy predictions', which may mislead the performance evaluation of different algorithms. To conduct a fair evaluation, we adopted a cluster-based strategy for cross-validation splitting, which was a reasonable strategy that had already been successfully used by sequence-based machine learning models, such as MONN[56] and CAMP[57], for performance evaluation. This strategy can ensure that no sequences between training and validation datasets shared similarities greater than a threshold, and thus simulate a more realistic setting for evaluating the prediction models. Here, the similarity between two TCR (or epitope) sequences p_i and p_j is defined as:

$$\frac{\mathrm{SW}(p_i, p_j)}{\sqrt{(\mathrm{SW}(p_i, p_i)\mathrm{SW}(p_j, p_j))}}, \tag{1}$$

where SW(·, ·) stands for the Smith–Waterman alignment score (calculated using the SSW library[58]) between two sequences. Then a single-linkage clustering algorithm was applied and the maximal similarities between any two sequences from different clusters were less than a threshold.

The threshold of sequence similarity should be small enough to distinguish the TCRs (epitopes) between training and validation datasets, but cannot be too small as it can lead to a small number of clusters with extremely large cluster sizes, which may influence the data splitting process during k-fold cross-validation. On the other hand, splitting the data using an overlarge threshold is not acceptable either, as it may yield a nearly random splitting result. Following the previous studies[56,57], we made a trade-off and chose similarity thresholds 0.5 for the sequence-level dataset and 0.8 for the residue-level dataset. We used the package Scikit-Learn 0.24.1 to cluster sequences.

The both-new splitting setting was evaluated through threefold cross-validation while the new-CDR3 and new-epitope splitting settings were through fivefold cross-validation. Detailed analyses of performance evaluation were conducted based on the predictions of the new-epitope setting. The metrics were calculated using Scipy 1.5.2 and Scikit-Learn 0.24.1.

We used the webserver of GalaxyPepDock to dock TCRs and epitopes. It takes protein structures and peptide sequences as input. We uploaded the structure of the whole TCR β chain and the epitope sequence for docking and calculated their predicted distance matrix. Then we calculated the corresponding predicted contact matrix by defining those residue pairs within 5 Å as contact sites. The contact score threshold of 0.1 was then used to predict contact sites. We downloaded and installed PepNN on a Linux server. PepNN-Seq requires sequences as input, while PepNN-Struc additionally requires the protein structure as input. Since PepNN only predicts the contact sites of

CDR3β, to make comparisons, we calculated the maximal values of the predicted contact matrices of TEIM-Res and average baseline over the epitope dimension to obtain the predicted contact scores for CDR3β.

Class activation map of CNN

Class activation map (CAM) is a common visualization method for explaining the prediction of CNN architectures in the computer vision area[59,60] and is usually interpreted as the contributions of individual pixels at the input image to the final classification. Here in TEIM-Res, CAM can be interpreted as how much each residue pair contributes to the final binary binding prediction result. For all TCR–epitope samples with known complex structures in the dataset, we calculated the correlations between the CAMs derived by TEIM-Seq and the true distances between residues to validate the intuition that residue pairs with closer distances contribute more to the binding prediction.

We modified the Grad-CAM algorithm to compute the activation of each residue pair for binding prediction[59]. Grad-CAM calculates the gradients of the predicted binding probability with respect to the outputs of the last CNN layer of TEIM-Seq. Then the outputs of the last CNN layer are averaged along the channel dimension with the gradients as the weights and negative values are clipped to zeros. Next, the values are normalized to be between zero and one. Finally, an average pooling layer with a window size 5 × 5 is applied to the matrix to obtain the final class activation map.

Analysis of the epitope-specific TCR pools

We assembled three epitope-specific TCR pools from ref. 7, which were specific to three HLA-A:0201-presented epitopes, that is, NLVPMVATV, GLCTLVAML and GILGFVFTL. We removed those samples that existed in our training set and used TEIM-Res to predict the contact site score matrix of epitopes with all CDR3βs in the corresponding pools. Then for each motif reported in ref. 7, we collected all CDR3s involved in the interactions in the corresponding pool and then calculated the average contact scores of both motifs and non-motif sites. To explore the contact site scores of a given epitope, we averaged the corresponding contact scores over all CDR3s in the pool.

Discovery of residue binding patterns

To validate whether the predictions of TEIM-Res can be applied for binding pattern discovery, we calculated the patterns from both the true crystal structure and the corresponding predicted conformation for each of 122 TCR–epitope samples derived from the STCRDab dataset. The validated patterns included the distances or contacts of different types of residue pairs or different relative positions. When considering the distances, only pairs closer than 10 Å were considered because such pairs are generally more important for conformation analyses. For distances or contacts of different types of residue pairs, only those pairs with more than 30 counts were considered to avoid randomness and bias. Since the sequences often have different lengths, we calculated the relative positions of individual sites in the sequences ranging from 0 to 1 and then discretized the relative positions into 10 bins. In the end, we evaluated the Pearson correlation coefficients between the patterns derived from ground truth and predictions.

After validation, we applied TEIM-Res to predict the binding conformations of TCR–epitope samples using only sequences, which contained 45,481 positive samples previously used in the pre-training stage, to derive a relatively large TCR–epitope binding conformation dataset. Next, for this augmented dataset, we calculated the distances/contacts of different types of residue pairs. Similarly, we only considered those residue pairs closer than 10 Å when computing the distance patterns. We also calculated the contact scores of specific residue pairs at different relative positions for the nine types of residue pairs that had the most contact pairs in the augmented dataset. Those position bins that contained less than five pairs were not considered. Finally, we calculated the distance/contact propensities of 3-mer pairs. The distances and contact scores of a 3-mer pair were defined as the average distances and contact scores of the nine pairwise combinations of corresponding residues. Only those 3-mer pairs that had more than 100 counts were considered in our analyses and only 3-mer pairs closer than 10 Å were used for distance pattern discovery.

Validation of TEIM-Seq

TEIM-Seq, ImRex, ERGO and NetTCR were compared through fivefold cross-validation on the VDJdb[47] and McPAS-TCR[49] datasets. Here, we only used the new-epitope splitting setting, which guaranteed that no epitopes between training and validation sets shared similarities greater than 0.5.

Since pMTnet further required the MHC sequences as input, we cannot directly compare it with TEIM-seq on our dataset. Instead, we trained and validated TEIM-Seq on the datasets used by pMTnet during comparison. In total there were 1,002 TCR–epitope-MHC pairs in this test set, among which there were 178 unique epitopes. We calculated the area under the receiver operating characteristic (AUC) and AUPR scores per epitope over the test set.

We also compared TEIM-Seq with tcrdist3, which was originally designed for a slightly different setting. More specifically, tcrdist3 analyses the sequence differences between TCRs and classifies the unobserved TCRs by measuring their sequence distances with the epitope-specific TCRs in the training data. We used the three human epitope-specific datasets provided by tcrdist3 during the comparison. To compare with tcrdist3, we first generated a set of negative TCR samples for each epitope through random shuffling. Then for each epitope, we randomly held out 20% TCRs as the test data. The prediction scores of tcrdist3 were calculated in the following way: we first used tcrdist3 to calculate the minimal distance between the query TCR and all positive TCRs of a given epitope in the training set, denoted as d_{epi}. We then calculated the minimal and maximal distances between the query TCR and all positive TCRs of all epitopes in the training set, denoted as d_{min} and d_{max}, respectively. Then prediction score was defined as:

$$\mathrm{score} = -\frac{d_{\mathrm{epi}} - d_{\mathrm{min}}}{d_{\mathrm{max}} - d_{\mathrm{min}}}.$$

We then calculated the AUC and AUPR values for each epitope based on the definition of this prediction score.

Applications of TEIM-Seq

We first fine-tuned TEIM-Seq for several novel epitopes in the ImmuneCODE dataset[50,51] with different numbers of extra training samples. We calculated the AUC and AUPR scores for each epitope and then used the average scores as the final metrics.

As the second application example, we chose the epitope GILGFVFTL to show how TEIM-Seq generated the binding motifs. To validate whether TEIM-Seq was able to discover the corresponding motifs for the binding TCR–epitope pairs, we compared the predicted binding scores of existing motifs in VDJdb versus the background CDR3β position weight matrices (PWMs) that were calculated from all CDR3s in the three binding datasets. For each motif or background CDR3β PWM, we randomly sampled 10,000 CDR3s according to the position weights and then calculated their average binding scores for the given epitope. We then generated new motifs for the epitope. In particular, we first randomly started from an initial CDR3β sequence from the background PWM and then used simulated annealing to improve the binding scores. More specifically, in each iteration, we randomly mutated a position of a CDR3β sequence and predicted its binding score with the epitope. If the predicted binding score was higher after mutation, we accepted the sequence otherwise we accepted it with a certain probability. After 10,000 iterations, all CDR3s with binding scores greater than 0.9 were used to calculate new motifs. The motif sequence logos were then plotted with Logomaker[61].

Next, to validate whether TEIM-Seq can identify the cross-reactivity of TCR CDR3βs, we predicted the binding scores of DMF4 TCR CDR3β with the ground-truth binding epitopes and the randomly sampled epitopes. The random epitopes were generated through sampling from the background epitope PWM which was calculated from the IEDB epitope sequence dataset[52]. We then used the following similar simulated annealing strategy to identify the cross-reactivity of all CDR3βs. In particular, given a CDR3β sequence, we first sampled an initial epitope from the IEDB epitope dataset, and then for each iteration we randomly mutated a position and predicted the corresponding binding score. If the binding score was improved after mutation, we accepted the change otherwise we only accepted it with a small probability. After 1,000 iterations, all epitopes with binding scores greater than 0.8 were used to evaluate the cross-reactivity of a given TCR CDR3β.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

We provide the processed data for our model training and evaluation in the GitHub repository at https://github.com/pengxingang/TEIM. The raw data were all downloaded from public websites. The sequence-level binding datasets were downloaded from VDJdb (https://vdjdb.cdr3.net/search), McPAS-TCR (complete database at http://friedmanlab.weizmann.ac.il/McPAS-TCR/) and ImmuneCODE (https://clients.adaptivebiotech.com/pub/covid-2020). The structures of TCR–epitope complexes were downloaded from STCRDab (https://opig.stats.ox.ac.uk/webapps/stcrdab/Browser?all=true#downloads). The epitope sequence dataset was retrieved from https://www.iedb.org/ by setting three filters ('Epitope Structure: Linear Sequence', 'No B cell assays' and 'MHC Restriction Type I') and pressing 'Export Results' for epitopes. The processed data for training the models (contact maps, sequence-level pairs, and all epitope sequences) is available at https://github.com/pengxingang/TEIM/tree/main/data. The affinity changes and sequences of the mutated A6-Tax sequences were retrieved from http://atlas.wenglab.org/web/search.php by searching TCR name A6 and also validated from their original papers (Supplementary Table 1). The TCR repertoire data for our analyses were retrieved from Supplementary Table 1 of https://www.nature.com/articles/nature22976. The crystal structures with the mentioned PDB IDs (5SWS, 5SWZ, 6UZI, 1AO7, 2VLJ, 3O4L, and 3GSN) were downloaded from the STCRDab dataset (https://opig.stats.ox.ac.uk/webapps/stcrdab/Browser?all=true#dbsearch).

Code availability

The source codes and model weights of TEIM-Res and TEIM-Seq are available on GitHub (https://github.com/pengxingang/TEIM) and Zenodo (https://zenodo.org/record/7604787)[62].

References

1. Peters, B., Nielsen, M. & Sette, A. T cell epitope predictions. *Ann. Rev. Immunol.* **38**, 123–145 (2020).
2. He, Q., Jiang, X., Zhou, X. & Weng, J. Targeting cancers through TCR-peptide/MHC interactions. *J. Hematol. Oncol.* **12**, 1–17 (2019).
3. Huppa, J. B. et al. TCR–peptide–MHC interactions in situ show accelerated kinetics and increased affinity. *Nature* **463**, 963–967 (2010).
4. Yamamoto, T., Kishton, R. & Restifo, N. Developing neoantigen-targeted T cell–based treatments for solid tumors. *Nat. Med.* **25**, 1488–1499 (2019).
5. Candia, Martín, Kratzer, B. & Pickl, W. F. On peptides and altered peptide ligands: from origin, mode of action and design to clinical application (immunotherapy). *Int. Arch. Allergy Immunol.* **170**, 211–233 (2016).
6. Joglekar, A. & Li, G. T cell antigen discovery. *Nat. Methods*, **18**, 873–880 (2021).
7. Glanville, J. et al. Identifying specificity groups in the T cell receptor repertoire. *Nature* **547**, 94–98 (2017).
8. Lu, T. et al. Deep learning-based prediction of the T cell receptor–antigen binding specificity. *Nat. Mach. Intell.* **3**, 864–875 (2021).
9. Gielis, S. et al. Detection of enriched T cell epitope specificity in full t cell receptor sequence repertoires. *Front. Immunol.* **10**, 2820 (2019).
10. Jurtz, V. A. et al. NetTCR: sequence-based prediction of TCR binding to peptide-mhc complexes using convolutional neural networks. Preprint at *bioRxiv* https://doi.org/10.1101/433706 (2018).
11. Springer, I., Besser, H., Tickotsky-Moskovitz, N., Dvorkin, S. & Louzoun, Y. Prediction of specific TCR-peptide binding from large dictionaries of TCR-peptide pairs. *Front. Immunol.* **11**, 1803 (2020).
12. Moris, P. et al. Current challenges for unseen-epitope TCR interaction prediction and a new perspective derived from image classification. *Brief. Bioinform.* **12**, bbaa318 (2020).
13. Kjærgaard, J. K. et al. TCRpMHCmodels: Structural modelling of TCR-pMHC class I complexes. *Sci. Rep.* **9**, 14530 (2019).
14. Lanzarotti, E., Marcatili, P. & Nielsen, M. Identification of the cognate peptide-MHC target of t cell receptors using molecular modeling and force field scoring. *Mol. Immunol.* **94**, 91–97 (2018).
15. Jumper, J. & Hassabis, D. Protein structure predictions to atomic accuracy with AlphaFold. *Nat. Methods* **19**, 11–12 (2022).
16. Yin, R., Feng, B. Y., Varshney, A. & Pierce, B. G. Benchmarking AlphaFold for protein complex modeling reveals accuracy determinants. *Protein Sci.* **31**, e4379 (2022).
17. Lin, X. et al. Rapid assessment of T-cell receptor specificity of the immune repertoire. *Nat. Comput. Sci.* **1**, 362–373 (2021).
18. Lee, H., Heo, L., Lee, MyeongSup & Seok, C. GalaxyPepDock: a protein-peptide docking tool based on interaction similarity and energy optimization. *Nucleic Acids Res.* **43**, W431–W435 (2015).
19. Ciemny, M. et al. Protein–peptide docking: opportunities and challenges. *Drug Discov. Today* **23**, 1530–1537 (2018).
20. Antunes, D. A. et al. Dinc 2.0: a new protein–peptide docking webserver using an incremental approach. *Cancer Res.* **77**, e55–e57 (2017).
21. Blaszczyk, M., Ciemny, MaciejPawel, Kolinski, A., Kurcinski, M. & Kmiecik, S. Protein–peptide docking using CABS-dock and contact information. *Brief. Bioinform.* **20**, 2299–2305 (2019).
22. Abdin, O., Nim, S., Wen, H. & Kim, P. M. PepNN: a deep attention model for the identification of peptide binding sites. *Commun. Biol.* 5, 503 (2022).
23. Ragoza, M., Hochuli, J., Idrobo, E., Sunseri, J. & Koes, DavidRyan Protein–ligand scoring with convolutional neural networks. *J. Chem. Inf. Model.* **57**, 942–957 (2017).
24. Yan, C. & Zou, X. Predicting peptide binding sites on protein surfaces by clustering chemical interactions. *J. Comput. Chem.* **36**, 49–61 (2015).
25. Zhao, Z., Peng, Z. & Yang, J. Improving sequence-based prediction of protein–peptide binding residues by introducing intrinsic disorder and a consensus method. *J. Chem. Inf. Model.* **58**, 1459–1468 (2018).
26. Wang, Y., Yao, Q., Kwok, J. T. & Ni, L. M. Generalizing from a few examples: a survey on few-shot learning. *ACM Comput. Surv.* **53**, 1–34 (2020).
27. Donahue, J. et al. DeCAF: a deep convolutional activation feature for generic visual recognition. In *Int. Conf. Machine Learning* 647–655 (PMLR, 2014).
28. Gras, S. et al. Reversed T cell receptor docking on a major histocompatibility class I complex limits involvement in the immune response. *Immunity* **45**, 749–760 (2016).

29. Adasme, M. F. et al. PLIP 2021: expanding the scope of the protein-ligand interaction profiler to DNA and RNA. *Nucleic Acids Res.* **5**, gkab294 (2021).
30. Garboczi, D. N. et al. Structure of the complex between human T-cell receptor, viral peptide and HLA-A2. *Nature* **384**, 134–141 (1996).
31. Borrman, T. et al. ATLAS: a database linking binding affinities with structures for wild-type and mutant TCR-PMHC complexes. *Proteins* **85**, 908–916 (2017).
32. Scott, D. R., Borbulevych, O. Y., Piepenbrink, K. H., Corcelli, S. A. & Baker, B. M. Disparate degrees of hypervariable loop flexibility control t-cell receptor cross-reactivity, specificity, and binding mechanism. *J. Mol. Biol.* **414**, 385–400 (2011).
33. Borbulevych, O. Y. et al. T cell receptor cross-reactivity directed by antigen-dependent tuning of peptide-mhc molecular flexibility. *Immunity* **31**, 885–896 (2009).
34. Haidar, J. N. et al. Structure-based design of a T-cell receptor leads to nearly 100-fold improvement in binding affinity for pepMHC. *Proteins* **74**, 948–960 (2009).
35. Li, Y. et al. Directed evolution of human T-cell receptors with picomolar affinities by phage display. *Nat. Biotechnol.* **23**, 349–354 (2005).
36. Pierce, B. G., Haidar, J. N., Yu, Y. & Weng, Z. Combinations of affinity-enhancing mutations in a T cell receptor reveal highly nonadditive effects within and between complementarity determining regions and chains. *Biochemistry* **49**, 7050–7059 (2010).
37. Borg, N. A. et al. The CDR3 regions of an immunodominant T cell receptor dictate the'energetic landscape'of peptide-MHC recognition. *Nat. Immunol.* **6**, 171–180 (2005).
38. Cole, DavidKenneth Increased peptide contacts govern high affinity binding of a modified TCR whilst maintaining a native PMHC docking mode. *Front. Immunol.* **4**, 168 (2013).
39. Piepenbrink, K. H., Blevins, S. J., Scott, D. R. & Baker, B. M. The basis for limited specificity and MHC restriction in a T cell receptor interface. *Nat. Commun.* **4**, 1948 (2013).
40. Ding, Yuan-Hua, Baker, B. M., Garboczi, D. N., Biddison, W. E. & Wiley, D. C. Four A6-TCR/peptide/HLA-A2 structures that generate very different T cell signals are nearly identical. *Immunity* **11**, 45–56 (1999).
41. Shang, X. et al. Rational optimization of tumor epitopes using *in silico* analysis-assisted substitution of TCR contact residues: molecular immunology. *Eur. J. Immunol.* **39**, 2248–2258 (2009).
42. Ochi, T. et al. Optimization of T-cell reactivity by exploiting TCR chain centricity for the purpose of safe and effective antitumor TCR gene therapy. *Cancer Immunol. Res.* **3**, 1070–1081 (2015).
43. Bassan, D. et al. Avidity optimization of a MAGE-A1-specific TCR with somatic hypermutation. *Eur. J. Immunol.* **51**, 1505–1518 (2021).
44. Gutierrez, L., Beckford, J. & Alachkar, H. Deciphering the TCR repertoire to solve the COVID-19 mystery. *Trends Pharmacol. Sci.* **41**, 518–530 (2020).
45. Leem, J., de Oliveira, SauloH. P., Krawczyk, K. & Deane, C. M. STCRDab: the Structural T-cell Receptor Database. *Nucleic Acids Res.* **46**, D406–D412 (2017).
46. Dash, P. et al. Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature* **547**, 89–93 (2017).
47. Bagaev, D. V. et al. VDJdb in 2019: database extension, new analysis infrastructure and a T-cell receptor motif compendium. *Nucleic Acids Res.* **48**, D1057–D1062 (2020).
48. Sewell, A. K. Why must T cells be cross-reactive? *Nat. Rev. Immunol.* **12**, 669–677 (2012).
49. Tickotsky, N., Sagiv, T., Prilusky, J., Shifrut, E. & Friedman, N. McPAS-TCR: a manually curated catalogue of pathology-associated T cell receptor sequences. *Bioinformatics* **33**, 2924–2929 (2017).
50. Klinger, M. et al. Multiplex identification of antigen-specific T cell receptors using a combination of immune assays and immune receptor sequencing. *PLoS ONE* **10**, e0141561 (2015).
51. Sidhom, J.-W. & Baras, A. S. Analysis of SARS-CoV-2 specific T-cell receptors in immunecode reveals cross-reactivity to immunodominant influenza M1 epitope. Preprint at *bioRxiv* https://doi.org/10.1101/2020.06.20.160499 (2020).
52. Vita, R. et al. The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res.* **47**, D339–D343 (2018).
53. Lefranc, M.-P. et al. IMGT unique numbering for immunoglobulin and T cell receptor constant domains and Ig superfamily C-like domains. *Dev. Comp. Immunol.* **29**, 185–203 (2005).
54. Dunbar, J. & Deane, C. M. ANARCI: antigen receptor numbering and receptor classification. *Bioinformatics* **32**, 298–300 (2015).
55. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. *Proc. of the 3rd International Conference on Learning Representations, ICLR 2015* (eds Bengio, Y. & LeCun, Y.) (2015).
56. Li, S. et al. MONN: a multi-objective neural network for predicting compound-protein interactions and affinities. *Cell Syst.* **10**, 308–322.e11 (2020).
57. Lei, Y. et al. A deep-learning framework for multi-level peptide–protein interaction prediction. *Nat. Commun.* **12**, 5465 (2021).
58. Zhao, M., Lee, W.-P., Garrison, E. P. & Marth, G. T. SSW Library: an SIMD Smith–Waterman C/C++ library for use in genomic applications. *PLoS ONE* **8**, e82138 (2013).
59. Selvaraju, R. R. et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vision* **128**, 336–359 (2019).
60. Qin, Z., Yu, F., Liu, C. & Chen, X. How convolutional neural networks see the world—a survey of convolutional neural network visualization methods. *Math. Found. Comput.* **1**, 149–180 (2018).
61. Tareen, A. & Kinney, J. B. Logomaker: beautiful sequence logos in Python. *Bioinformatics* **36**, 2272–2274 (2019).
62. Xingang, P. pengxingang/TEIM: TEIM. *Zenodo* https://zenodo.org/record/7604787 (2023).

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (T2125007 and 61872216 to J.Z.; 31900862 to DZ), the National Key Research and Development Program of China (2021YFF1201300), the Turing AI Institute of Nanjing, and the Tsinghua-Toyota Joint Research Fund.

Author contributions

X.P., Y.L., D.Z. and J.Z. conceived the concept. X.P. and Y.L. implemented the model and performed computational experiments. Y.L. and P.F. prepared and processed all data. X.P., Y.L., L.J., J.M., D.Z. and J.Z. analysed the results. X.P., Y.L., D.Z. and J.Z. wrote the paper with help from all the authors.

Competing interests

J.Z. is a founder of Silexon AI Technology and has an equity interest. All other authors declare no competing interests.

Additional information

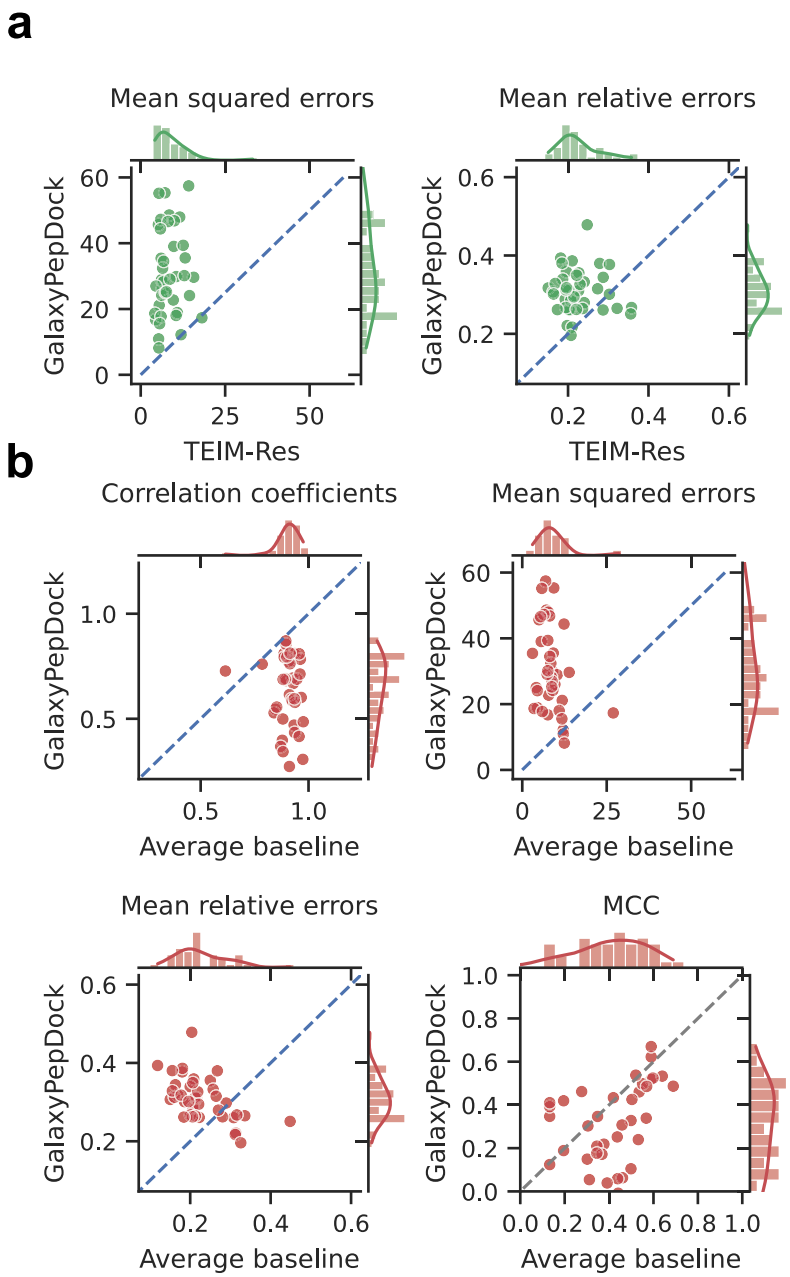
Extended data is available for this paper at https://doi.org/10.1038/s42256-023-00634-4.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-023-00634-4.

Correspondence and requests for materials should be addressed to Dan Zhao or Jianyang Zeng.

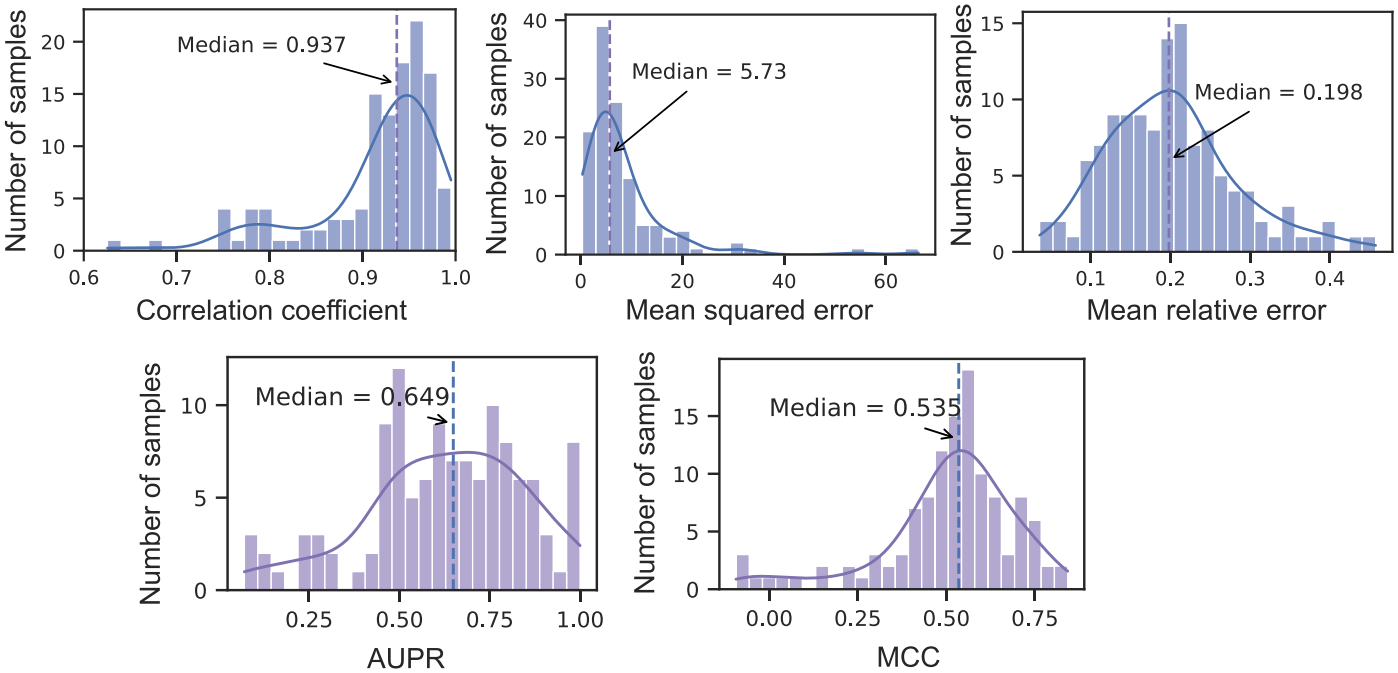
Peer review information *Nature Machine Intelligence* thanks Geir Kjetil Sandve, Pieter Meysman and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

https://doi.org/10.1038/s42256-023-00634-4



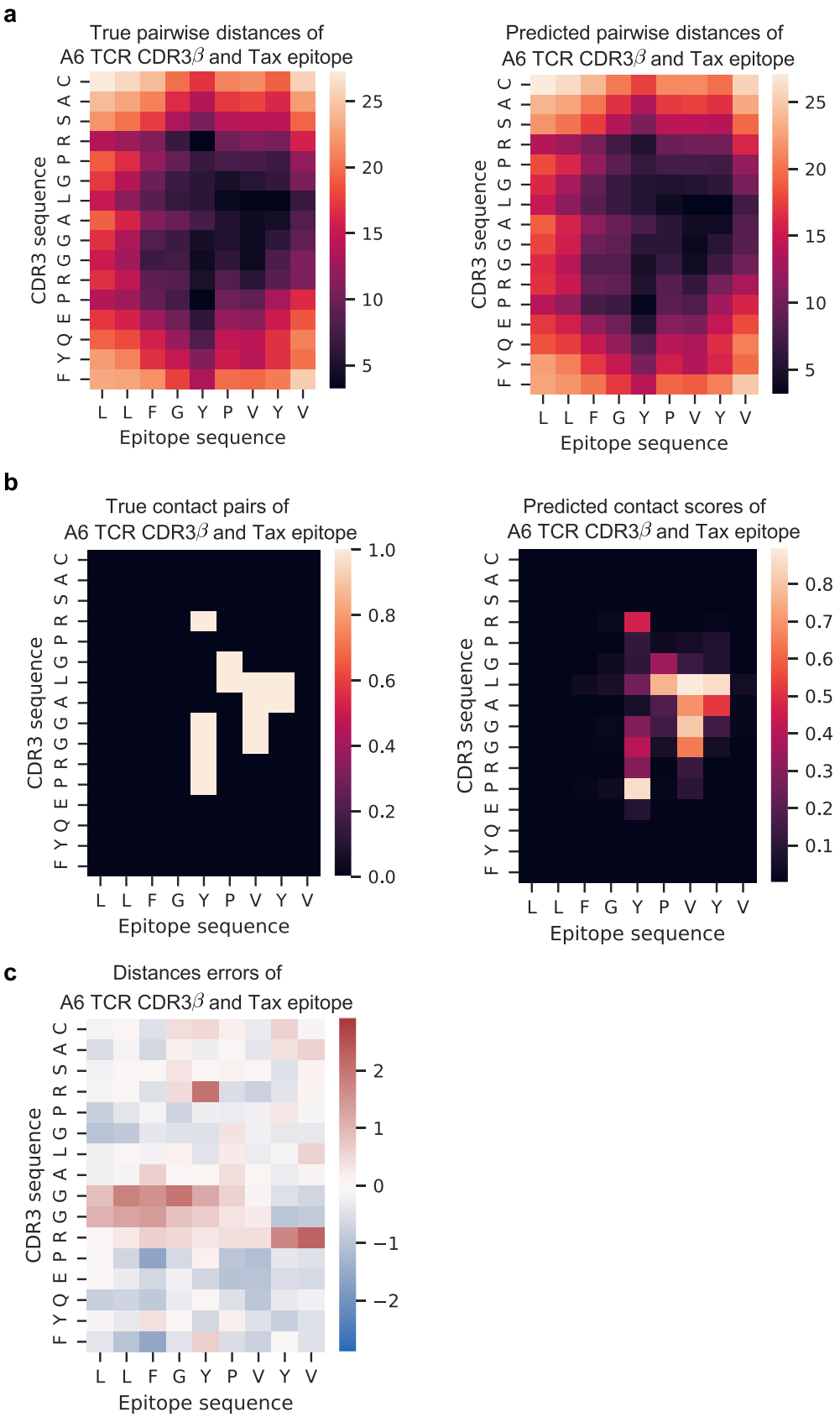
Extended Data Fig. 1 | Comparison among GalaxyPepDock, TEIM-Res and the average baseline. a, Comparison between GalaxyPepDock and TEIM-Res in terms of the mean squared errors and the mean relative errors. **b**, Comparison between GalaxyPepDock and the average baseline in terms of the correlation coefficients, the mean squared errors, the mean relative errors, and MCCs.

https://doi.org/10.1038/s42256-023-00634-4



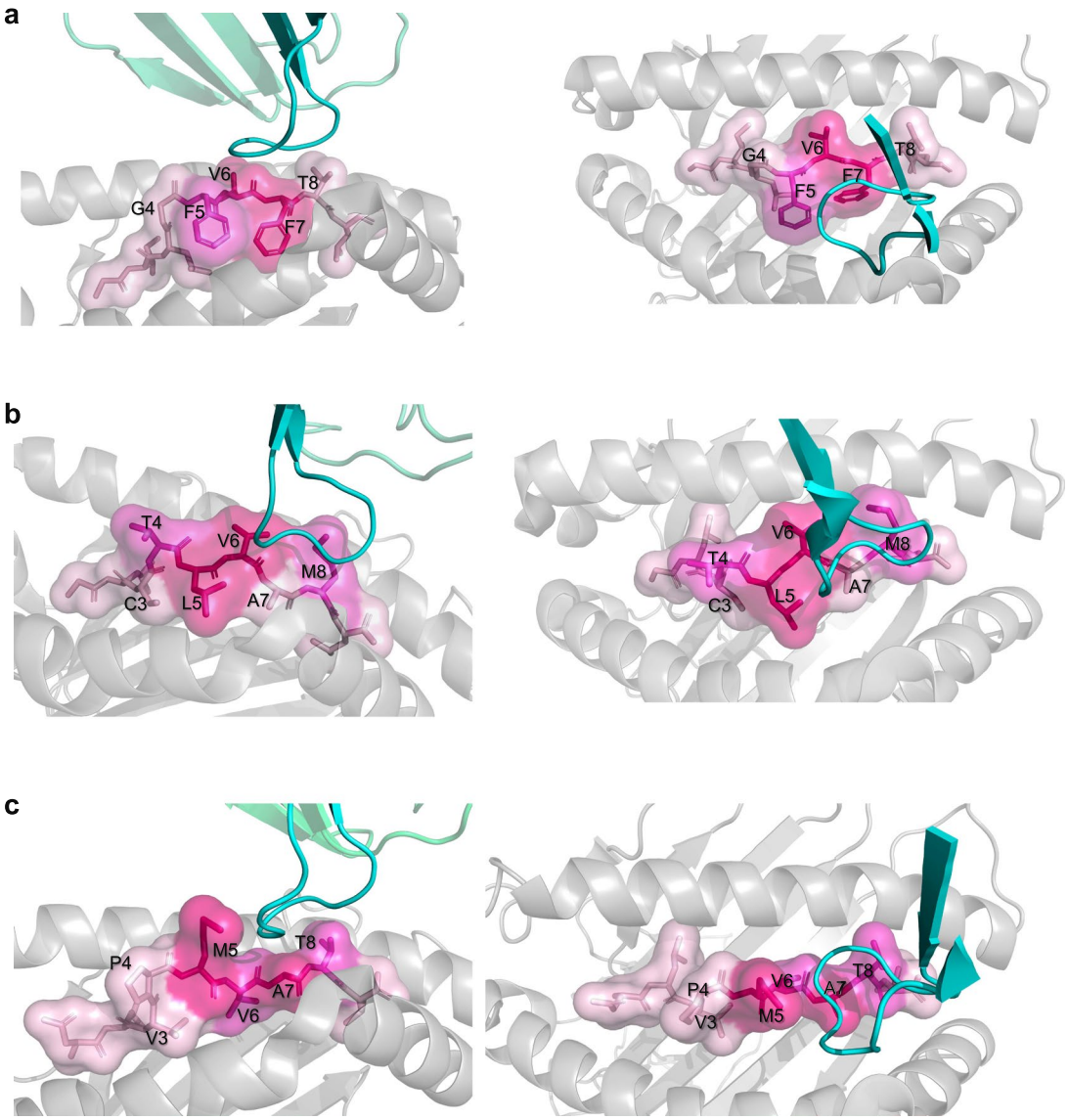
Extended Data Fig. 2 | The distributions of the individual evaluation metrics per sample under the new-epitope splitting setting. The five subfigures show the distributions of the correlation coefficient, mean squared error, mean relative error, AUPR, and MCC, respectively.

https://doi.org/10.1038/s42256-023-00634-4



Extended Data Fig. 3 | The true and predicted distances/contacts of the wild-type sample A6-Tax. a, The true and predicted pairwise distances of the A6-Tax sample. **b**, The true contact and predicted contact scores of the A6-Tax sample. **c**, The distance errors of the A6-Tax sample. The errors are defined as the predicted distances minus the corresponding true distances.

https://doi.org/10.1038/s42256-023-00634-4



Extended Data Fig. 4 | The crystal structures of the three epitopes interacting with the CDR3βs. a, Different views of the epitope GILGFVFTL interacting with CDR3β (PDB ID: 2VLJ). **b**, Different views of the epitope GLCTLVAML interacting with the CDR3β (PDB ID: 3O4L). **c**, Different views of the epitope NLVPMVATV interacting with the CDR3β (PDB ID: 3GSN). The CDR3βs are shown in cyan and the MHCs are shown in grey. The epitopes are shown in red and the darker emphasizes the important binding residues.

nature research

Corresponding author(s): Jianyang Zeng

Last updated by author(s): Feb 13, 2023

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection.
Data analysis	Our TEIM is available at https://github.com/pengxingang/TEIM. The software was built on Python 3.8.5, PyTorch 1.6, PyTorch-Lightning 1.0.3, Numpy 1.19.1, Scipy 1.5.2, Pandas 1.1.3, Scikit-Learn 0.24.1 and Biopython 1.78. PyMOL 2.4.1 was used to visualize structures. Logomaker 0.8 was used to draw the sequence logo plots. ANARCI version 2021.02.04 (http://opig.stats.ox.ac.uk/webapps/newsabdab/sabpred/anarci/) was used to number TCR sequences. The Smith-Waterman algorithm was implemented using SSW library v1.1 (https://github.com/mengyao/Complete-Striped-Smith-Waterman-Library). PLIP version 2.2.2 (https://github.com/pharmai/plip) was used to annotate non-covalent bonds. GLIPH version 2 (http://50.255.35.37:8080/) was used to annotate binding motifs for TCR pools.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

We provide the processed data for our model training and evaluation in the GitHub repository at https://github.com/pengxingang/TEIM. The raw data were all downloaded from the public websites. The sequence-level binding datasets were downloaded from VDJdb (https://vdjdb.cdr3.net/search), McPAS-TCR (complete database at http://friedmanlab.weizmann.ac.il/McPAS-TCR/) and ImmuneCODE (https://clients.adaptivebiotech.com/pub/covid-2020). The structures of TCR-

epitope complexes were downloaded from STCRDab (https://opig.stats.ox.ac.uk/webapps/stcrdab/Browser?all=true#downloads). The epitope sequence dataset was retrieved from https://www.iedb.org/ by setting three filters (``Epitope Structure: Linear Sequence'', ``No B cell assays'' and ``MHC Restriction Type I'') and pressing ``Export Results'' for epitopes. The processed data for training the models (contact maps, sequence-level pairs, and all epitope sequences) is available at https://github.com/pengxingang/TEIM/tree/main/data. The affinity changes and sequences of the mutated A6-Tax sequences were retrieved from http://atlas.wenglab.org/web/search.php by searching TCR name A6 and also validated from their original papers (Supplementary Table 1). The TCR repertoire data for our analyses were retrieved from Supplementary Table 1 of https://www.nature.com/articles/nature22976. The crystal structures with the mentioned PDB IDs (5SWS, 5SWZ, 6UZI, 1AO7, 2VLJ, 3O4L, and 3GSN) were downloaded from the STCRDab dataset (https://opig.stats.ox.ac.uk/webapps/stcrdab/Browser?all=true#dbsearch).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to predetermine sample size. The data used for model training and validation were downloaded from the corresponding databases and filtered according to the criteria described in the main text. There remained 122 samples for the residue-level dataset and 45,481 positive samples for the sequence-level dataset. The residue-level data were retrieved from all the available TCR-peptide-MHC structures.
Data exclusions	For the residue-level data preparation, the noise data point (PDB ID: 6UZI) was excluded because its CDR3 did not bind to the epitope. For the sequence-level data preparation, the 10X Genomics data in VDJdb were excluded due to the controversial post-processing cut-offs and the sequences with lengths dissatisfying the requirement were deleted.
Replication	For model validation, we used k-fold cross-validation, i.e., training and validating the model k times on different data splits. we chose k=3 for validation under the both-new splitting settings and k=5 for others. We prepared our code on GitHub and reran it. The results were consistent with those presented in the manuscript.
Randomization	The data used for training and validation are randomly shuffled and split. The cross-validation splitting guarantees that the sequence similarities between training and validation sets are lower than a threshold to avoid data redundancy.
Blinding	Authors were blinded to the validation set. We were blinded to the group allocation during data collection and analyses. The group allocation process was performed by computer scripy withou any manual intervention.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging