

Identifying T cell antigen at the atomic level with graph convolutional network

Corresponding Author: Professor Qinghua Jiang

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

This paper introduces deepAntigen, a novel computational method designed to identify T cell antigens at the atom level. Unlike existing methods that operate at the sequence or residue level, deepAntigen leverages graph convolutional networks (GCNs) to transform antigen, HLA, and TCR sequences into atom graphs. This allows for the extraction of crucial atom features, enhancing the prediction of antigen-HLA binding and antigen-TCR interactions.

Key Contributions:

Atom-Level Identification: deepAntigen is a new method to identify T cell antigens at the atom level, providing a more detailed understanding of antigen interactions.

Few-Shot Learning: The method employs a few-shot learning strategy to train the model with limited structural data, improving the accuracy of atom-level contact predictions.

Top-K Pooling Layer: To reduce computational workload, a top-k pooling layer is designed to select crucial atoms involved in binding or interaction, balancing accuracy and efficiency.

Clinical Applications: Deepantigen has shown promise in identifying tumor neoantigens, which successfully elicit T cell immune responses in vitro, highlighting its potential for personalized immunotherapies.

Major Gaps and Areas for Clarification:

Information Loss in Top-K Pooling: While the top-k pooling layer reduces computational workload, it can lead to information loss. Clarifying how this trade-off impacts overall prediction accuracy would benefit readers.

Limited Training Data: The availability of mono-allelic HLA II restricted antigens for training is limited. More details on how this limitation affects the model's performance and potential strategies to mitigate it would be helpful.

TCR Sequence Consideration: Currently, deepAntigen only considers the sequence of TCR β CDR3 due to technical constraints in single-cell TCR sequencing. Discussing how the inclusion of TCR α chains and other regions could enhance the model's performance would provide valuable insights.

Overall, deepAntigen represents a significant advancement in the field of T cell antigen identification, with promising applications for vaccine development and personalized immunotherapies. Addressing the identified gaps could further strengthen the method's robustness and applicability.

Other Comments:

Line 43-44 Most of the identified

Line 43-44 "fail to elicit" Suggestion: Please use common terminology for a general audience and out-of-the field readers for better understanding.

Line 46 Before launching into the description of the AI method, I advise the authors to provide a brief overview of the current identification method, weigh its advantages and disadvantages, and identify the shortcomings of the previous methods this paper aims to resolve.

Line 48 - 50 It's unclear what the authors intend to convey in this statement.

Line 51 Suggestion: Please provide references regarding the limitations of these methods.

Line 51 When talking about interactions, the more general term would be atomic-level, and authors repeatedly state singular terms like contact or interaction. Suggestion would be to use appropriate plural terms.

Line 51 This statement lacks clarity and specificity.

Overall, the introduction requires significant work to improve its structure and flow, which lack clarity.

Line 54 - 63 Elaborate on David Baker's method. Discuss potential challenges. Overall, the authors are addressing a significant issue, but they have not clearly state and discuss the current methods and limitations. Suggestion would be to elaborate and keenly discuss every technical term introduced in the intro.

Line 66 "atom level" Please go through the paper for consistency.

Consider adding background on deepAntigen's development and methodology details to provide context and clarify its innovations.

Overall, the introduction needs work, and Discuss limitations of the current study and future directions.

Line 77 - 82 Why only the CDR3 sequence? What about other CDRs, and which CDR3 alpha or beta chain for the general audience? The author needs to clarify and elaborate on these specific details.

Line 79 - 78 For the benefit of readers, authors should introduce deepAntigen_seq and deepAntigen_Atom in the introduction.

Line 120 Elaborate on few-shot learning methodology.

Line 119 - 122 .There is a lack of clarity, as each method appears to have utilized three separate datasets. However, the figure demonstrates the use of three distinct data sets for comparison with all the previously mentioned methods. Please clarify

Line 132 - 133 please elaborate

Line 188 -189 Discuss deepAntigen's scalability and limitations.

Line 216 - 217 elaborate deepAntigen's atom-level contact prediction methodology details.

Line 264 Discuss potential applications beyond immunogenic neoantigen identification.

Line 265 Compare deepAntigen with atomic-level prediction methods.

Line 281- 282 Citation and please elaborate on why it is challenging.

Discussion Comments.

1. Elaborate on deepAntigen's scalability and computational efficiency.
2. Discuss potential integration with experimental validation.
3. Performance of algorithm with other pooling method.
4. Provide system requirement and installation guide document. As a supplementary document.

(Remarks on code availability)

I have reviewed the code and the GitHub repository, and I am pleased to inform you that everything is in order. The code is well structured, and the repository is organized effectively.

Reviewer #2

(Remarks to the Author)

The authors present deepAntigen, a comprehensive computational framework developed to predict interactions between antigens, human leukocyte antigen (HLA) molecules, and T-cell receptors (TCRs) at both the sequence and atomic levels, aiming to advance the precision of neoantigen identification for personalized cancer immunotherapy. The deepAntigen framework comprises two main modules: deepAntigen_Seq and deepAntigen_Atom. deepAntigen_Seq is designed for sequence-level binding prediction between antigens and HLA molecules, as well as between antigens and TCRs.

Employing a 10-fold cross-validation approach, it fine-tunes hyperparameters, such as determining the optimal number of top-k atoms for analysis. deepAntigen_Atom targets the prediction of atom-level contacts within antigen-HLA/TCR complexes. Due to the limited data available at this level, it uses a leave-one-out cross-validation strategy, with performance assessed by the Negative Pearson Correlation Coefficient (NPCC) between predicted joint scores and actual atomic distances.

Core methodologies in deepAntigen incorporate Graph Neural Networks (GNNs) with positional encoding, transforming sequences into graph structures where atoms serve as nodes and bonds as edges. The integration of positional encoding embeds spatial information into atom features, enhancing the model's capability to distinguish atoms based on their specific molecular locations. Additionally, deepAntigen_Atom employs fine-tuning and multi-head attention mechanisms, focusing on oxygen and nitrogen atoms essential for non-covalent interactions, which optimizes the model's ability to discern critical atom pairs between antigens and HLA/TCR molecules. Evaluation metrics for deepAntigen include AUROC and AUPR for binary binding predictions, along with NPCC to atom-level accuracy. Benchmarking was conducted against other state-of-the-art antigen-HLA and antigen-TCR prediction tools using standardized parameters to ensure fair comparisons.

Experimental validation of deepAntigen involved patient-derived blood and tumor samples from breast and lung cancer cases, collected under stringent protocols. Peripheral blood mononuclear cells (PBMCs) were isolated and subjected to neoantigen validation. Whole exome sequencing (WES) and RNA sequencing on tumor samples identified mutations for neoantigen prediction, followed by TCR sequencing on patient blood to determine prevalent TCR clonotypes for computational neoantigen interaction analysis. An ELISPOT assay further validated neoantigen immunogenicity by evaluating the ability of neoantigen peptides to stimulate PBMCs to produce interferon-gamma (IFN- γ), indicating an

immune response.

In general, the paper is clear, and both the model and mathematical formulations are consistent and thoughtfully constructed. The work is very relevant, original, and timely. Perhaps, adding more detailed discussions on certain areas, such as performance variability across different tasks, the impact of hyperparameters, and potential limitations in specific settings (e.g., zero-shot learning and long-tail distributions), could strengthen the paper further. Providing this context would address our questions regarding the generalizability and real-world applicability of deepAntigen and help highlight the unique strengths of the approach while transparently acknowledging its limitations.

Major notes:

Selecting the right model architecture and tuning parameters are critical for the success of any predictive model. The choice of architecture directly impacts how well the model captures the biological complexity of interactions, while hyperparameters affect the model's accuracy and generalization.

Q1: Since there are multiple potential architectures (e.g., GNNs, transformers, recurrent neural networks) and different parameterization approaches, why did the authors choose GCNs with positional encoding over other architectures (e.g., transformers)?

Q2: How did the authors determine the optimal top-k value and other hyperparameters?

One of the primary concerns in bioinformatics and immunoinformatics is whether a model trained on a specific dataset can generalize to novel or diverse data. On lines 373 to 375 the authors say: "For each positive pair, a non-binding antigen was generated as a negative pair by cleaving a segment of equal length from the protein where the binding antigen is located."

Simply cleaving a segment of equal length from the same protein does not guarantee it will be a non-binder. The same protein can have multiple peptides capable of binding the same HLA, with varying affinities:

Blum, J. S., Wearsch, P. A., & Cresswell, P. (2013). "Pathways of antigen processing." *Annual Review of Immunology*, 31, 443–473;

This review discusses antigen processing pathways and how certain peptides, even from regions initially considered non-binding, can still be presented, especially when proteins are processed and presented via non-canonical pathways.) Without experimental validation or computational predictions specifically indicating that these segments do not bind, labeling them as non-binding is mainly speculative.

Neefjes, J., Ovaa, H., & Liu, Y. C. (2020). "A guide to major histocompatibility complex (MHC) class I and MHC class II antigen processing pathways in infectious diseases, autoimmune diseases, cancers, and vaccines." *Frontiers in Immunology*, 11, 1–16.

This paper shows the idea that antigen presentation is not limited to a single binding motif, and that various regions, including secondary and tertiary sites, may be involved in immune recognition.

Therefore, if any of these "negative" samples do in fact bind to HLA, they introduce false negatives into the dataset. This could confuse the model during training, as it may learn to associate features of binding segments with non-binding labels, potentially reducing model accuracy. The presence of false negatives can also skew performance metrics, making the model appear less sensitive to true binders. Plus, biological datasets, especially those involving antigens, HLAs, and TCRs, are notoriously heterogeneous and may not fully capture the diversity seen in real-world applications.

Q3: deepAntigen is capable of handling unseen or rare antigens and TCRs? (This would determine the model's utility in clinical and research settings where encountering novel epitopes is common)

Q4: How does deepAntigen generalize to antigens, HLAs, and TCRs not represented in the training data?

Q5: How do you address the risk of false negatives in negative samples?

Computational models, especially those involving deep learning and complex datasets, can be resource-intensive. Since deepAntigen uses advanced techniques such as CUDA acceleration and GNNs, I wonder about its practicality and efficiency, especially if intended for high-throughput or clinical settings.

Q6: How computationally intensive is deepAntigen, and can it be scaled to larger datasets or applied in real-time clinical settings?

I believe this would benefit from ICS or Tetramer Staining, that would allow for more detailed characterization of responding T-cell subsets but might miss low-frequency responses or require more resources and preparation. Also ELISA or Luminex would provide quantitative cytokine data but without single-cell resolution, potentially reducing sensitivity and making it harder to detect rare antigen-specific responses for further studies. However, IFN- γ ELISPOT is a well-suited choice for validating neoantigen immunogenicity in this study, especially given its sensitivity and direct relevance to T-cell activation.

Benchmarking new methods against established ones is crucial for demonstrating their utility and uniqueness. While deepAntigen was compared to ten other methods, I still have question on the specific strengths and weaknesses relative to these models. If deepAntigen shows variable performance across different tasks, I would like to see more clear statements about the trade-offs inherent in the tool design or if certain tasks are better handled by other specialized methods.

Q7: "Then deepAntigen was applied to identify neoantigens that could be presented by HLA and recognized by TCR." Which TCRs? The authors use patient data that includes TCRseq, but they do not mention having obtained matched antigen-TCR data from these samples. The validation here seems to be regarding the capacity to induce the activation of PBMCs, but without a direct link to a specific TCR sequence.

Q8: The authors briefly acknowledge in their Discussion the limitation of their methods regarding focus on CDR3 beta only. Is there a technical limitation of deepAntigen that would prevent it from working with the other CDR loops, or is really just a matter of lack of data for paired alpha-beta sequences and cognate peptides? In addition, when doing the structural analysis of the TCR peptide bonds, like in the case of NY-ESO, do the authors also see contacts with CDR3 alpha? In the authors perspective, how much the TCR-peptide accuracy could be improved by adding information of the other CDR loops?

Minor notes:

N1. Intro: "But it is currently an unsolved problem: most of identified T cell antigens fail to elicit T cell immune responses⁹." The text seems to merge the challenges of HLA-binding prediction with those of TCR-specificity prediction. Although deepAntigen targets both sides of the peptide interaction in a TCRpMHC complex, this is not often the case for other tools that target these different problems independently. In this case, is the aforementioned unsolved problem understood by the authors as mostly a fail of epitope discovery or of TCR specificity prediction? What is the core contribution of deepAntigen in this context, and what are secondary yet useful capabilities?

N2. "However, developing computational methods to predict antigen-HLA binding and antigen-TCR interaction at the atom level and further to infer the immunogenicity of the antigen, has not been explored yet." There is a long history of efforts to conduct structure-based HLA-binding prediction, going back to Rognan et al., 1999 (10579827) and Altuvia et al., 2004 (15542371), and more recently with tools such as PANDORA (38143769) and APE-Gen (38415656). Although such structure based methods have not fulfilled their potential to outperform sequence-based methods, in great part due to shortcomings of scoring functions, this is important work that should not be characterized as a problem that "has not been explored yet". Even structure-based immunogenicity prediction has been attempted before (31555277, 39276157).

N3. "With minimum modifications, deepAntigen can be used to mimic interactions between nucleic acids, proteins, and other molecules at the atom level". Is mimic the proper term here? I believe the authors mean to say that deepAntigen can be adapted to produce similar types of binding predictions for other molecules.

N4. Although the manuscript is generally well-written and clear, there are several passages that contain typos/mistakes. Given the high standards of the Journal and potential high impact of this publication, a more detailed revision of the text regarding style/grammar would be appreciated. Some examples of sentences that could be improved/fixed include:

- Abstract: "Precise identification of T cell antigens in silico is crucial for the development of cancer mRNA vaccine(s)"
- Intro: "T cell antigens refer to short peptide epitopes presented on Major Histocompatibility Complex (MHC) molecules, also known as Human Leukocyte Antigens (HLA) in humans, and recognized by the T cell receptor (TCR) on the surface of T cells." Although there is nothing wrong with this sentence, and I understand its meaning, it has some ambiguity. It could be interpreted as if the antigens are in the surface of the T cells.
- "Thus, deepAntigen could get out of trouble encountered by other methods..."
- "We analyzed this antigen library and identified motifs at positions 1, 4, 6, and 9, where are corresponding to pockets on the surface of HLA II molecules^{34,35}"
- "Atom-level contact prediction enable deepAntigen to discovery antigen-TCR motifs."
- "inspired by the biological fact that T cell immune response is determined by the catch-bond^{22,27–29}." This sentence should be revised both in terms of style and immunology. The role of catch bonds are significant and interesting, but authors should not state as if the presence of a single catch-bond is the only determining feature of T-cell activation.

This manuscript was co-reviewed by one of the reviewers and a Early Career Researcher. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Installation on a Google Colab environment with GPUs went well. The documentation was incredibly helpful in supporting my understanding of the package's features and applications in real-world scenarios. It provided clear guidance on best practices and effectively explained different use cases, making it much easier to start applying the package.

Reviewer #3

(Remarks to the Author)

This article presents a graph convolutional network-based T-cell antigen identification method, which has a significant

innovative contribution to atom-level antigen-HLA/TCR contact prediction. This is the first work showing the ability to recognize and learn atom-level information about the antigen, which potentially stimulates new biological findings in the related area.

Overall, this manuscript is well-organized. However, there are some remaining concerns that I am glad to highlight for the authors' notice:

Main concerns:

Insufficient novelty in terms of the machine learning methods: The proposed deepAntigen uses mostly a standard graph convolutional network (GCN). I can see it is certain that the proposed deepAntigen model presents some novelty in considering the atom-level feature to enhance the model's robustness in predicting, but the model design itself is still very basic. Technically speaking, the model design and training strategy are more like a real-world application demonstration of the well-known GCN, rather than a methodological contribution. I suppose the lack of a specific model design is one of the reasons of the observed limited improvement in predicting antigen-HLA binding is not significant (with only a 0.01 improvement in AUC for the recall compared to the MixMHCpred), as well as in other tasks.

Clinical verification is based on a very limited scale: One of the applaudable parts of this study is the proposed method has been used in practical clinical experiments, but only two patients' tissue is used to report the performance, which cannot generate trustworthy conclusions for me. The reason is simple, it is not convincing enough to avoid any coincidental biased reported results in neoantigens effectiveness, which is not also comparable to the figure reported by the tumor neoantigen selection alliance.

Minor concerns:

Some of the figures are quite confusing and even misleading: For instance, Fig.2 a) compares the ROC curves across different models, but they are so close to each other, and the claimed best one - deepAntigen even performs worse in some of the regions (e.g. When 1-specificity is small in the left subfigure, and when the recall is small in the right subfigure). Another misleading example is Fig. 2 b), the gap between the deepAntigen performance for the antigen length 8, 10 and 11 seems to significantly differ from the others, but this effect is obvious just because the y-axis scale is too small roughly from 0.9 to 1. As a result, it may not be justice to claim that the deepAntigen really generates a stable prediction accuracy but others do not. I would recommend, if possible, testing a wider range of antigen length to observe if this decreasing trend is more significant somehow. Moreover, Fig.2 c), most of the scatters concentrate in the top-right region, leading to badly distinguishable demonstration. I will be happy to check a more detailed plot about the data concentrating region for better comparison across the results. A problematic figure for me is Supplementary Fig.4. I get confused by the annotations shown in the figure, I suppose the right black vertical line should be the higher quantile at 96.71% the left one should be a smaller quantile at 70.24%, but I am really not sure if the annotations have been mis-located.

Some equations are not well explained and justified: For example, I guess the equation (10)-(12) demonstrates the core of the positional information for atom features encoding, but what do the symbols i , d , h , and PE actually mean? These equations really need more explanation and justification. For example, why this encoding approach is informative to include the positional information, and why sinusoidal functions are chosen? Similar to the equation (14), I cannot find an explanation for the two quantities e^s_i and e^s_j .

(Remarks on code availability)

Reviewer #4

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #2

(Remarks to the Author)

I have reviewed the revised manuscript and I am satisfied with all the corrections made. The authors have adequately addressed all my previous comments and concerns or indicated explicitly in the text the connections I have pointed in the first review. I believe that the paper can be accepted in its current form.

(Remarks on code availability)

Reviewer #4

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

I have been able to login on CodeOcean and get access to documentation.

I created a non-identifiable email for that.

I also tested the package installation on a Google Colab environment.

GPUs had to be used on the runtime.

I don't know if same can be done under a free account.

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Dear Referees,

Thank you very much for the constructive comments and the opportunity to resubmit the revised manuscript to **Nature Communications**. Enclosed are the point-by-point responses to the comments from four reviewers. In the point-by-point responses, the original comments from reviewers are marked **blue**, and our responses in black. In the revised manuscript, we marked all the changes in the manuscript with **dark red**.

Responses to Reviewer #1

Reviewer #1 (Remarks to the Author):

Overview

This paper introduces deepAntigen, a novel computational method designed to identify T cell antigens at the atom level. Unlike existing methods that operate at the sequence or residue level, deepAntigen leverages graph convolutional networks (GCNs) to transform antigen, HLA, and TCR sequences into atom graphs. This allows for the extraction of crucial atom features, enhancing the prediction of antigen-HLA binding and antigen-TCR interactions.

Key Contributions:

Atom-Level Identification: deepAntigen is a new method to identify T cell antigens at the atom level, providing a more detailed understanding of antigen interactions.

Few-Shot Learning: The method employs a few-shot learning strategy to train the model with limited structural data, improving the accuracy of atom-level contact predictions.

Top-K Pooling Layer: To reduce computational workload, a top-k pooling layer is designed to select crucial atoms involved in binding or interaction, balancing accuracy and efficiency.

Clinical Applications: deepAntigen has shown promise in identifying tumor neoantigens, which successfully elicit T cell immune responses in vitro, highlighting its potential for personalized immunotherapies.

Response: Thank you very much for your support of our work and for providing such insightful comments. We have carefully considered each of your suggestions and used them to revise our manuscript. We are confident that these changes have significantly enhanced the quality and clarity of our paper. Below, you will find detailed responses to each of your valuable comments.

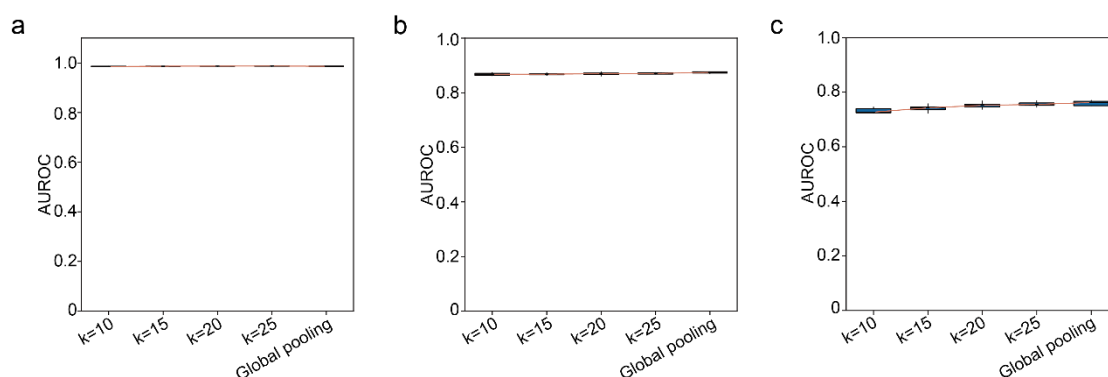
#Major Gaps and Areas for Clarification:

1.Information Loss in Top-K Pooling: While the top-k pooling layer reduces computational workload, it can lead to information loss. Clarifying how this trade-off impacts overall prediction accuracy would benefit readers.

Response: We appreciate the comment raised by the reviewer. We compared the impact of the Top-k Pooling with the commonly used Global Max Pooling methods on model performance. Additionally, we analyzed how varying k-values influence the performance. Our results indicated that Top-k pooling results in a slightly lower overall prediction

performance, **with an AUROC approximately 0.01 lower** compared to Global Max Pooling. In general, **as the k-value increases, the accuracy gradually improves** (Supplementary Fig.19).

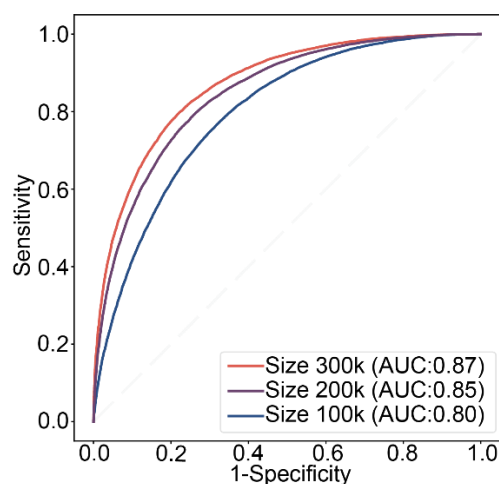
In the revised manuscript, we have clarified the information loss in Top-k Pooling in the **Discussion** to make it clearer and have included the relevant results in the **Supplementary Fig.19**.



Supplementary Fig.19 | The impact of top-k pooling on the performance of deepAntigen a, in prediction of antigen-HLA I binding; **b**, in prediction of antigen-HLA II binding; **c**, in prediction of antigen-TCR binding.

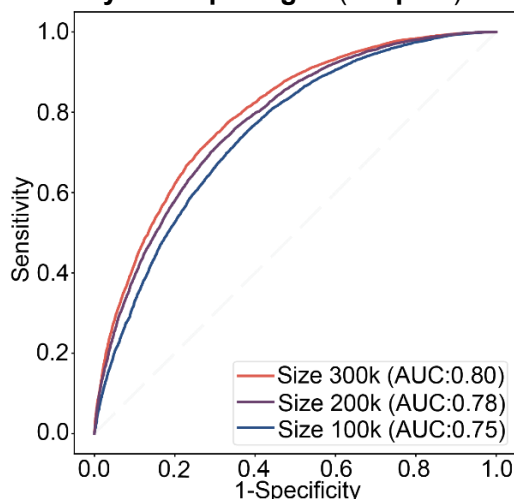
2.Limited Training Data: The availability of mono-allelic HLA II restricted antigens for training is limited. More details on how this limitation affects the model's performance and potential strategies to mitigate it would be helpful.

Response: We are truly grateful for the reviewer to raise this suggestion. We revisited the literature to collect a mono-allelic HLA II dataset and discovered a recently published study by Martin et al. (**PMID: 37301199**). This study expanded the dataset of Abelin et al. (**PMID: 31495665**) that was used by deepAntigen. The dataset of Abelin et al. contains only **52,188** positive antigen-HLA II pairs, covering **43** HLA II alleles, while the dataset of Martin et al. includes **175,291** positive antigen-HLA II pairs, covering **83** HLA II alleles after removing redundant antigens. Then, we generated negative antigens to balance the dataset and split it into a training dataset and a test dataset with an 8:2 ratio. The portions of the test dataset that overlap with the training dataset of other methods (**PMID: 37023751; 37085710**) were extracted and added to our training dataset. Finally, our training dataset contains 299,324 antigen-HLAII pairs, and the test dataset contains 50,956 antigen-HLAII pairs. We randomly sampled 100k, 200k and 300k pairs from our training dataset to construct three new training datasets with different size. We retrained deepAntigen three times using three new training datasets and tested three models on the test dataset. The experimental results show that **as the size of the training dataset increases, the performance of deepAntigen on the test dataset gradually improves** (Supplementary Fig.20).



Supplementary Fig.20 | The impact of size of the antigen-HLA II training dataset on performance of deepAntigen.

We further selected antigens from the above test dataset that did not appear in any of the three training datasets of different sizes, and analyzed the impact of data scale on the ability of deepAntigen to recognize unseen antigens. We obtained a test dataset containing 19,322 antigen-HLA II pairs. The test results show that **limited training data does indeed affect the generalization ability of deepAntigen (Graph 1).**



Graph 1 | The impact of size of the antigen-HLA II training dataset on the generalization ability of deepAntigen.

Taken together, our analysis reveals that limited training data affects the overall prediction accuracy and generalization performance of deepAntigen. This issue improves as the data scale increases. Other studies have used deconvolution algorithms, such as NNAlign_MA (**PMID: 31578220**) and MoDec (**PMID: 31611696**), to simultaneously resolve both the binding core and allele assignment of each antigen from multi-allelic HLA II ligandome. These methods can expand the number of antigen-HLA II pairs, but it is important to carefully evaluate the data quality to avoid bias that are not consistent with actual data distribution (**PMID: 38000035**). Therefore, this strategy was not included in our study; instead, we used mass spectrometry-based mono-allelic HLA II ligandome, which provides the “gold standard” for identification of HLA II binders.

In the revised manuscript, we have clarified how limited training data influences

the performance of deepAntigen and discussed potential strategy to mitigate it in the **Discussion and Supplementary Fig.20**. In addition, we updated the performance of retrained deepAntigen in prediction of antigen-HLA II binding on new test dataset in the **Result and Fig.2**.

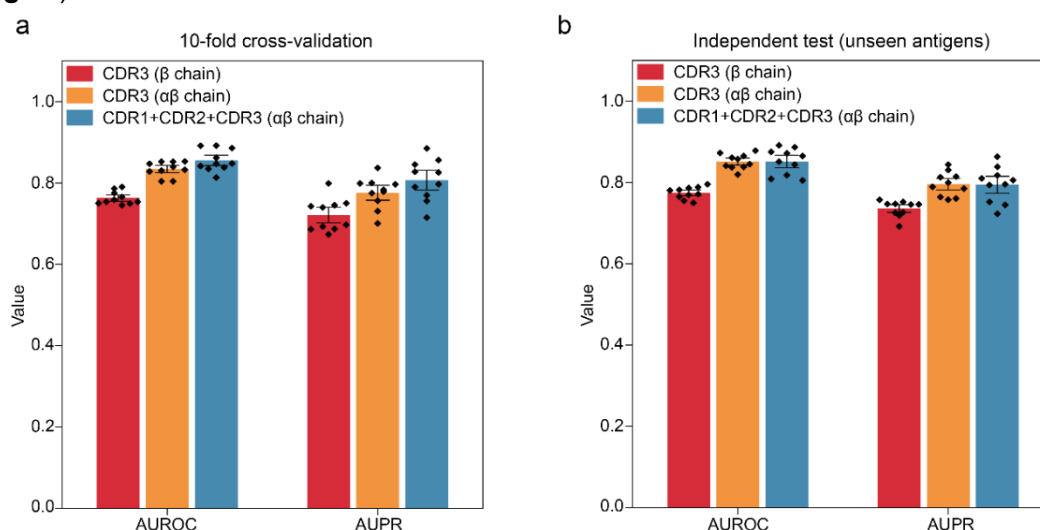
3.TCR Sequence Consideration: Currently, deepAntigen only considers the sequence of TCR β CDR3 due to technical constraints in single-cell TCR sequencing. Discussing how the inclusion of TCR α chains and other regions could enhance the model's performance would provide valuable insights.

Response: We thank the reviewer for suggesting further clarification on how the inclusion of TCR α chains and other regions enhances the performance of deepAntigen. TCRs are heterodimers composed of one α and one β chain. The diversity of TCRs is achieved during the V(D)J recombination when a unique combination of V and J (for the α chain) and V, D, and J (for the β chain) germline-encoded segments is selected and assembled. Additional diversity occurs through N- and P- nucleotide insertions at the V(D)J junctions. These regions, referred to as complementarity-determining regions 3 (CDR3), are mainly involved in the recognition of the antigen, while two other CDRs (CDR1 and CDR2) located on the V segments mediate contact primarily with the MHC. **Both chains of the TCR ultimately determine the TCR specificity (PMID: 38615042; PMID: 33981311)**. A previous study of GLIPH reported that **both CDR3 α and CDR3 β chain are involved in contacting with antigen through many 3D structure analysis cases while the CDR3 β rather than CDR3 α are necessarily required (PMID: 28636589)**.

Conventional bulk sequencing methods have been applied to sequence the TCRs in antigen-specific T cell populations one chain at a time or only the β chain. More recently, single-cell TCR sequencing has enabled the acquisition of paired $\alpha\beta$ TCR sequences. One of the most comprehensive $\alpha\beta$ TCR sequence datasets of antigen-specific T cells was generated by the 10x Genomics immune profiling platform, coupling DNA barcoded pMHCs multimers with single-cell TCR-Seq. These data constitute around 70% of all paired $\alpha\beta$ TCR-antigens currently stored in VDJdb. Recently developed quality control tools suggest that **not all these interactions are of equal quality (PMID: 33990328)**.

We collected antigen-CDR3 $\alpha\beta$ pairs from VDJdb and found that **only 3,011 antigen-CDR3 $\alpha\beta$ pairs** had high confidence (Score>0). After remove TCRs without known TRAV/TRBV genes, we retained **1,734** antigen-CDR3 $\alpha\beta$ pairs, which contains 345 unique antigens, 1,377 unique CDR3 α sequences and 1,536 unique CDR3 β sequences. In this dataset, 49 antigens with more than 5 known interacting TCRs were used as the training dataset, and the remaining 296 antigens were used as the independent test dataset. Then we collected 700,660 TCRs (including paired CDR3 $\alpha\beta$ sequence and TRAV, TRBV gene) derived from healthy donors as control TCRs, obtained from a previous study (**PMID: 39216000**). Non-interacting TCRs in negative antigen-CDR3 $\alpha\beta$ pairs were randomly sampled from the control TCRs, which had a very low probability of encountering TCRs interacting with the given antigen. CDR1 and CDR2 sequences are assembled using the TRAV/TRBV gene reference sequences from IMGT. Then, we split the training dataset into 10 subsets and established three different scenarios: using only CDR3 β , using paired CDR3 $\alpha\beta$ and using paired CDR $\alpha\beta$. Due to the small size of the dataset, we adjusted the parameters of deepAntigen to retrain, using 3 graph convolutional layers with 16 hidden units. The results of the 10-fold cross-validation indicated that **incorporating TCR α chains**

and other regions can improved the performance of deepAntigen in prediction of antigen-TCR interactions. However, the results of the independent test indicated that the inclusion of other regions did not significantly improve performance (**Supplementary Fig.21**).



Supplementary Fig.21 | The performance improvement with inclusion of TCRα chains and other regions. a, The performance comparison of 10-fold cross-validation. **b,** The performance comparison of independent test.

Taken together, TCRα chain and other regions are useful for recognizing antigens. However, due to the limited antigen-TCRαβ pair data and uneven data quality, we have not incorporated the CDR3α chain and other regions into the modeling of deepAntigen. Future updates are expected when more information on alpha chains becomes available. We have included these results and explanations in the **Discussion** and **Supplementary Fig.21**.

Overall, deepAntigen represents a significant advancement in the field of T cell antigen identification, with promising applications for vaccine development and personalized immunotherapies. Addressing the identified gaps could further strengthen the method's robustness and applicability.

Response: Thank you very much for your feedback and constructive criticism of our work. We are confident that this will greatly enhance the robustness and practicality of our manuscript, and we thank you once again.

#Other Comments:

1.Line 43-44 Most of the identified

Response: Thank you for your significant reminding. According to your suggestion, we corrected the above grammatical error.

2.Line 43-44 "fail to elicit" Suggestion: Please use common terminology for a general audience and out-of-the field readers for better understanding.

Response: We apologize for any confusion caused by our wording. We changed "fail to elicit" to "did not activate", which is more understandable to general readers.

3.Line 46 Before launching into the description of the AI method, I advise the authors to provide a brief overview of the current identification method, weigh its advantages and disadvantages, and identify the shortcomings of the previous methods this paper aims to resolve.

Response: We thank the reviewer for this suggestion and apologize for the lack of clarity in explaining the advantages and disadvantages of the current identification method, as well as the problems our paper aims to address. Based on the reviewer's suggestion, we have reorganized the Introduction section, providing a detailed discussion of the advantages and disadvantages of the current methods, as well as the existing gaps. This leads to the presentation of the problems our paper aims to address and the main research topics. Here is the revision:

“Early computational methods for antigen-HLA binding prediction relied on antigen sequence similarity and amino acid position-specific scoring matrices to develop scoring functions for estimating antigen-HLA binding affinity, including SYFPEITHI¹⁴, MixMHCpred¹⁵, SMM-align¹⁶ and MixMHC2pred¹⁷. Alternatively, other computational approaches for antigen-TCR interaction prediction grouped TCRs based on TCR sequence similarity to determine antigen specificity, with notable methods such as GLIPH¹⁸, TCRdist¹⁹, GIANA²⁰ and DeepTCR²¹. These methods are generally tailored to specific HLA alleles or antigens and they cannot efficiently capture the non-linear relations and complicated patterns, resulting in limited generalizability.”

4.Line 48 - 50 It's unclear what the authors intend to convey in this statement.

Response: We appreciate the comment raised by the reviewer and apologize for the lack of clarity in conveying our intended message. What we intended to convey is that Incorporating structural information on residue contacts, represent an advancement over methods that rely solely on sequence data. We have revised this statement and adjusted it to the section discussing structure-based methods in lines 74-75.

“Incorporating information about the HLA-antigen-TCR complex structure could provide new insights into T cell specificity^{41–46}.”

5.Line 51 Suggestion: Please provide references regarding the limitations of these methods.

Response: Thank you for your suggestion. After reorganizing the **Introduction**, we elaborated more on the limitations of current methods when summarizing the existing issues, and added relevant references. Please see lines 83-93.

6.Line 51 When talking about interactions, the more general term would be atomic-level, and authors repeatedly state singular terms like contact or interaction. Suggestion would be to use appropriate plural terms.

Response: We appreciate the reviewer's important suggestion. We have changed “*atom-level*” to “*atomic-level*” throughout the manuscript and have proofread the use of singular and plural forms in the text.

7.Line 51 This statement lacks clarity and specificity.

Response: Thank you for your suggestion. In the revised manuscript, we first highlighted the importance of atomic-level structural information for predicting antigen-HLA binding and

antigen-TCR interactions and then pointed out that the inability of sequence-based methods to capture atomic-level interactions is a current limitation. This should make the point clearer. The revised sentence can be found in lines 86-87:

“Second, sequence-based methods are insufficient to characterize the refined atomic-level interactions between antigen and TCR⁴¹.”

Overall, the introduction requires significant work to improve its structure and flow, which lack clarity.

Response: We greatly appreciate the reviewer's valuable comments and patient guidance, which have been very helpful in improving the logic and flow of our paper. We have revised to our **Introduction** according to the reviewer's suggestions, addressing each point one by one.

8.Line 54 - 63 Elaborate on David Baker's method. Discuss potential challenges. Overall, the authors are addressing a significant issue, but they have not clearly state and discuss the current methods and limitations. Suggestion would be to elaborate and keenly discuss every technical term introduced in the intro.

Response: Thank you for pointing out this important aspect. We apologize for this lack of clarity. In this revision, we have added the clarification and discussion about structure-based methods in lines 75-72 and lines 87-89:

“Recently, revolutionary AI tools like RosettaFold All-Atom⁴⁷ and AlphaFold3⁴⁸ have provided atomic-level insights into studying general molecular interactions. RosettaFold All-Atom⁴⁷ utilizes SE(3)-Transformers to generate 3D backbone from homologous structure templates and performs all-atom expansion according physicochemical principles. AlphaFold3⁴⁸ leverages a diffusion network to transform an atomic cloud into the accurate molecular structure through iterative refinement. But they still struggle to capture the relative contributions of details such as hydrogen bonding, electrostatic forces, and hydrophobic effects to the overall stability and specificity of interactions^{49,50}.”

“Third, the scarcity of high-quality HLA-antigen-TCR complex crystal structures limits the effectiveness of template-based structure prediction tools in accurately modeling the diverse range of HLA-antigen-TCR interactions⁵⁰.”

In our study, we incorporated potential atomic-level interaction information from abundant sequence-level data, which can both improve the resolution of identification of antigen-TCR interactions and alleviate the issue of insufficient structural data.

9.Line 66 “atom level” Please go through the paper for consistency.

Consider adding background on deepAntigen's development and methodology details to provide context and clarify its innovations.

Response: Thank you for the reviewer's suggestion. We have updated "atom level" to "atomic level" throughout the text. In addition, we have added a description of the challenges that deepAntigen aims to address and how the methodology resolves these issues, as follows:

“Despite these progresses, challenges remain in accurately identifying T cell antigens for several reasons. First, except for pMTnet, most tools consider either antigen-HLA binding or antigen-TCR interactions when identifying T cell antigens, without comprehensively modeling both sides of the HLA-antigen-TCR complex. Second,

sequence-based methods are insufficient to characterize the refined atomic-level interactions between antigen and TCR⁴¹. Third, the scarcity of high-quality HLA-antigen-TCR complex crystal structures limits the effectiveness of template-based structure prediction tools in accurately modeling the diverse range of HLA-antigen-TCR interactions⁵⁰. Fourth, the independent encoders separately encode the antigen and HLA, which limits the powerful representation ability of AI. In other words, a specific HLA allele can usually present thousands of different antigens, but the embedding of HLA cannot be adaptively adjusted according to the antigen by current methods, leading to a lack of flexibility⁵¹.

To address these challenges, we proposed deepAntigen, a deep learning framework based on graph convolutional networks (GCN), to predict **not only antigen-HLA binding but also antigen-TCR interactions at the atomic level**. Our framework transforms the amino acid sequences of antigens, HLAs and TCRs into graphs, in which nodes correspond to atoms and edges correspond to chemical bonds. deepAntigen utilizes GCN and cross-attention to characterize crucial atomic-level contacts between antigens and HLAs/TCRs. It is worth noting that in each graph convolution layer, an **atomic-level communication mechanism** between antigens and HLAs is introduced, which allows the embedding of a specific HLA can be adaptively adjusted according to the antigen. In addition, to tackle the data deficiency problem of crystal structure, we applied a **few-shot learning strategy by incorporating sequence-level interaction information into atomic-level contact prediction**. We first trained deepAntigen_Seq on abundant sequence-level data to implicitly capture atomic-level information and infer antigen-HLA binding or antigen-TCR interactions. Then, we fine-tuned deepAntigen_Seq using limited HLA-antigen-TCR complex crystal structure data to obtain deepAntigen_Atom, which can predict atomic-level contacts between antigens and HLAs/TCRs...”

Overall, the introduction needs work, and Discuss limitations of the current study and future directions.

Response: We greatly appreciate the reviewer’s valuable comments. We have revised our Introduction according to the reviewer’s suggestions, highlighting the limitations and challenges of current research.

10.Line 77 - 82 Why only the CDR3 sequence? What about other CDRs, and which CDR3 alpha or beta chain for the general audience? The author needs to clarify and elaborate on these specific details.

Response: We apologize for this lack of clarity and appreciate the reviewer’s suggestion. We have added explanations for the CDRs and the TCR $\alpha\beta$ chains, and referenced literature to explain why only the CDR3 β sequence is used. Here is the revision in lines 120-125:

“TCRs are heterodimers composed of one α and one β chain, where the complementarity-determining region 3 (CDR3) is primarily involved in antigen recognition, while the other two CDRs (CDR1 and CDR2) mediate contact primarily with the HLA³¹. In this study, we only considerate the CDR3 of TCR β chain due to **the importance of the CDR3 β in antigen-TCR interactions¹⁸ and the number of antigen-TCR $\alpha\beta$ pairs is much smaller than the antigen-TCR β pairs^{29,54}**.”

11.Line 79 - 78 For the benefit of readers, authors should introduce deepAntigen_seq and deepAntigen_Atom in the introduction.

Response: Thank you for your suggestion. We have added the description about deepAntigen_Seq and deepAntigen_Atom in lines 101-108:

"In addition, to tackle the data deficiency problem of crystal structure, we applied a few-shot learning strategy by incorporating sequence-level interaction information into atomic-level contact prediction. We first trained deepAntigen_Seq on abundant sequence-level data to implicitly capture atomic-level information and infer antigen-HLA binding or antigen-TCR interactions. Then, we fine-tuned deepAntigen_Seq using limited HLA-antigen-TCR complex crystal structure data to obtain deepAntigen_Atom, which can predict atomic-level contacts between antigens and HLAs/TCRs."

12.Line 102 Elaborate on few-shot learning methodology.

Response: Thank you for bringing up this suggestion. Benefiting from the transformation of sequences into atomic graphs, deepAntigen can learn the potential atomic-level contacts as prior knowledge from abundant sequence-level interaction data, enabling it to quickly generalize with limited structural data. This strategy **extends training data beyond structural databases**, differing from other few-shot learning methods such as meta-learning and GANs. We have elaborated on few-shot learning method in lines 148-150 and lines 156-157 of the revised manuscript.

"We applied a few-shot learning strategy by incorporating sequence-level interaction information into atomic-level contact prediction, which benefited from the transformation of sequences into atomic graphs."

"Through this strategy, we can extend the training data beyond structural databases for atomic-level contact prediction."

13.Line 119 - 122. There is a lack of clarity, as each method appears to have utilized three separate datasets. However, the figure demonstrates the use of three distinct data sets for comparison with all the previously mentioned methods. Please clarify

Response: We thank the reviewer for your attention to detail and sincerely apologize for any confusion this may have caused. In the revised paper, we have numbered the different datasets in **Fig.1d** and **Supplementary Table 1,2,3**. Additionally, we have clarified the corresponding datasets information in the context of describing the performance. For example, in **Fig.2a,b,c**, we show the performance of deepAntigen and other methods on the **antigen-HLA I** benchmark test dataset (**Dataset 1**); in **Fig.2d,e,f**, we show the performance of deepAntigen and other methods on the **antigen-HLA II** benchmark test dataset (**Dataset 5**). Here are some examples of the corresponding clarifications in our manuscript. NetMHCPan and MixMHCpred are antigen-HLA I binding prediction methods. NetMHCIIpan and MixMHC2pred are antigen-HLA II binding prediction methods.

*"We compared deepAntigen with other state-of-the-art antigen-HLA I binding prediction methods, including MixMHCpred2.2¹⁵, NetMHCpan4.1²² and TransPHLA²³, using the benchmark test dataset for antigen-HLA I binding (**Dataset 1**). **This dataset consists of 72,936 antigen-HLA I pairs covering 67,951 unique antigens and 95 HLA I alleles** (Supplementary Table 1)."*

*"We further stratify **Dataset 1** by both antigen length and HLA I allele type to analyze their*

impact on the performance.”

“We evaluate the performance of deepAntigen in predicting antigen-HLA II binding on the benchmark test dataset for antigen-HLA II binding (**Dataset 5**), which consists of 50,956 antigen-HLA II pairs covering 46,182 unique antigens and 83 HLA II allele types (Supplementary Table 2).”

d

Task	Dataset	Level	Dataset size	#Antigens	#HLAs or TCRs	Available methods
Antigen-HLA I	Dataset 1	Sequence	72,936	67,591	95	deepAntigen MixMHCpred2.2 TransPhLA NetMHCpan4.1
	Dataset 2	Sequence	27,092	26,974	86	deepAntigen MixMHCpred2.2 TransPhLA NetMHCpan4.1
	Dataset 3	Sequence	608	605	13	deepAntigen
	Dataset 4	Atom	130	93	61	deepAntigen
Antigen-HLA II	Dataset 5	Sequence	50,956	46,182	83	deepAntigen MixMHC2pred2.0 NetMHCIIpan4.2
	Dataset 6	Sequence	19,322	19,091	83	deepAntigen MixMHC2pred2.0 NetMHCIIpan4.2
	Dataset 7	Atom	73	66	18	deepAntigen
Antigen-TCR	Dataset 8	Sequence	1,714	491	1,400	deepAntigen Panpep TEIM ImRex ERGO2
	Dataset 9	Sequence	1,129,028	518	698,034	deepAntigen Panpep TEIM ImRex ERGO2
	Dataset 10	Sequence	384	104	357	deepAntigen Panpep TEIM ImRex ERGO2
	Dataset 11	Atom	134	94	91	deepAntigen AlphaFold3 RosettaFold All-Atom

Fig.1d | Comparison between deepAntigen and other state-of-the-art methods on diverse independent test dataset.

Supplementary Table 1. Summary of antigen-HLA I datasets used in this study.

Dataset	Size	Description	Related figure	Source
Training dataset 1	293,746	Training deepAntigen_Seq to predict antigen-HLA I binding	Supplementary Fig.3, 4, 5, 6, 7, 12, 19, 22 Extended Fig.1a	PMID: 30357391 PMID: 28228285 PMID: 31844290
Dataset 1	72,936	Benchmarking deepAntigen_Seq and other state-of-the-art methods	Fig.2a, b, c	PMID: 30357391 PMID: 28228285 PMID: 31844290
Dataset 2	27,092	Unseen antigens	Extended Fig.1c	PMID: 39414796
Dataset 3	608	Tumor neoantigen presented by HLA I	Extended Fig.3	PMID: 33038342
Dataset 4	130	Antigen-HLA I complex structure used to fine-tune deepAntigen_Atom	Fig.4a	PMID: 29087479 Protein Data Bank

Supplementary Table 2. Summary of antigen-HLA II datasets used in this study.

Dataset	Size	Description	Related figure	Source
Training dataset 2	299,324	Training deepAntigen_Seq to predict antigen-HLA II binding	Supplementary Fig.3, 4, 8, 9, 12, 19, 20, 22 Extended Fig.1b	PMID: 37301199
Dataset 5	50,956	Benchmarking deepAntigen_Seq and other state-of-the-art methods	Fig.2d, e, f	
Dataset 6	19,322	Unseen antigens	Extended Fig.1d	
Dataset 7	73	Antigen-HLA II complex structure used to fine-tune deepAntigen_Atom	Fig.4a	PMID: 36318236 Protein Data Bank
HLA-DR401 antigen library	85,756	Peptides with known 9-mer binding region are used to analyze preference of deepAntigen_Atom	Supplementary Fig.10	PMID: 32887877

Supplementary Table 3. Summary of antigen-TCR datasets used in this study.

Dataset		Size	Description	Related figure	Source
Training dataset 3		62,466	Training deepAntigen_Seq to predict antigen-TCR interacting; at least 5 interacting TCRs for each antigen	Supplementary Fig.3, 4, 11, 12, 19, 22	PMID: 30357391 PMID: 28977646 PMID: 31373607 PMID: 28481982
Dataset 8	Zero-shot	1,714	Each antigen has no overlap with the training dataset; no more than 5 interacting TCRs for each antigen	Fig.3a, d	
	Few-shot	488 + 500	488 records for fine-tuning and 500 records for testing; generated from the zero-shot dataset; 2 interacting TCRs for each antigen are used to fine-tune, remaining interacting TCRs are used to test	Fig.3d	
	Majority	46,464 + 10,640	46,464 records for retraining and 10,640 records for testing; at least 100 interacting TCRs for each antigen are used to retrain	Fig.3d	
Dataset 9		384	Tumor neoantigen recognized by TCR	Supplementary Fig.15	PMID: 33927717
Dataset 10		1,129,028	SARS-CoV-2 antigens and TCRs	Fig.3b Supplementary Fig.13	ImmuneCODE
Dataset 11		133	Antigen-TCR complex structures used to fine-tune deepAntigen_Atom	Fig.4a	PMID: 29087479 Protein Data Bank
pMTnet test dataset		3,542	322 positive pairs and 10 times negative pairs	Fig.3c	PMID: 36003885
Mutation test dataset		134	133 mutated antigens plus one wild-type antigen, and a specific TCR	Fig.4d	PMID: 31695703
Control TCRs		60,333,379	Non-activated TCRs derived from peripheral blood of 587 healthy donors are used to balance interacting TCRs of each antigen through random sampling	/	PMID: 26596423

14.Line 132 - 133 please elaborate

Response: We apologize for this lack of clarity and appreciate the reviewer's suggestion. In the revised manuscript, we have added a description of "*In addition, deepAntigen outperformed current methods on the majority of the 95 alleles (above the diagonal). Other methods perform poorly on certain HLA I alleles (e.g. C*03:03, A*23:01 and C*16:01), but*

deepAntigen consistently maintained good performance (Fig.2c). " and highlighted the representative alleles in **Fig. 2c**. In addition, we provided the original data in the **Source Data** and statistical information in **Table 1,2**.

Table 1. Details about performance comparison on 95 HLA I alleles

Methods	Total alleles	deepAntigen better	deepAntigen worse
deepAntigen	95	/	/
MixMHCpred	95	89	6
NetMHCpan_EL	95	93	2
NetMHCpan_BA	95	93	2
TransPhLA	95	95	0

Table 2. Details about performance comparison on 83 HLA II alleles

Methods	Total alleles	deepAntigen better	deepAntigen worse
deepAntigen	83	/	/
NetMHCIIpan_EL	83	80	3
NetMHCIIpan_BA	83	83	0
MixMHC2pred	83	83	0

15.Line 188 -189 Discuss deepAntigen's scalability and limitations.

Response: We appreciate your feedback on this crucial aspect of our methodology. We acknowledge deepAntigen consumes more time because of sequence-to-graph transformation and utilize multiprocessing to accelerate this process. We added relevant discussion in lines 265-268 and further explained this in the **Discussion** section.

16.Line 216 - 217 elaborate deepAntigen's atom-level contact prediction methodology details.

Response: Thank you for the reviewer's suggestion. We have elaborated on the methodology of atomic-level contact prediction in lines 287-290.

"We froze the parameters of the graph convolutional layers in deepAntigen_Seq. Then we fine-tuned the scoring layer for atoms and the atomic-level contact prediction classifier using limited structural data to obtain deepAntigen_Atom (Methods)."

17.Line 264 Discuss potential applications beyond immunogenic neoantigen identification.

Response: Thank you for your thoughtful review and for helping us to improve our work. Here is the revised description:

"Such subtleties and nuances of these interactions captured by deepAntigen can provide valuable biological insights not only for identifying immunogenic neoantigen but also for engineering TCR-T cells. More importantly, the capacity for atomic-level contacts prediction can yield a broader understanding of the physical principles underlying molecular interactions, such as small molecule drug discovery, protein-ligand docking and antibody-antigen binding affinity."

18.Line 265 Compare deepAntigen with atomic-level prediction methods.

Response: We appreciate the reviewer's valuable suggestions. Following your guidance, we compared deepAntigen with AlphaFold3 (**PMID: 38718835**) and RosettaFold All-Atom

(PMID: 38452047) in atomic-level contact prediction between antigens and TCRs. In our manuscript, deepAntigen predicted the atomic-level contacts between the complex (PDB: 2BNQ), and the results showed good consistency with the true structure. Moreover, the subtle differences in antigen-TCR interactions captured by deepAntigen can explain the impact of single residue mutations on T cell immune responses (Fig.4b,c,e).

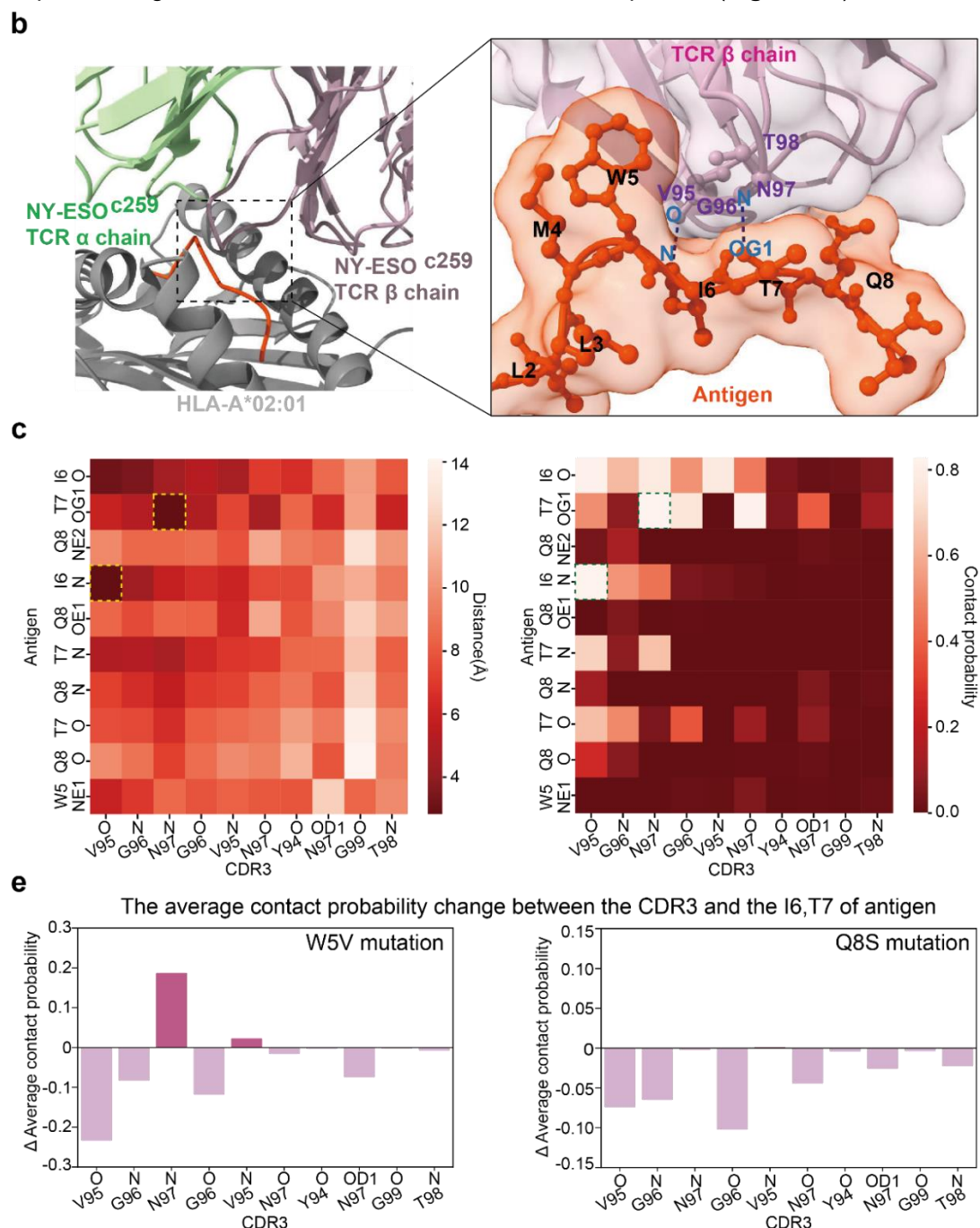
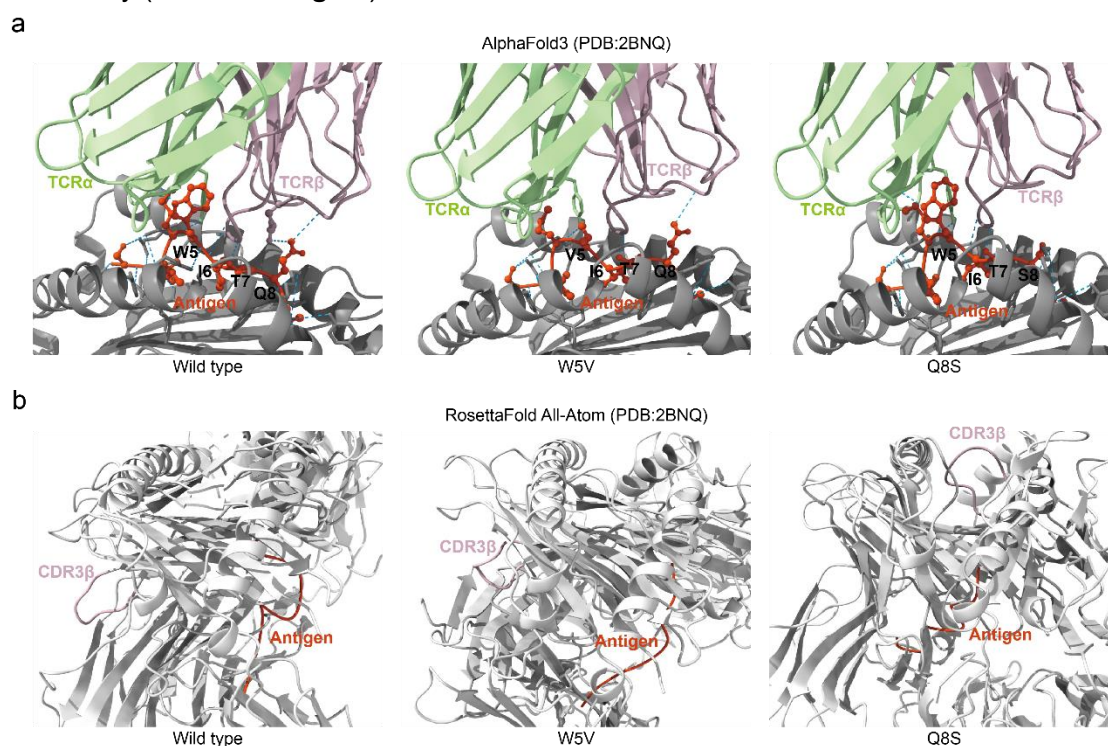


Fig.4 | Atomic-level contacts prediction. **b**, The crystal structure of C259-NYESO complex and the hydrogen bonds between the antigen and the TCR β chain (PDB:2BNQ). The α and β chains of NY-ESOc259 TCR are colored in green and purple, respectively. The HLA is colored in grey and the antigen is colored in orange. The hydrogen bonds between atoms are plotted by ChimeraX as dotted lines. **c**, The pairwise true distance and contact probability of crucial atoms that were identified by deepAntigen. The first line of tick labels indicates the atom name from PDB and the second line is the residue type and

position. I6 and T7 in the antigen are the core of contacting with CDR3. **e**, The average of contact probability change between the core of antigen and each crucial atom of CDR3 after the two mutations (W5V and Q8S).

Therefore, we also wanted to explore whether AlphaFold3 and RosettaFold All-Atom possess similar capabilities. First, we input the wild-type antigen, HLA $\alpha\beta$, and TCR $\alpha\beta$ sequences into the AlphaFold3 web server and the local RosettaFold All-Atom model, using default parameters to predict the structure of the complex. We found AlphaFold3 is able to predict the complex structure very well and correctly identified the hydrogen bonds between the TCR and I6, T7 of the antigen, mainly because it has seen all the structures in the PDB database. However, there are some issues with the details. For example, there is no hydrogen bond between Q8 and the TCR β chain, but AlphaFold3 predicts this as a false positive contact (**Extended Fig.2a**). The performance of RosettaFold All-Atom was disappointing, as it seems incapable of predicting the structure of complex complexes effectively (**Extended Fig.2b**).

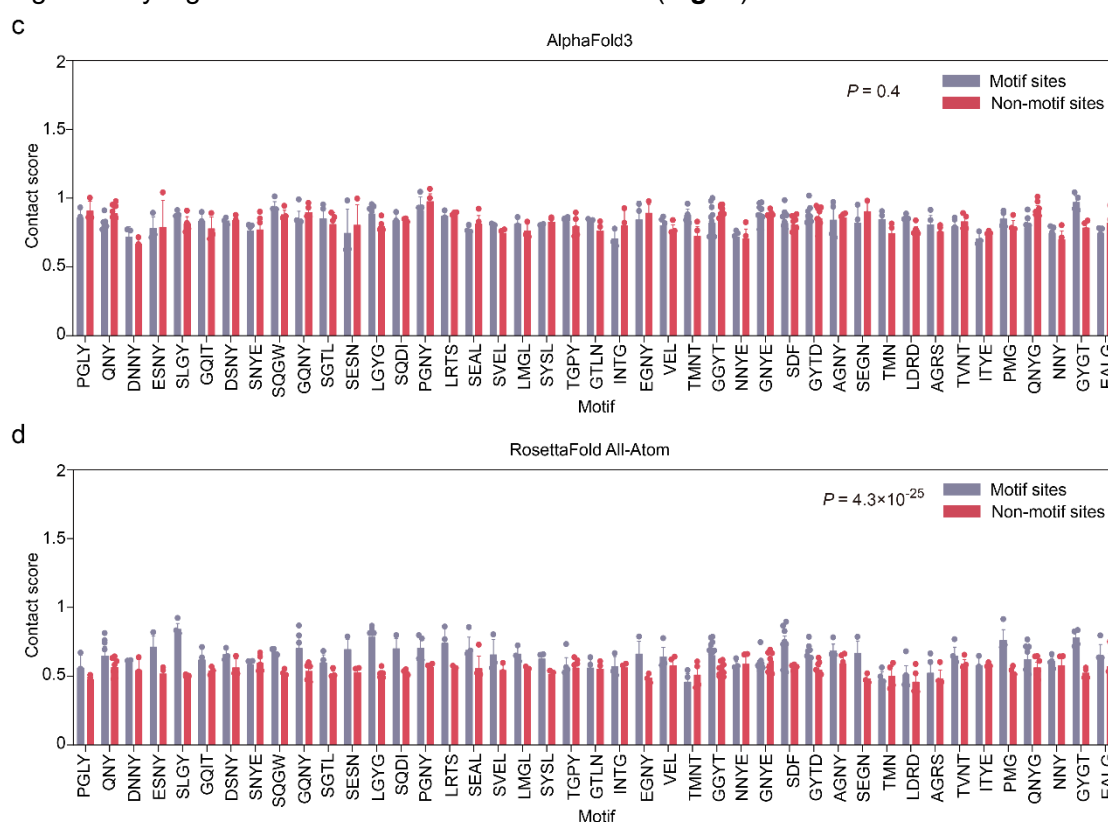


Extended Fig.2 | Analysis of atom-level contacts predicted by AlphaFold3 and RosettaFold All-Atom. a-b, The structures of C259-NYESO complex (wild type, W5V and Q8S) predicted by AlphaFold3 and RosettaFold All-Atom.

In addition, we used AlphaFold3 and RosettaFold All-Atom to analyze the interactions between TCR and two mutant antigens (W5V and Q8S) that were discussed in detail in our paper. W5V and Q8S lead to a significant reduction in the strength of T cell immune responses. In the structure of W5V predicted by AlphaFold3, the hydrogen bond between TCR and I6 was disrupted, weakening the interaction and leading to a reduced immune response. This is consistent with the experimental results and the predictions of deepAntigen. However, in the case of the Q8S mutation, AlphaFold3 predicted positive hydrogen bonds between TCR and both I6 and T7, which contradicts the experimental

results and the predictions of deepAntigen (**Extended Fig.2a**). The structures of W5V and Q8S predicted by RosettaFold All-Atom are still unexpected, even the conformations of other chains are also unstable across different prediction attempts (**Extended Fig.2b**).

We also tested the effectiveness of AlphaFold3 and RosettaFold All-Atom in identifying antigen-specific TCR motifs. Using the antigen and 126 specific CDR3 sequences (whole sequences and other chains were not available) as input, we predicted the structures by AlphaFold3 and RosettaFold All-Atom. The reciprocal average of interatomic distances was used as the contact score between residues, and we analyzed the difference in contact scores between motif sites and non-motif sites. The results showed that AlphaFold3 could not distinguish the motif and non-motif sites (**Extended Fig.2c**). Although RosettaFold All-Atom was able to capture this difference, the magnitude of the difference was very small (**Extended Fig.2d**). Atomic-level contact prediction enables deepAntigen to discover antigen-specific TCR motifs. The predicted contact scores of the motif sites were significantly higher than those of the non-motif sites (**Fig.4f**).



Extended Fig.2 | Analysis of atom-level contacts predicted by AlphaFold3 and RosettaFold All-Atom. c-d, The contact scores of motif sites and ‘non-motif’ sites in the LLLDRLNQL-specific TCR pools predicted by AlphaFold3 and RosettaFold All-Atom.

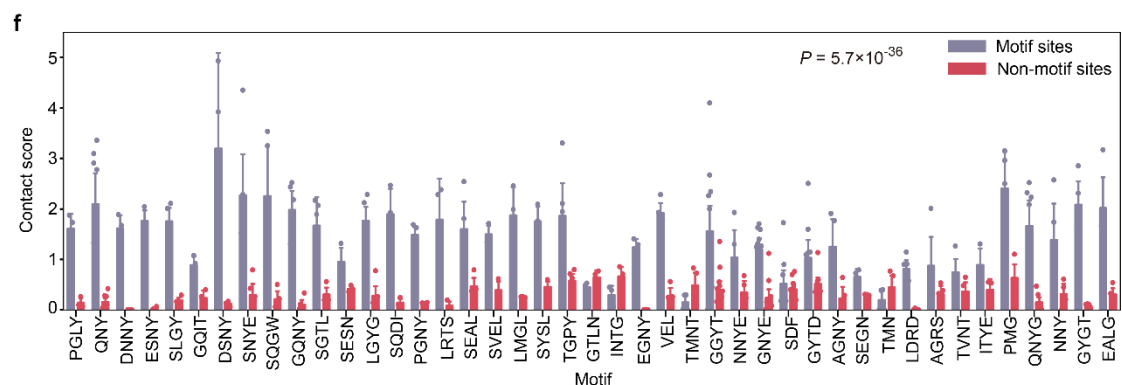


Fig.4 | Atomic-level contacts prediction. **f**, The contact scores of motif sites and ‘non-motif’ sites in the LLLDRLNQL-specific TCR pools predicted by deepAntigen.

19.Line 281- 282 Citation and please elaborate on why it is challenging.

Response: Thanks for your suggestion. We have elaborated on the reasons why it is challenging and added the reference in lines 382-384.

“Nevertheless, the high degree of heterogeneity in tumors poses a significant hurdle for identifying neoantigens as each tumor contains numerous unique putative neoantigens^{10,64}.”

Discussion Comments.

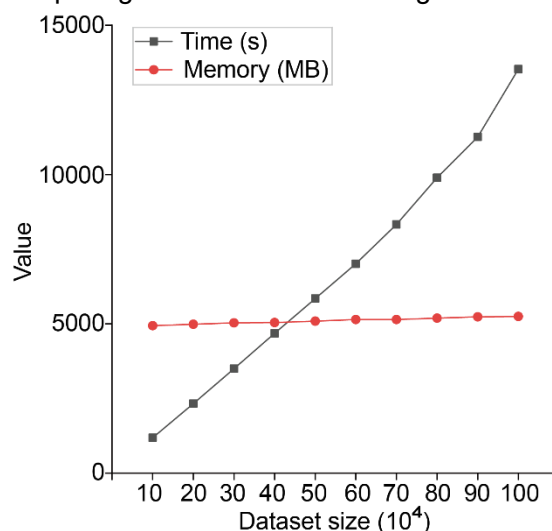
20. Elaborate on deepAntigen's scalability and computational efficiency.

Response: We are truly grateful for the reviewer to raise this comment and followed the reviewer’s suggestion. We tested the consumption of computational time and memory for deepAntigen and the current state-of-the-art methods. (**Supplementary Table 4,6,7**). On the million-scale antigen-TCR dataset, deepAntigen can significantly reduce the time consumed by utilizing multiprocessing (**Supplementary Table 4**). Despite the extremely fast speeds of ImRex, PanPep, and ERGO2, their overall prediction accuracy is comparable to that of random guessing (**Fig.4c**). With default parameters, TEIM encoded TCRs and antigens using **64** processes, whereas deepAntigen consumed a comparable amount of time utilizing **32** processes.

Supplementary Table 4. The time and memory consumption comparison between deepAntigen and other state-of-the-art methods in prediction of antigen-TCR interactions (Dataset size = 1,000,000).

Method	Time (s)	Memory (MB)
ImRex	262	27,232
PanPep	451	1,152
ERGO2	899	6,264
TEIM	4,002	1,048
deepAntigen (32 processes, GPU)	4,084	5,325
deepAntigen (16 processes, GPU)	6,480	5,290
deepAntigen (8 processes, GPU)	13,529	5,248
deepAntigen (8 processes, CPU)	15,173	5,018
deepAntigen (4 processes, GPU)	25,359	5,247
deepAntigen (2 processes, GPU)	50,084	5,239
deepAntigen (1 process, GPU)	98,049	5,240

Additionally, we generated 10 datasets of different sizes, including 100k, 200k, ..., and 1000k. We analyzed the time and memory consumption of deepAntigen as the dataset size increased. The results show that as the dataset size grows, the time consumption of deepAntigen increases linearly, while the memory consumption grows very gradually (**Graph 2**). Therefore, deepAntigen can be scaled to larger datasets.



Graph 2 | Time and memory consumption of deepAntigen as the dataset size increases.

We also compared the time and memory consumption of deepAntigen with the current state-of-the-art methods in antigen-HLA binding prediction (**Supplementary Table 6, 7**). deepAntigen is indeed slower than other methods, but we consider this gap to be acceptable. Taking clinical patient #P1017343 as an example, we constructed a pool containing 6043 antigen peptides based on called variants of the patient. Considering the six HLA alleles (2 HLA-A, 2 HLA-B, and 2 HLA-C) of the patient, we used deepAntigen with 8 processes to predict $6,043 \times 6$ antigen-HLA I pairs. This process took approximately 5 minutes. After the initial screening, we retained 610 peptides with an antigen-HLA binding prediction score greater than 0.5. Then, based on the top 10,000 high-frequency TCR clones of the patient, we used deepAntigen with 8 processes to predict $610 \times 10,000$ antigen-TCR pairs. Despite the need to process 6,100,000 pairs, there are only 610 unique antigens and 10,000 unique TCRs. As a result, after performing the sequence-to-graph transformation 10,610 times, deepAntigen can quickly predict antigen-TCR interactions through CUDA acceleration. It took **4 hours** to predict the interactions of these 6,100,000 antigen-TCR pairs. Therefore, deepAntigen is capable of handling clinical prediction tasks.

Supplementary Table 6. The time and memory consumption comparison between deepAntigen and other state-of-the-art methods in prediction of antigen-HLA I binding (Dataset size = 50,000).

Method	Time (s)	Memory (MB)
deepAntigen (8 processes, GPU)	410	5,597
NetMHCpan	332	6
MixMHCpred	10	6
TransPhLA	12	6,170

Supplementary Table 7. The time and memory consumption comparison between deepAntigen and other state-of-the-art methods in prediction of antigen-HLA II binding (Dataset size = 50,000).

Method	Time (s)	Memory (MB)
deepAntigen (8 processes, GPU)	951	5,701
NetMHCIIpan	620	6
MixMHC2pred	50	6

We have added a discussion on computational efficiency and scalability of the deepAntigen in the **Discussion** section.

“Sequence-to-graph transformation requires more time, so deepAntigen is less efficient in terms of computational speed compared to sequence-based methods (Supplementary Table 4,6,7). However, we can accelerate this process through multiprocessing (Supplementary Table 5), enabling deepAntigen to be scaled for large-scale data in clinical applications (Supplementary Note 4).”

21. Discuss potential integration with experimental validation.

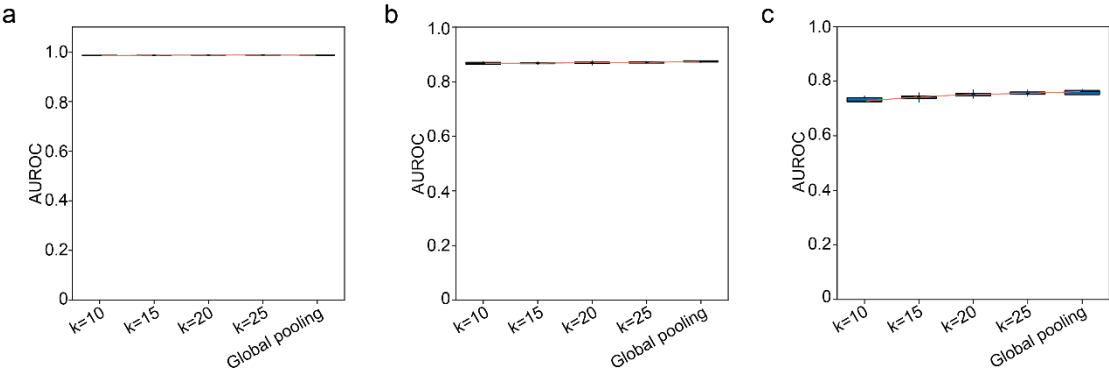
Response: Thank you very much for the comment provided by the reviewer. We have included more discussion on the integration of computational predictions with experimental validation.

“In the validation of neoantigens identified by deepAntigen, our main goal was to validate their immunogenicity using ELISPOT assays. However, we did not further explore the TCRs specific to these neoantigens, which would require antigen-specific T cell sequencing (such as BEAM-T sequencing reported by 10x Genomics⁷⁰) or TCR sequencing before and after antigen stimulation to detect neoantigen-expanded T cell clones². The accumulation of neoantigen-specific TCRs enables iterative algorithm optimization, while the optimized algorithm enhances the efficiency of biological experimental validation, further accelerating immunogenic neoantigen discovery.”

22. Performance of algorithm with other pooling method.

Response: Thank you for your insightful comment. We compared the impact of the Top-k Pooling with the commonly used Global Max Pooling methods on model performance. We have clarified the performance of deepAntigen with Global Max Pooling in **Discussion**.

“Our results indicate that top-k pooling results in slightly lower overall prediction accuracy, approximately 0.01 less, compared to global max pooling (Supplementary Fig.19).”



Supplementary Fig.19 | The impact of top-k pooling on the performance of

deepAntigen a, in prediction of antigen-HLA I binding; **b**, in prediction of antigen-HLA II binding; **c**, in prediction of antigen-TCR binding.

23. Provide system requirement and installation guide document. As a supplementary document.

Response: We thank the reviewer for their suggestions. We have provided an explanation of the system requirements and detailed the installation steps in the **Supplementary File**.

Reviewer #1 (Remarks on code availability):

I have reviewed the code and the GitHub repository, and I am pleased to inform you that everything is in order. The code is well structured, and the repository is organized effectively.

Response: Thank you for your kind words regarding the organization of our code and repository. We appreciate your feedback and will continue to strive to improve the quality and usability of our code and resources.

Responses to Reviewer #2

Reviewer #2 (Remarks to the Author):

Overview

The authors present deepAntigen, a comprehensive computational framework developed to predict interactions between antigens, human leukocyte antigen (HLA) molecules, and T-cell receptors (TCRs) at both the sequence and atomic levels, aiming to advance the precision of neoantigen identification for personalized cancer immunotherapy. The deepAntigen framework comprises two main modules: deepAntigen_Seq and deepAntigen_Atom. deepAntigen_Seq is designed for sequence-level binding prediction between antigens and HLA molecules, as well as between antigens and TCRs. Employing a 10-fold cross-validation approach, it fine-tunes hyperparameters, such as determining the optimal number of top-k atoms for analysis. deepAntigen_Atom targets the prediction of atom-level contacts within antigen-HLA/TCR complexes. Due to the limited data available at this level, it uses a leave-one-out cross-validation strategy, with performance assessed by the Negative Pearson Correlation Coefficient (NPCC) between predicted joint scores and actual atomic distances.

Core methodologies in deepAntigen incorporate Graph Neural Networks (GNNs) with positional encoding, transforming sequences into graph structures where atoms serve as nodes and bonds as edges. The integration of positional encoding embeds spatial information into atom features, enhancing the model's capability to distinguish atoms based on their specific molecular locations. Additionally, deepAntigen_Atom employs fine-tuning and multi-head attention mechanisms, focusing on oxygen and nitrogen atoms essential for non-covalent interactions, which optimizes the model's ability to discern critical atom pairs between antigens and HLA/TCR molecules. Evaluation metrics for deepAntigen include AUROC and AUPR for binary binding predictions, along with NPCC to atom-level accuracy. Benchmarking was conducted against other state-of-the-art antigen-HLA and antigen-TCR prediction tools using standardized parameters to ensure fair comparisons.

Experimental validation of deepAntigen involved patient-derived blood and tumor samples from breast and lung cancer cases, collected under stringent protocols. Peripheral blood mononuclear cells (PBMCs) were isolated and subjected to neoantigen validation. Whole exome sequencing (WES) and RNA sequencing on tumor samples identified mutations for neoantigen prediction, followed by TCR sequencing on patient blood to determine prevalent TCR clonotypes for computational neoantigen interaction analysis. An ELISPOT assay further validated neoantigen immunogenicity by evaluating the ability of neoantigen peptides to stimulate PBMCs to produce interferon-gamma (IFN- γ), indicating an immune response.

In general, the paper is clear, and both the model and mathematical formulations are consistent and thoughtfully constructed. The work is very relevant, original, and timely. Perhaps, adding more detailed discussions on certain areas, such as performance variability across different tasks, the impact of hyperparameters, and potential limitations

in specific settings (e.g., zero-shot learning and long-tail distributions), could strengthen the paper further. Providing this context would address our questions regarding the generalizability and real-world applicability of deepAntigen and help highlight the unique strengths of the approach while transparently acknowledging its limitations.

Response: Thank you very much for your feedback and constructive criticism of our work. We appreciate your comments on the clarity of the performance variability of deepAntigen across different tasks, the impact of hyperparameters and potential limitations in specific settings. We have carefully addressed each of your points and made substantial improvements to the presentation of our method. Below are our detailed responses to your comments.

#Major notes:

Selecting the right model architecture and tuning parameters are critical for the success of any predictive model. The choice of architecture directly impacts how well the model captures the biological complexity of interactions, while hyperparameters affect the model's accuracy and generalization.

Q1: Since there are multiple potential architectures (e.g., GNNs, transformers, recurrent neural networks) and different parameterization approaches, why did the authors choose GCNs with positional encoding over other architectures (e.g., transformers)?

Response: Thank you for your thoughtful comments and questions regarding our choice of model architecture. We appreciate your careful consideration of our work.

The primary objective of deepAntigen is to characterize the **atomic-level interactions** between antigens, HLAs, and TCRs. By transforming amino acid sequences into natural graph structures, where atoms serve as nodes and chemical bonds as edges, we are motivated to utilize GCNs to extract critical atomic features. Positional encoding integrates spatial information into atomic features, enhancing the ability of deepAntigen to differentiate atoms based on their precise specific locations.

Transformer is a popular and powerful model architecture widely used in various domains, particularly in natural language processing and computer vision. It is designed to handle sequential data efficiently by leveraging a self-attention mechanism, which allows the model to weigh the importance of different elements in a sequence dynamically. This enables the transformer to capture long-range dependencies and contextual relationships effectively, overcoming the limitations of traditional recurrent or convolutional neural networks. Transformer represents residues in the sequence as tokens, employing self-attention to capture interactions between residues. **However, it is limited in its ability to capture atomic-level spatial topological information.**

We designed a transformer encoder-based model backbone to build a framework for modeling antigen-HLA binding. In this framework, we establish a 6-layer transformer encoder, with 64 hidden units per layer and 4 heads in the multi-head attention mechanism. We concatenated the antigen and HLA/TCR sequences as the input for this framework. The predictive performance of the framework was evaluated using 10-fold cross-validation. We acknowledge that in antigen-HLA binding prediction, the transformer encoder-based architecture achieves comparable overall accuracy to deepAntigen (**Graph 3a,b**). However, in antigen-TCR interaction prediction, deepAntigen significantly outperforms the

a Antigen-HLA I binding

Model	AUROC (approx.)
Transformer encoder	0.99
deepAntigen	0.99

b Antigen-HLA II binding

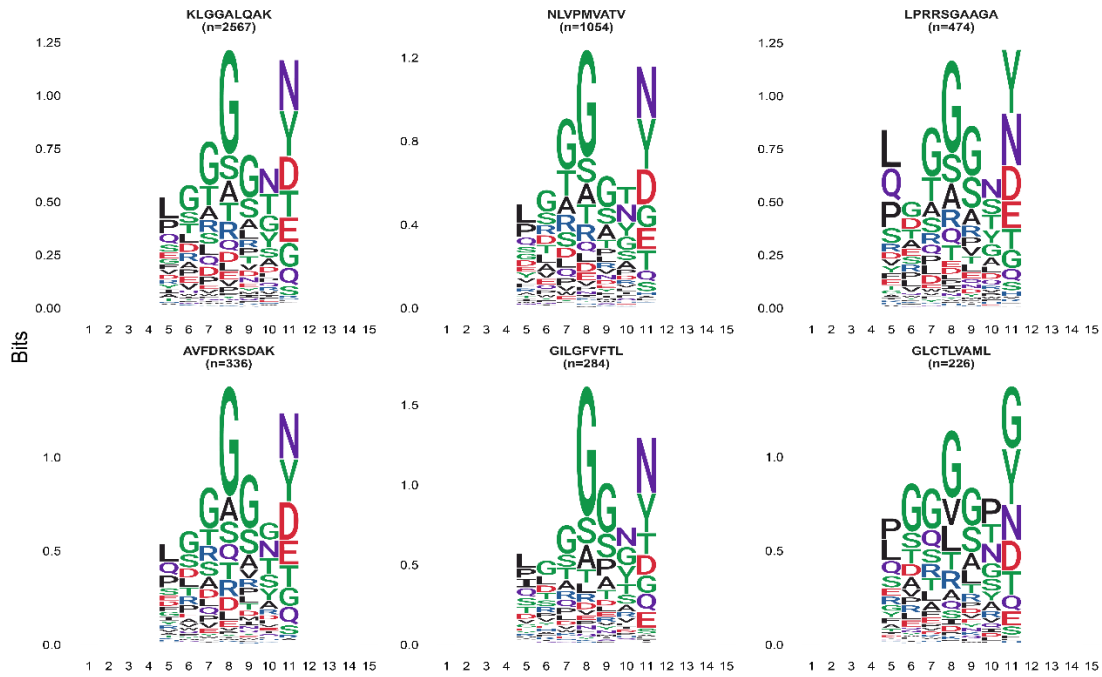
Model	AUROC (approx.)
Transformer encoder	0.87
deepAntigen	0.87

c Antigen-TCR interaction

Model	AUROC (approx.)
Transformer encoder	0.70
deepAntigen	0.75

$P = 1.7 \times 10^{-7}$

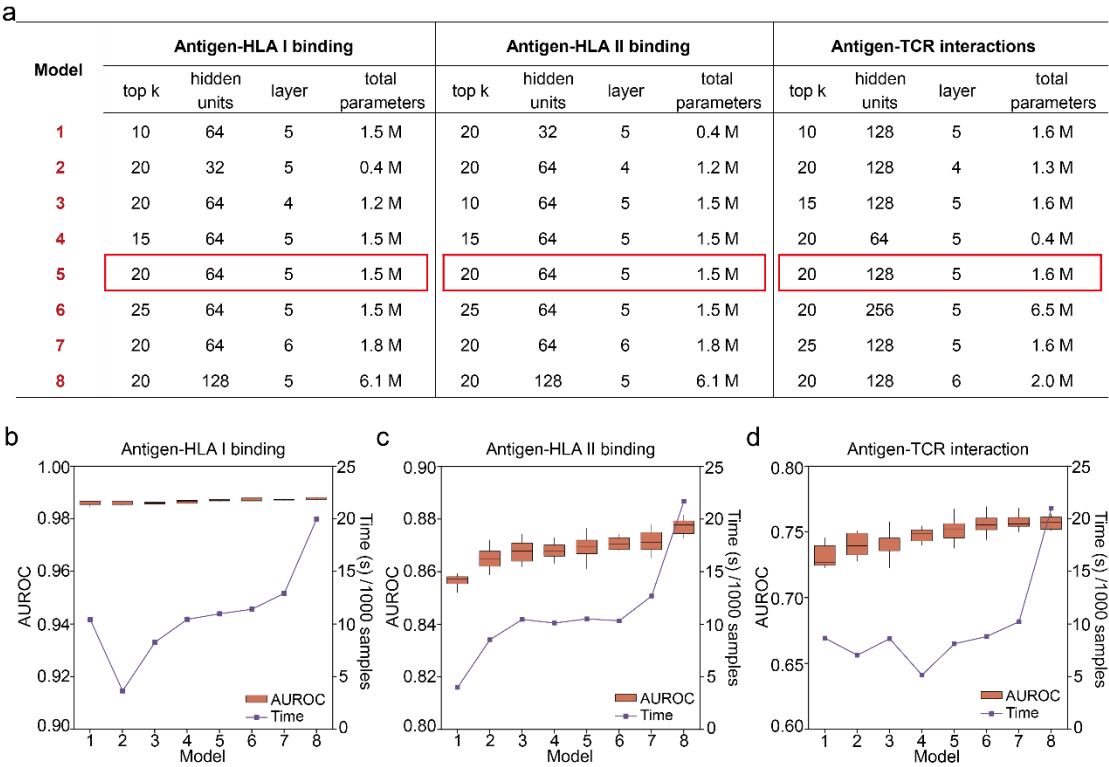
Supplementary Fig.8 (partial) | Sequence logo of positive HLA class II-restricted antigens



Graph 4 | Sequence logo of antigen-specific TCRs.

Q2: How did the authors determine the optimal top-k value and other hyperparameters?

Response: Thank you for pointing out this important aspect of hyperparameters determination. We did not actually perform an explicit hyperparameter search, as a comprehensive hyperparameter search is resource-intensive and may offer limited performance improvement. We experimented with different combinations of hyperparameters (top k-value, the number of hidden units, and the number of graph convolutional layers) across various tasks. We observed that hyperparameters had a relatively minor influence on the model's performance. However, increasing the number of hidden units and graph convolutional layers led to greater model complexity. Balancing performance and complexity, we carefully selected appropriate hyperparameters tailored to each task: for antigen-HLA I and antigen HLA II binding, a top k-value of 20, 64 hidden units and 5 graph convolutional layers; for antigen-TCR interaction, a top k-value of 20, 128 hidden units and 5 graph convolutional layers (**Supplementary Fig.3**).



Supplementary Fig.3 | Performance evaluation of deepAntigen with different hyperparameters in predicting antigen-HLA I binding, antigen-HLA II binding and antigen-TCR interactions through 10-fold cross-validation. a, Different combinations of hyperparameters and the corresponding number of parameters. **b-d**, The box plots of AUROC and the line charts of consumed time per 1000 samples for deepAntigen with different hyperparameters.

We have provided a detailed explanation of the hyperparameter selection in the **Methods** section. Please see lines 701-710. We once again extend our sincere gratitude for your feedback, which has been invaluable in improving the clarity of our work.

One of the primary concerns in bioinformatics and immunoinformatics is whether a model trained on a specific dataset can generalize to novel or diverse data. On lines 373 to 375 the authors say: "For each positive pair, a non-binding antigen was generated as a negative pair by cleaving a segment of equal length from the protein where the binding antigen is located."

Simply cleaving a segment of equal length from the same protein does not guarantee it will be a non-binder. The same protein can have multiple peptides capable of binding the same HLA, with varying affinities:

Blum, J. S., Wearsch, P. A., & Cresswell, P. (2013). "Pathways of antigen processing." Annual Review of Immunology, 31, 443–473;

This review discusses antigen processing pathways and how certain peptides, even from regions initially considered non-binding, can still be presented, especially when proteins

are processed and presented via non-canonical pathways.) Without experimental validation or computational predictions specifically indicating that these segments do not bind, labeling them as non-binding is mainly speculative.

Neefjes, J., Ovaa, H., & Liu, Y. C. (2020). "A guide to major histocompatibility complex (MHC) class I and MHC class II antigen processing pathways in infectious diseases, autoimmune diseases, cancers, and vaccines." *Frontiers in Immunology*, 11, 1–16.

This paper shows the idea that antigen presentation is not limited to a single binding motif, and that various regions, including secondary and tertiary sites, may be involved in immune recognition.

Therefore, if any of these "negative" samples do in fact bind to HLA, they introduce false negatives into the dataset. This could confuse the model during training, as it may learn to associate features of binding segments with non-binding labels, potentially reducing model accuracy. The presence of false negatives can also skew performance metrics, making the model appear less sensitive to true binders. Plus, biological datasets, especially those involving antigens, HLAs, and TCRs, are notoriously heterogeneous and may not fully capture the diversity seen in real-world applications.

Q3: deepAntigen is capable of handling unseen or rare antigens and TCRs? (This would determine the model's utility in clinical and research settings where encountering novel epitopes is common)

Response: Thank you for your valuable insights regarding the benchmarking of deepAntigen. We appreciate your suggestion to include additional benchmarks, particularly handling unseen antigens and TCRs. As you mentioned, the model's ability to handle unseen antigens and TCRs directly determines its clinical practicality. Indeed, we have evaluated the performance of deepAntigen to handle unseen antigens. We sincerely apologize for not describing this clearly, which may have caused confusion.

In the **zero-shot setting** (Dataset 8), the antigens in the test dataset were all not present in the Training dataset 3. In addition, the **COVID-19 antigens** (Dataset 9) were exogenous and the **tumor neoantigens** (Dataset 10) are rare for the immune system. During data processing, we also retained only antigens that were not present in the Training dataset 3. Therefore, the tests on Dataset 8, Dataset 9 and Dataset 10 were conducted on unseen antigens. Our results demonstrated that deepAntigen outperforms current state-of-the-art methods in handling unseen antigens (**Fig.3a,c and Supplementary Fig.13,15**).

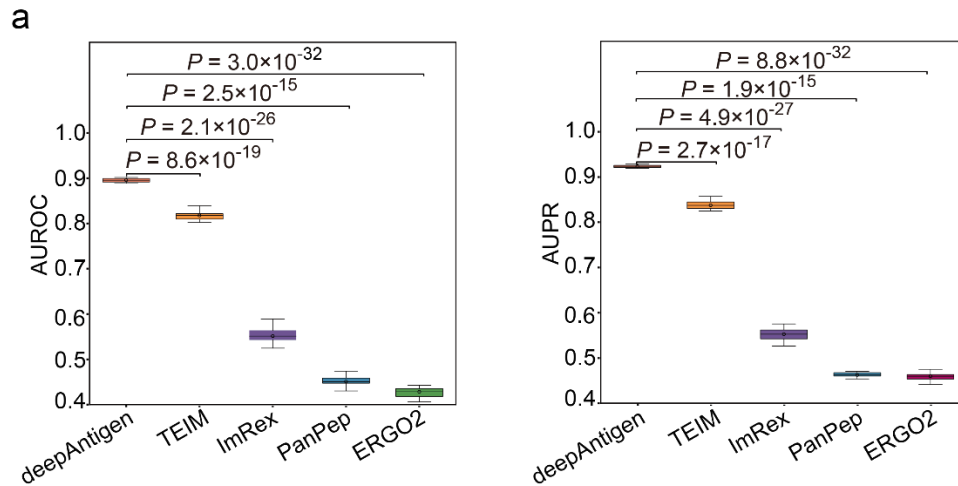


Fig.3 | Performance evaluation of deepAntigen in predicting antigen-TCR interactions. a, Box plots AUROC and AUPR of deepAntigen and state-of-the-art on the reconstructed zero-shot datasets with 20 group negative samples (Dataset 8).

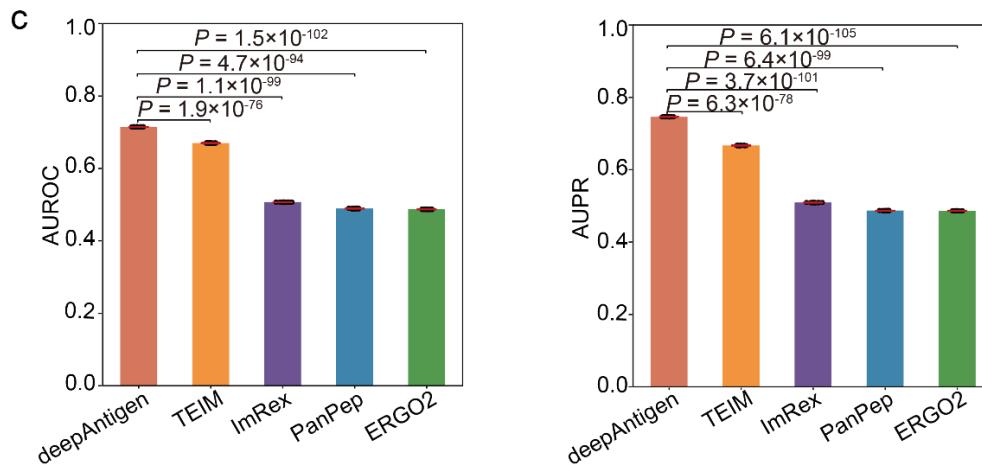
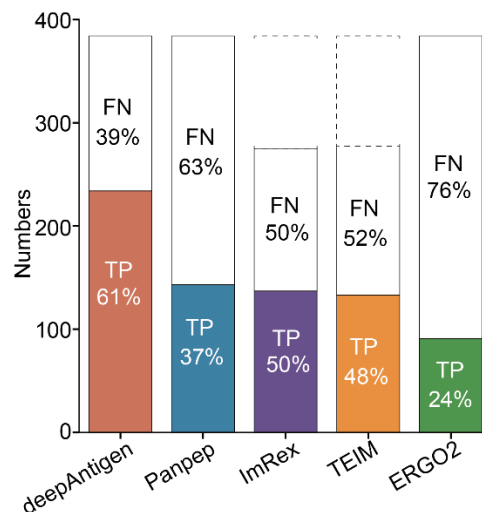
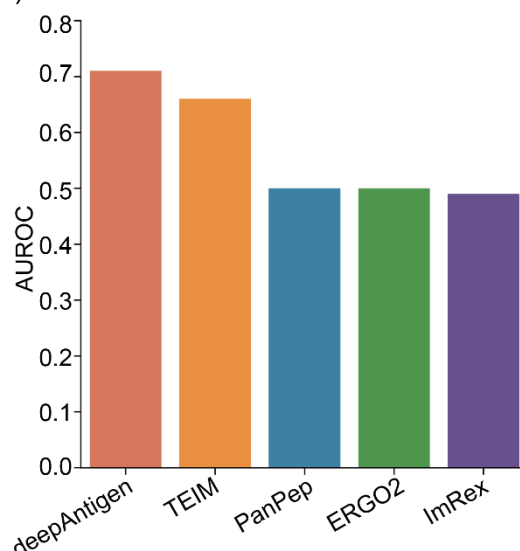


Fig.3 | Performance evaluation of deepAntigen in predicting antigen-TCR interactions. c, AUROC and AUPR for deepAntigen, TEIM, PanPep, ImRex and ERGO2 on the large independent COVID-19 test dataset (Dataset 9) covering 1,129,028 SARS-CoV-2 antigen-TCR pairs. Black points are results of twenty times repeated evaluations with different negative samples. Red lines are error bars. P values were calculated by a one-sided paired t-test.



Supplementary Fig.15 | The true positive (TP) and false negative (FN) rate of deepAntigen and other state-of-the-art methods on the neoantigen dataset (Dataset 10).

To further validate our approach and address your concerns regarding deepAntigen's capacity to handle unseen TCRs, we further filtered Dataset 10 by excluding TCRs present in the training dataset, generating a dataset where both antigens and TCRs are unseen. The test results indicate that deepAntigen still outperforms current state-of-the-art methods (Supplementary Fig.13).

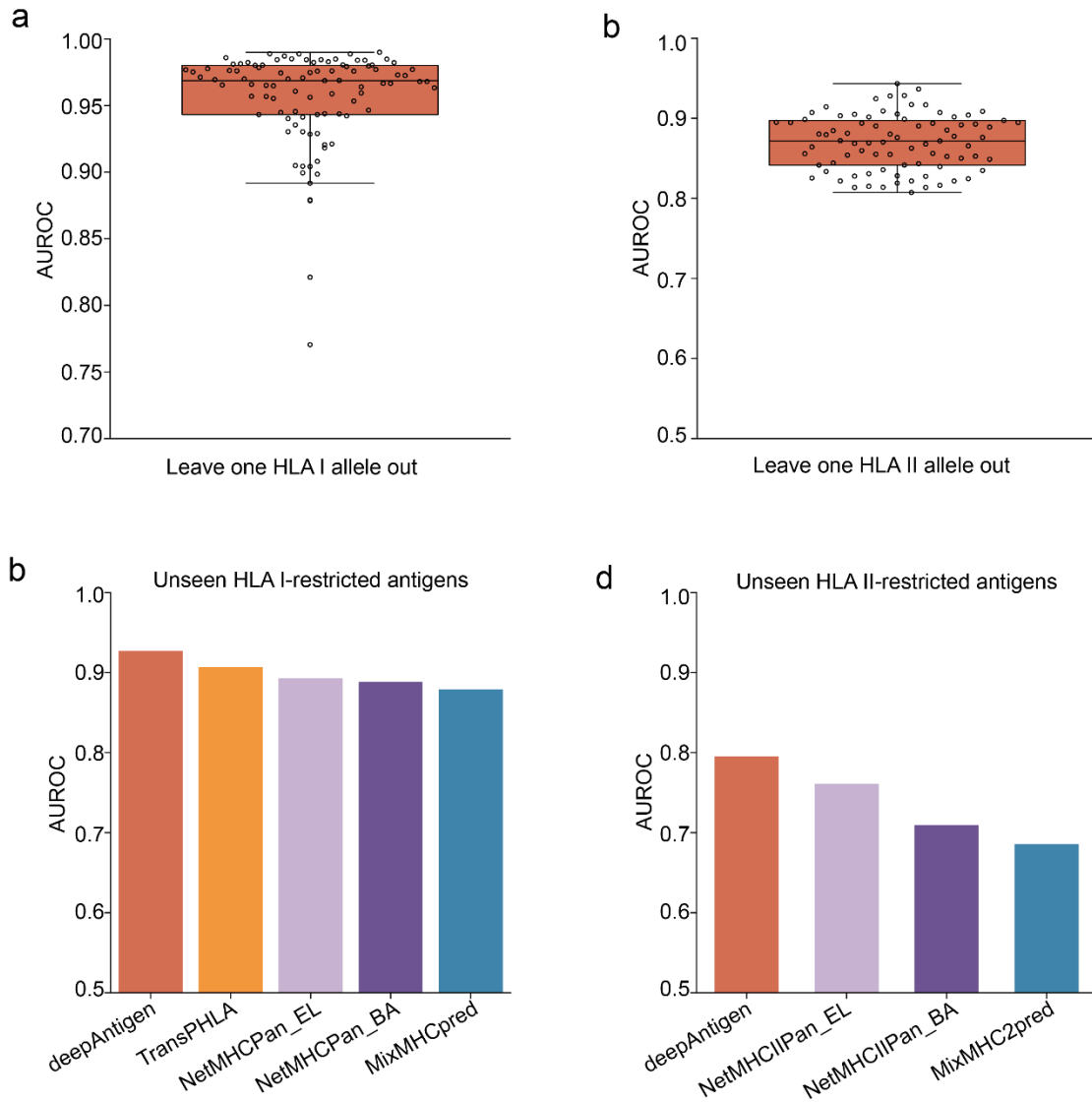


Supplementary Fig.13 | Performance of deepAntigen and other state-of-the-art methods on the dataset where antigens and TCRs are both unseen.

Q4: How does deepAntigen generalize to antigens, HLAs, and TCRs not represented in the training data?

Response: Thank you for your significant question regarding the generalization of deepAntigen to antigens, HLAs and TCRs.

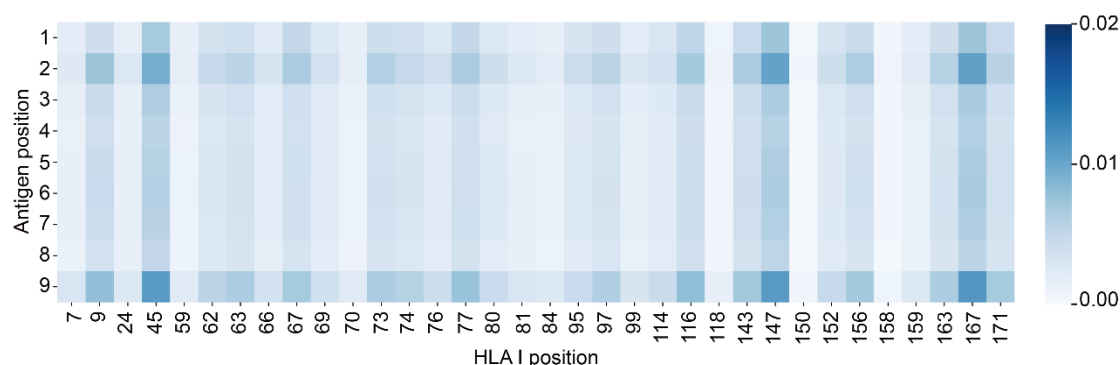
In our response to your Question 2, we have discussed the generalization of deepAntigen to unseen antigens and TCRs. Here, we also evaluated its performance in predicting antigen-HLA I binding on Training Dataset 1 and antigen-HLA II binding on Training Dataset 2 using **leave-one allele out cross-validation**. Our results showed that deepAntigen achieved **a median AUROC of 0.97** in predicting antigen-HLA I binding and **a median AUROC of 0.87** in predicting antigen-HLA II binding (**Extended Fig.1a,b**). In addition, we evaluated the performance of deepAntigen and current state-of-the-art methods on the Dataset 2 and Dataset 6, where **all antigens were not present in the training datasets** of any methods mentioned above. As shown in **Extended Fig.1c,d**, deepAntigen had the highest AUROC in both prediction of unseen antigen-HLA I binding and unseen antigen-HLA II binding, demonstrating the powerful predictive performance of deepAntigen.



Extended Fig.1 | Performance evaluation of deepAntigen generalized to unseen HLAs and antigens. **a**, The box plot of the AUROC for deepAntigen through leave-one HLA I allele out cross-validation on Training Dataset 1. **b**, Performance comparison between deepAntigen and current state-of-the-art methods on the unseen antigen-HLA I dataset (Dataset 2). **c**, The box plot of the AUROC for deepAntigen through leave-one HLA II allele out cross-validation on Training Dataset 2. **d**, Performance comparison between deepAntigen and current state-of-the-art methods on the unseen antigen-HLA II dataset (Dataset 6).

When designing deepAntigen, we considered its generalization ability to unseen antigens, HLAs, and TCRs. Rather than building specialized models for predicting antigen-TCR interactions or antigen-HLA bindings targeting specific antigens or HLAs, even though this might yield better benchmark results, we refrained from doing so. The reason is that such models would lose their ability to generalize to unseen antigens and HLAs. Instead, we aimed to utilize deepAntigen to discover more generalized patterns of antigen-HLA binding or antigen-TCR binding from datasets containing diverse antigens, HLAs, and TCRs. Taking antigen-HLA I binding as an example, we verified that deepAntigen indeed learned generalized patterns of antigen-HLA I binding by analyzing attention weights. As shown in the **Supplementary Fig.7**, we summarized the attention weights of deepAntigen

focusing on the binding sites of all positive antigen-HLA pairs in the Training dataset 1. The attention map strongly highlighted the second and the last positions of the antigen. These positions are referred to as anchor positions, which are supported by previous studies (PMID: 1323878; 1709722). Additionally, we also analyzed the capability of deepAntigen to recognize the binding core of HLA II-restricted antigens (Supplementary Fig.10) and antigen-specific TCR motifs (Fig.4f) in our manuscript. Therefore, the generalized binding patterns captured by deepAntigen can support its generalization to unseen antigens, HLAs, or TCRs.



Supplementary Fig.7 | Visualization of attention weights in heatmaps for antigen-HLA I binding. The heatmap showed the contribution of each position on the antigen and HLA sequence.

We have incorporated the relevant results into our manuscript. Once again, thank you for your valuable suggestions, which have greatly contributed to enriching the content and improving the quality of our article.

Q5: How do you address the risk of false negatives in negative samples?

Response: We sincerely appreciate the significant effort you have put into explaining the issue of potential false negatives when using randomly segmented peptides as negative samples. We understand your concern about that false negatives could confuse the model during training and skew performance metrics.

First, although false negative peptides may be generated, the possibility and proportion of such peptides are very low (PMID: 10358753; 25576665). Bennink et al. (PMID: 10358753) observed that “only approximately 1/200 of potential determinants bind at greater than the threshold affinity ($K_d > 500$ nM) associated with immunogenicity” and Nasser et al. (PMID: 25576665) pointed that “Cathepsins present in processing compartments contribute by cutting and trimming of the protein antigens. Somehow during these processes few peptides from many peptides that can bind to MHC II are selected to represent the antigen to T cells”.

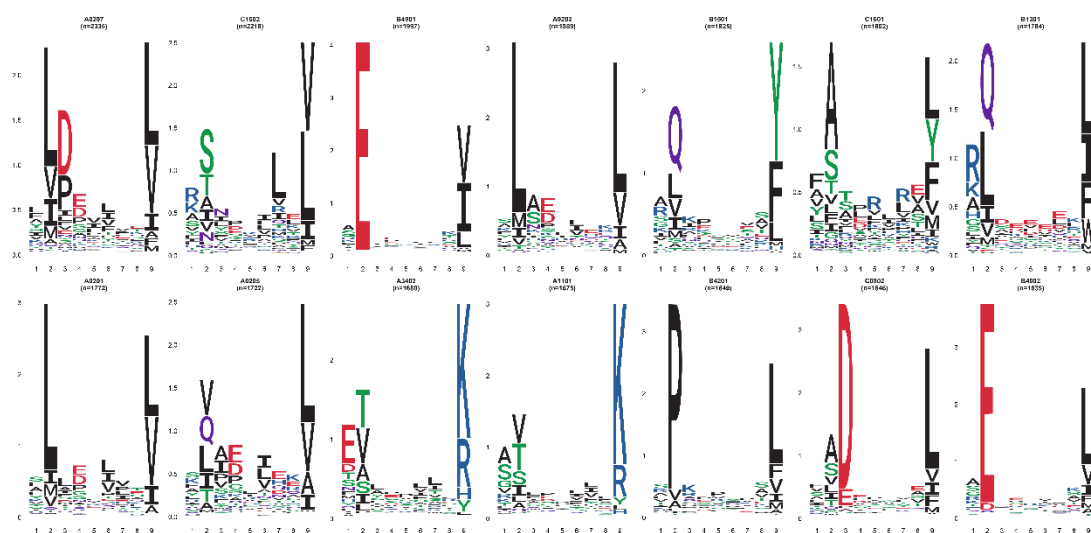
Second, previous study (PMID: 32406916; 31611696; 37023751) randomly sampling peptides from natural protein as negative background and calculate the rank of positive peptides, approximately 1%-5% Rank peptides are recognized positive peptides. This indirectly suggests that the likelihood of false negatives among randomly generated peptides is relatively low.

Third, we ensured that negative peptides were not redundant with positive peptides for a specific HLA by conducting thorough checks during the generation of negative peptides. In addition, we analyzed the motifs of 9-mer HLA I peptides and 9-mer HLA II peptides deconvoluted by Modec⁴. We observed distinct position-specific motifs in positive

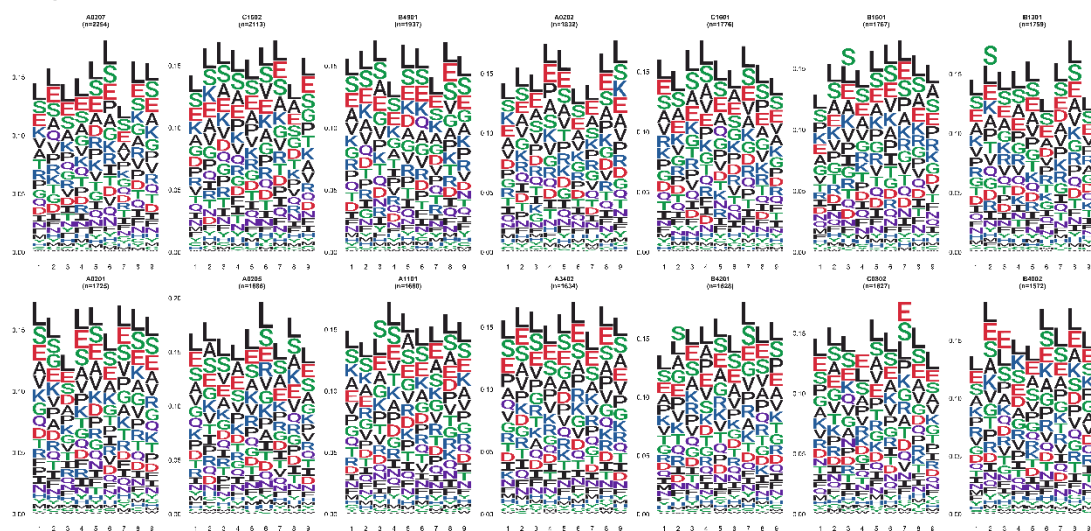
peptides, whereas no clear patterns were identified in negative peptides (**Supplementary Fig.5,6,8,9**). Therefore, even if unexpected false negatives did occur, they would not affect the differences in data distribution between positive and negative peptides. Deep learning models with strong generalization capabilities do not intentionally learn such noise from the dataset. Our results indicated that deepAntigen captured the differences in motifs between positive and negative antigen sequences, which focused on the 2nd and 9th position of the antigen in prediction of antigen-HLA I binding (**Supplementary Fig.7**).

Finally, we carefully examined whether the negative pairs in the test dataset conflicted with the positive pairs from training datasets of other methods. However, we found no unexpected conflicts. Therefore, the performance metrics remain unbiased.

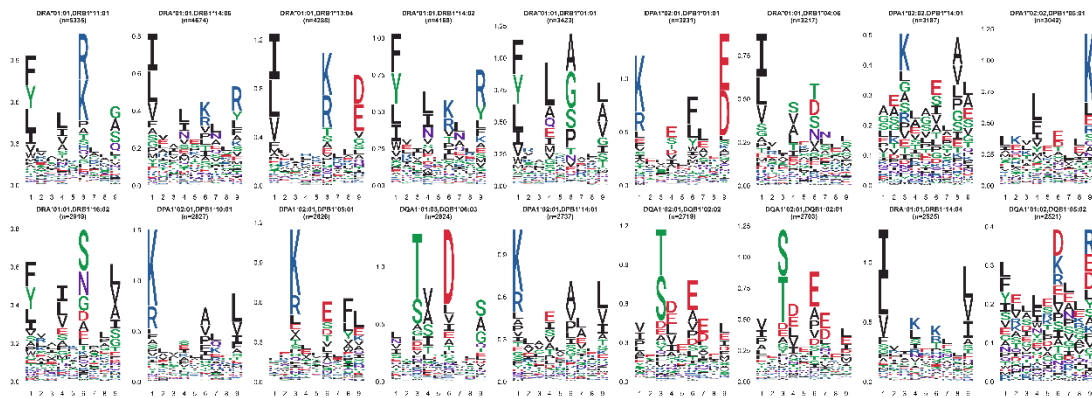
We sincerely hope that this detailed explanation, coupled with the provided data visualization, addresses your question and deepens your understanding of how we control the risk of false negatives. We also incorporated these analyses into our **Data curation** to clarify that the risk of false negatives is low and will not compromise the performance of deepAntigen. Please see lines 499-503 and lines 517-518.



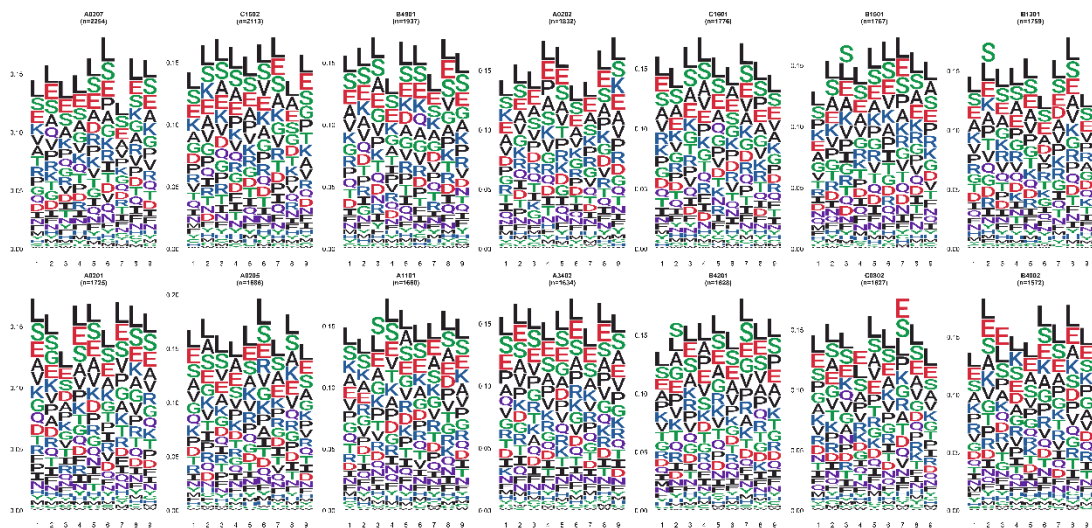
Supplementary Fig.5 (partial) | Sequence logo of positive HLA class I-restricted antigens.



Supplementary Fig.6 (partial) | Sequence logo of negative HLA class I-restricted antigens.



Supplementary Fig.7 (partial) | Sequence logo of positive HLA class II-restricted antigens.



Supplementary Fig.8 (partial) | Sequence logo of negative HLA class II-restricted antigens.

Computational models, especially those involving deep learning and complex datasets, can be resource-intensive. Since deepAntigen uses advanced techniques such as CUDA acceleration and GNNs, I wonder about its practicality and efficiency, especially if intended for high-throughput or clinical settings.

Q6: How computationally intensive is deepAntigen, and can it be scaled to larger datasets or applied in real-time clinical settings?

Response: We are truly grateful for the reviewer raising the concern about the computational efficiency and scalability of deepAntigen. deepAntigen transforms antigen, HLA, and TCR sequences into atomic graphs, utilizing graph convolutional neural networks to characterize the atomic-level interactions between antigens and HLAs/TCRs. While deepAntigen has enhanced its capabilities in predicting antigen-HLA binding and antigen-TCR interactions, it indeed consumes more computation time and memory. This is because the sequence-to-graph transformation requires more time and memory. However, deepAntigen supports accelerating this process through multiprocessing. The time consumed in the computation process of GNNs is actually not much, as the atomic graph

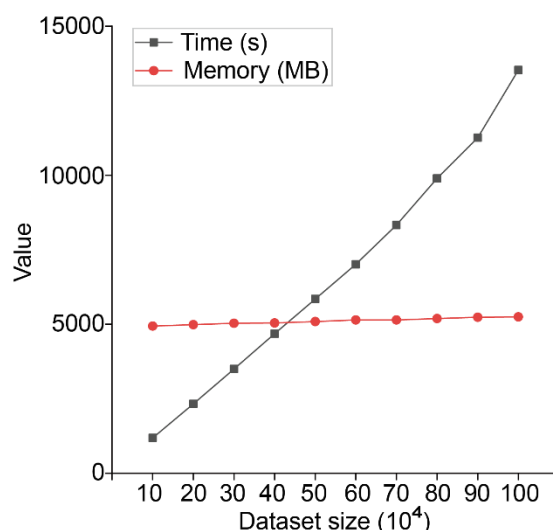
structures of antigens, HLAs, and TCRs are simple. Whether to utilize CUDA for acceleration of GNNs is optional. With the continuous improvement of computing power, most hosts now include multiple CPU cores, making the clinical application of deepAntigen feasible.

We tested the consumption of computational time and memory for deepAntigen and the current state-of-the-art methods. (**Supplementary Table 4,6,7**). On the million-scale antigen-TCR dataset, deepAntigen can significantly reduce the time consumed by utilizing multiprocessing (**Supplementary Table 4**). Despite the extremely fast speeds of ImRex, PanPep, and ERGO2, their overall prediction accuracy is comparable to that of random guessing (**Fig.4c**). With default parameters, TEIM encoded TCRs and antigens using **64** processes, whereas deepAntigen consumed a comparable amount of time utilizing **32** processes.

Supplementary Table 4. The time and memory consumption comparison between deepAntigen and other state-of-the-art methods in prediction of antigen-TCR interactions (Dataset size = 1,000,000).

Method	Time (s)	Memory (MB)
ImRex	262	27,232
PanPep	451	1,152
ERGO2	899	6,264
TEIM	4,002	1,048
deepAntigen (32 processes, GPU)	4,084	5,325
deepAntigen (16 processes, GPU)	6,480	5,290
deepAntigen (8 processes, GPU)	13,529	5,248
deepAntigen (8 processes, CPU)	15,173	5,018
deepAntigen (4 processes, GPU)	25,359	5,247
deepAntigen (2 processes, GPU)	50,084	5,239
deepAntigen (1 process, GPU)	98,049	5,240

Additionally, we generated 10 datasets of different sizes, including 100k, 200k, ..., and 1000k. We analyzed the time and memory consumption of deepAntigen as the dataset size increased. The results show that as the dataset size grows, the time consumption of deepAntigen increases linearly, while the memory consumption grows very gradually (**Graph 2**). Therefore, deepAntigen can be scaled to larger datasets.



Graph 2 | Time and memory consumption of deepAntigen as the dataset size increases.

We also compared the time and memory consumption of deepAntigen with the current state-of-the-art methods in antigen-HLA binding prediction (**Supplementary Table 6,7**). deepAntigen is indeed slower than other methods, but we consider this gap to be acceptable. Taking clinical patient #P1017343 as an example, we constructed a pool containing 6043 antigen peptides based on called variants of the patient. Considering the six HLA alleles (2 HLA-A, 2 HLA-B, and 2 HLA-C) of the patient, we used deepAntigen with 8 processes to predict $6,043 \times 6$ antigen-HLA I pairs. This process took approximately 5 minutes. After the initial screening, we retained 610 peptides with an antigen-HLA binding prediction score greater than 0.5. Then, based on the top 10,000 high-frequency TCR clones of the patient, we used deepAntigen with 8 processes to predict $610 \times 10,000$ antigen-TCR pairs. Despite the need to process 6,100,000 pairs, there are only 610 unique antigens and 10,000 unique TCRs. As a result, after performing the sequence-to-graph transformation 10,610 times, deepAntigen can quickly predict antigen-TCR interactions through CUDA acceleration. It took **4 hours** to predict the interactions of these 6,100,000 antigen-TCR pairs. Therefore, deepAntigen is capable of handling clinical prediction tasks.

We have clarified the scalability and computational efficiency in **Discussion**. Please see lines 453-456.

Supplementary Table 6. The time and memory consumption comparison between deepAntigen and other state-of-the-art methods in prediction of antigen-HLA I binding (Dataset size = 50,000).

Method	Time (s)	Memory (MB)
deepAntigen (8 processes, GPU)	410	5597
NetMHCPan	332	6
MixMHCpred	10	6
TransPHLA	12	6,170

Supplementary Table 7. The time and memory consumption comparison between deepAntigen and other state-of-the-art methods in prediction of antigen-HLA II binding (Dataset size = 50,000).

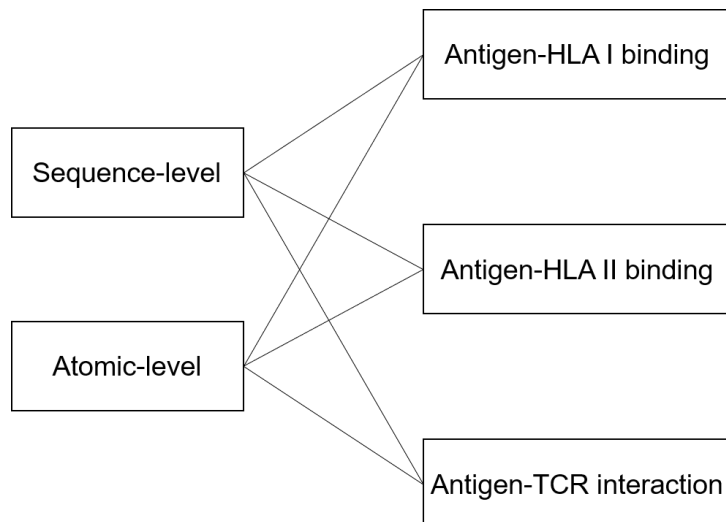
Method	Time (s)	Memory (MB)
deepAntigen (8 processes, GPU)	951	5,701
NetMHCIIPan	620	6
MixMHC2pred	50	6

I believe this would benefit from ICS or Tetramer Staining, that would allow for more detailed characterization of responding T-cell subsets but might miss low-frequency responses or require more resources and preparation. Also, ELISA or Luminex would provide quantitative cytokine data but without single-cell resolution, potentially reducing sensitivity and making it harder to detect rare antigen-specific responses for further studies. However, IFN- γ ELISPOT is a well-suited choice for validating neoantigen immunogenicity in this study, especially given its sensitivity and direct relevance to T-cell activation.

Benchmarking new methods against established ones is crucial for demonstrating their utility and uniqueness. While deepAntigen was compared to ten other methods, I still have question on the specific strengths and weaknesses relative to these models. If deepAntigen shows variable performance across different tasks, I would like to see more clear statements about the trade-offs inherent in the tool design or if certain tasks are better handled by other specialized methods.

Response: We sincerely apologize for not clearly explaining the uniqueness and utility of deepAntigen. The design of deepAntigen is inspired by the catch bond (**PMID: 30053426; 30711376;35389803**), in which the subtle atomic-level interactions are crucial for activating T-cell immune responses. Existing computational frameworks are all sequence-based methods that overlook atomic-level interactions. deepAntigen transforms antigen, HLA, and TCR sequences into atomic graphs to characterize antigen-HLA binding and antigen-TCR interactions at the atomic level. deepAntigen demonstrated better generalization capability for unseen antigens, HLAs, and TCRs compared to other methods (**Extended Fig.1, Fig.3 and Supplementary Fig.13**). The atomic-level characterization can overcome the challenges encountered by sequence-based methods, such as the impact of sequence length and sequence similarity on model performance (**Fig.2b,e and Supplementary Fig.12**). Atomic-level representations can encompass information at the sequence level. For example, deepAntigen primarily focuses on the second and ninth residues of HLA I-restricted antigens (**Supplementary Fig.7**), which correspond to HLA I-restricted antigen sequence motifs (**Supplementary Fig.5**). In addition, atomic-level contact prediction is a unique capability of deepAntigen, which allows for the atomic-level interpretation of mutation effects (**Fig.4b,c,e**). However, sequence-to-graph transformation introduces additional computational complexity, resulting in higher time and resource consumption compared to other methods (**Supplementary Table 4,6,7**).

deepAntigen can handle six different tasks (**Graph 5**), but these tasks are essentially the same, so we have integrated them into a unified framework. In reality, there are significant differences between the data for the different tasks, so deepAntigen is trained on the data from each task separately within this unified framework. In other words, deepAntigen contains **six distinct sets of parameters to handle different tasks**, and these **parameter sets are independent and do not affect each other**. Therefore, we did not discuss the trade-offs in the tool design in our manuscript. We apologize once again for any confusion caused and appreciate your suggestion.



Graph 5 | Different tasks involved in deepAntigen.

Q7: “Then deepAntigen was applied to identify neoantigens that could be presented by HLA and recognized by TCR.” Which TCRs? The authors use patient data that includes TCRseq, but they do not mention having obtained matched antigen-TCR data from these samples. The validation here seems to be regarding the capacity to induce the activation of PBMCs, but without a direct link to a specific TCR sequence.

Response: Thanks for your suggestion and we sincerely apologize for any misunderstanding this description may have caused. We applied deepAntigen to identify candidate neoantigens with the potential to activate T cell immune responses. Specifically, we ranked the neoantigens based on antigen-HLA binding prediction scores and antigen-TCR interaction-derived immunogenicity scores. Here, the interaction between antigens and TCRs is predictive and has not been experimentally validated.

The main goal of our biological experiments is to validate the immunogenicity of the neoantigens identified by deepAntigen through ELISPOT. Obtaining matched antigen-TCR data is an exciting development, as it will expand the current neoantigen database. However, after the ELISPOT experiments, the cell viability was insufficient to guarantee successful sorting using flow cytometry or magnetic separation. Two other feasible approaches are antigen-specific T cell sequencing (such as BEAM-T sequencing reported by 10x Genomics) or TCR sequencing before and after antigen stimulation to detect neoantigen-expanded T cell clones. Both of these approaches can obtain paired antigen-TCR data, but they cannot directly quantify the immunogenicity of neoantigens. Instead, the immunogenicity can only be indirectly estimated through the clonal frequency of T cells. Therefore, in this study, we validated the immunogenicity of neoantigens only through ELISPOT. In future research, we will explore which TCRs can recognize neoantigens and contribute matched neoantigen-TCR data. We sincerely apologize for the misunderstanding caused by our wording and sincerely appreciate you for pointing the direction for our future experimental plans.

We have revised our description in lines 417-418 and added discussion on how to obtain paired neoantigen-TCR data in lines 474-481.

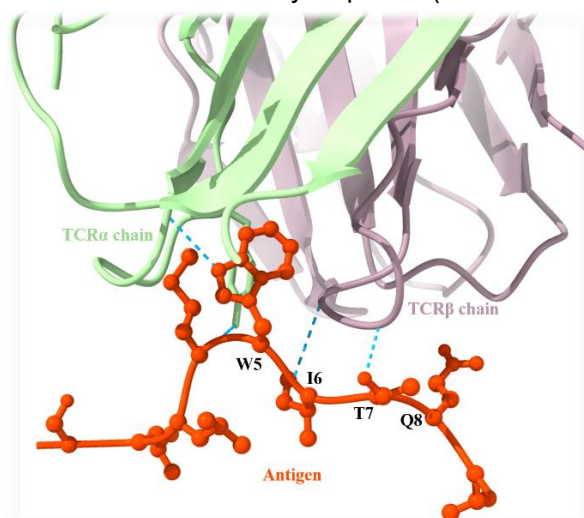
“Then deepAntigen was applied to identify candidate neoantigens with the potential to activate T cell immune responses.”

“In the validation of neoantigens identified by deepAntigen, our main goal was to validate their immunogenicity using ELISPOT assays. However, we did not explore the TCRs specific to these neoantigens, which would require antigen-specific T cell sequencing (such as BEAM-T sequencing reported by 10x Genomics⁷⁰) or TCR sequencing before and after antigen stimulation to detect neoantigen-expanded T cell clones². The accumulation of neoantigen-specific TCRs enables iterative algorithm optimization, while the optimized algorithm enhances the efficiency of biological experimental validation, further accelerating immunogenic neoantigen discovery.”

Q8: 1) The authors briefly acknowledge in their Discussion the limitation of their methods regarding focus on CDR3 beta only. Is there a technical limitation of deepAntigen that would prevent it from working with the other CDR loops, or is really just a matter of lack of data for paired alfa-beta sequences and cognate peptides? 2) In addition, when doing the structural analysis of the TCR peptide bonds, like in the case of NY-ESO, do the authors also see contacts with CDR3 alpha? 3) In the authors perspective, how much the TCR-peptide accuracy could be improved by adding information of the other CDR loops?

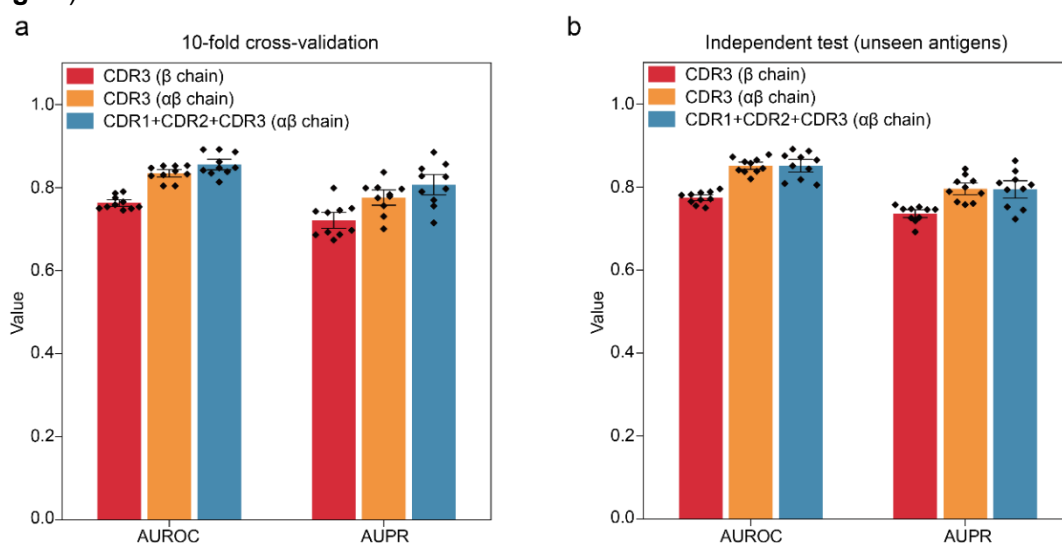
Response: Thank you very much for your feedback and constructive criticism regarding the consideration of TCR sequences. 1) A previous study of GLIPH reported that both CDR3 α and CDR3 β chain are involved in contacting with antigen through many 3D structure analysis cases while the CDR3 β rather than CDR3 α are necessarily required (**PMID: 28636589**). Additionally, the number of antigen-TCR $\alpha\beta$ pairs is much smaller than the antigen-TCR β pairs and the data quality of antigen-TCR $\alpha\beta$ pairs is uneven (**PMID: 33990328**). Therefore, we only consider CDR3 β in deepAntigen. In fact, there is no technical limitation of deepAntigen that prevent it from working with the other CDR loops. deepAntigen can be readily extended to handle CDR3 $\alpha\beta$ by concatenating the interacting map between CDR3 α and the peptide with the interacting map between CDR3 β and the peptide. We provided the reference code of deepAntigen how to model antigen-CDR3 $\alpha\beta$ interactions on Github.

2) We found that the alpha chain may also contact with the antigen, as seen in cases such as NY-ESO (**Graph 6**). However, many 3D structure analysis cases show no CDR3 α contact. Therefore, CDR3 α is not necessarily required. (**PMID: 28636589**).



Graph 6 | H-bonds between antigen and TCR $\alpha\beta$ in the case of NY-ESO.

3) We collected antigen-CDR3 $\alpha\beta$ pairs from VDJdb and found that **only 3,011 antigen-CDR3 $\alpha\beta$ pairs** had high confidence (Score>0). After remove TCRs without known TRAV/TRBV genes, we retained **1,734 antigen-CDR3 $\alpha\beta$ pairs**, which contains 345 unique antigens, 1,377 unique CDR3 α sequences and 1,536 unique CDR3 β sequences. In this dataset, 49 antigens with more than 5 known interacting TCRs were used as the training dataset, and the remaining 296 antigens were used as the independent test dataset. Then we collected 700,660 TCRs (including paired CDR3 $\alpha\beta$ sequence and TRAV, TRBV gene) derived from healthy donors as control TCRs, obtained from a previous study (**PMID: 39216000**). Non-interacting TCRs in negative antigen-CDR3 $\alpha\beta$ pairs were randomly sampled from the control TCRs, which had a very low probability of encountering TCRs interacting with the given antigen. CDR1 and CDR2 sequences are assembled using the TRAV/TRBV gene reference sequences from IMGT. Then, we split the training dataset into 10 subsets and established three different scenarios: using only CDR3 β , using paired CDR3 $\alpha\beta$ and using paired CDR $\alpha\beta$. Due to the small size of the dataset, we adjusted the parameters of deepAntigen to retrain, using 3 graph convolutional layers with 16 hidden units. The results of the 10-fold cross-validation indicated that **incorporating TCR α chains and other regions can improved the performance of deepAntigen** in prediction of antigen-TCR interactions. However, the results of the independent test indicated that the inclusion of other regions did not significantly improve performance (**Supplementary Fig.21**).



Supplementary Fig.21 | The performance improvement with inclusion of TCR α chains and other regions. **a**, The performance comparison of 10-fold cross-validation. **b**, The performance comparison of independent test.

We have clarified the inclusion of TCR α and other loops how impacts on the performance of deepAntigen in predicting antigen-TCR interactions in **Discussion**. Please see lines 473-474.

#Minor notes:

N1. Intro: "But it is currently an unsolved problem: most of identified T cell antigens fail to elicit T cell immune responses⁹." The text seems to merge the challenges of HLA-binding

prediction with those of TCR-specificity prediction. Although deepAntigen targets both sides of the peptide interaction in a TCRpMHC complex, this is not often the case for other tools that target these different problems independently. In this case, is the aforementioned unsolved problem understood by the authors as mostly a fail of epitope discovery or of TCR specificity prediction? What is the core contribution of deepAntigen in this context, and what are secondary yet useful capabilities?

Response: Thank you for bringing up this suggestion. From the perspective of immune response activation, both antigen presentation by HLA and recognition by TCR are crucial processes. From a computational prediction standpoint, antigen-HLA binding prediction is relatively straightforward because there are HLA-specific sequence patterns, and existing antigen-HLA prediction algorithms have achieved good performance. In contrast, predicting antigen-TCR interactions is more challenging. We believe that the lack of consideration for antigen-TCR interactions in neoantigen screening is a possible reason for the failure of most antigens to activate T cell immune responses. By applying deepAntigen to antigen-TCR binding prediction in neoantigen screening, we can further refine the selection of candidate antigens, thereby increasing the success rate of identifying immunogenic neoantigens. We believe that TCR specificity prediction is a core contribution of deepAntigen, but in fact, deepAntigen has shown improved performance in both tasks.

N2. "However, developing computational methods to predict antigen-HLA binding and antigen-TCR interaction at the atom level and further to infer the immunogenicity of the antigen, has not been explored yet." There is a long history of efforts to conduct structure-based HLA-binding prediction, going back to Rognan et al., 1999 (10579827) and Altuvia et al., 2004 (15542371), and more recently with tools such as PANDORA (38143769) and APE-Gen (38415656). Although such structure-based methods have not fulfilled their potential to outperform sequence-based methods, in great part due to shortcomings of scoring functions, this is important work that should not be characterized as a problem that "has not been explored yet". Even structure-based immunogenicity prediction has been attempted before (31555277, 39276157).

Response: Thank you for highlighting previous studies relevant to our work and we apologized for this not rigorous description. We added a brief overview about these methods and limitations of structure-based methods in the **Introduction**.

"Incorporating information about the HLA-antigen-TCR complex structure could provide new insights into T cell specificity^{41–46}. Recently, revolutionary AI tools like RosettaFold All-Atom⁴⁷ and AlphaFold3⁴⁸ have provided atomic-level insights into studying general molecular interactions. RosettaFold All-Atom⁴⁷ utilizes SE(3)-Transformers to generate 3D backbone from homologous structure templates and performs all-atom expansion according physicochemical principles. AlphaFold3⁴⁸ leverages a diffusion network to transform an atomic cloud into the accurate molecular structure through iterative refinement. But they still struggle to capture the relative contributions of details such as hydrogen bonding, electrostatic forces, and hydrophobic effects to the overall stability and specificity of interactions^{49,50}."

"Third, the scarcity of high-quality HLA-antigen-TCR complex crystal structures limits the effectiveness of template-based structure prediction tools in accurately modeling the diverse range of HLA-antigen-TCR interactions⁵⁰."

N3. "With minimum modifications, deepAntigen can be used to mimic interactions between nucleic acids, proteins, and other molecules at the atom level". Is mimic the proper term here? I believe the authors mean to say that deepAntigen can be adapted to produce similar types of binding predictions for other molecules.

Response: We are truly grateful for the reviewer to raise this problem. Indeed, what we intended to convey is that deepAntigen can also be used to predict other types of molecular interactions. We have replaced "mimic" with "predict."

N4. Although the manuscript is generally well-written and clear, there are several passages that contain typos/mistakes. Given the high standards of the Journal and potential high impact of this publication, a more detailed revision of the text regarding style/grammar would be appreciated. Some examples of sentences that could be improved/fixed include:

- Abstract: "Precise identification of T cell antigens in silico is crucial for the development of cancer mRNA vaccine(s)"

Response: We apologize for this mistake and appreciate the reviewer's careful examination. We have made the correction.

- Intro: "T cell antigens refer to short peptide epitopes presented on Major Histocompatibility Complex (MHC) molecules, also known as Human Leukocyte Antigens (HLA) in humans, and recognized by the T cell receptor (TCR) on the surface of T cells." Although there is nothing wrong with this sentence, and I understand its meaning, it has some ambiguity. It could be interpreted as if the antigens are in the surface of the T cells.

Response: We greatly appreciate the comment provided by the reviewer and apologize for the lack of clarity here. We have revised this sentence to "*T cell antigens refer to short peptide epitopes presented on Major Histocompatibility Complex (MHC) molecules, also known as Human Leukocyte Antigens (HLA) in humans, and recognized by the T cell receptor (TCR).*"

- "Thus, deepAntigen could get out of trouble encountered by other methods, ..."

Response: Thanks for your correction. We have revised this sentence to: "*Thus, deepAntigen could overcome the trouble encountered by other methods...*".

- "We analyzed this antigen library and identified motifs at positions 1, 4, 6, and 9, where are corresponding to pockets on the surface of HLA II molecules^{34,35}"

Response: We thank the reviewer for bringing this confusion to our attention. We have clarified the relationship between the motifs and the pockets. This is revised description: "*We identified motifs at positions 1, 4, 6, and 9, which are accommodated by the pockets on the surface of HLA II molecules^{56,57}.*"

- "Atom-level contact prediction enable deepAntigen to discovery antigen-TCR motifs."

Response: We are truly grateful to the reviewer for pointing out this mistake and apologize for our oversight. We have corrected this sentence to: "*Atomic-level contact prediction enables deepAntigen to discover antigen-specific TCR motifs.*"

- “inspired by the biological fact that T cell immune response is determined by the catch-bond^{22,27–29}.” This sentence should be revised both in terms of style and immunology. The role of catch bonds are significant and interesting, but authors should not state as if the presence of a single catch-bond is the only determining feature of T-cell activation.

Response: We appreciate the comment raised by the reviewer. The sentence is indeed not rigorous enough. We have rephrased the sentence. The revised sentence is as follows: *“inspired by the biological fact that catch-bonds play a crucial role in the activation of T cell immune responses.”*

We believe these changes have strengthened the manuscript and made it more engaging for the reader. Thank you for guiding us in refining our language usage to enhance the overall quality of the paper.

This manuscript was co-reviewed by one of the reviewers and an Early Career Researcher. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer #2 (Remarks on code availability):

Installation on a Google Colab environment with GPUs went well. The documentation was incredibly helpful in supporting my understanding of the package’s features and applications in real-world scenarios. It provided clear guidance on best practices and effectively explained different use cases, making it much easier to start applying the package.

Response: Thank you for your kind words regarding the reproducibility of our code and the clarity of our online documentation. We are pleased to hear that you find them clear and useful for users applying the package. We appreciate your feedback and will continue striving to improve the quality and usability of our code and resources.

Responses to Reviewer #3

Reviewer #3 (Remarks to the Author):

#Overview

This article presents a graph convolutional network-based T-cell antigen identification method, which has a significant innovative contribution to atom-level antigen-HLA/TCR contact prediction. This is the first work showing the ability to recognize and learn atom-level information about the antigen, which potentially stimulates new biological findings in the related area.

Response: Thank you sincerely for your encouragement and invaluable comments regarding our work. We have carefully considered to each of your suggestions and incorporated them into our manuscript. We are confident that these revisions have greatly improved the quality and clarity of our paper. Below, you will find detailed responses to each of your comments.

Overall, this manuscript is well-organized. However, there are some remaining concerns that I am glad to highlight for the authors' notice:

#Main concerns:

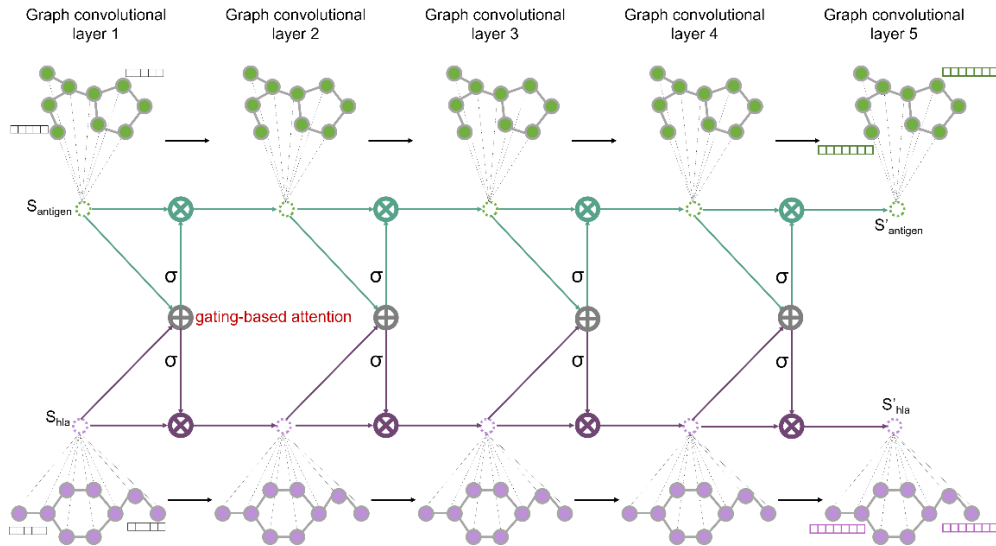
1. Insufficient novelty in terms of the machine learning methods: The proposed deepAntigen uses mostly a standard graph convolutional network (GCN). I can see it is certain that the proposed deepAntigen model presents some novelty in considering the atom-level feature to enhance the model's robustness in predicting, but the model design itself is still very basic. Technically speaking, the model design and training strategy are more like a real-world application demonstration of the well-known GCN, rather than a methodological contribution. I suppose the lack of a specific model design is one of the reasons of the observed limited improvement in predicting antigen-HLA binding is not significant (with only a 0.01 improvement in AUC for the recall compared to the MixMHCpred), as well as in other tasks.

Response: We greatly appreciate the reviewer's critical suggestions, which have been invaluable in enhancing the methodological contributions and overall quality of our paper. Indeed, we have provided a basic framework for atomic-level identification of T cell antigen, primarily to demonstrate the feasibility of this approach. However, as you rightly pointed out, the performance improvement in the antigen-HLA binding prediction task was not significant, mainly due to the lower sensitivity at high specificity compared to current state-of-the-art methods. This suggested that our model may not have sufficiently learned the binding patterns of antigen-HLA pairs.

We analyzed a possible reason, which is about the architecture of deepAntigen. deepAntigen is a representation-based model. Specifically, the embedding of an HLA allele remains fixed through GCN-based encoders, despite the fact that a single HLA allele can bind to thousands of antigens. The binding patterns are mainly captured by the cross-attention mechanism. The full potential of GCN has yet to be realized.

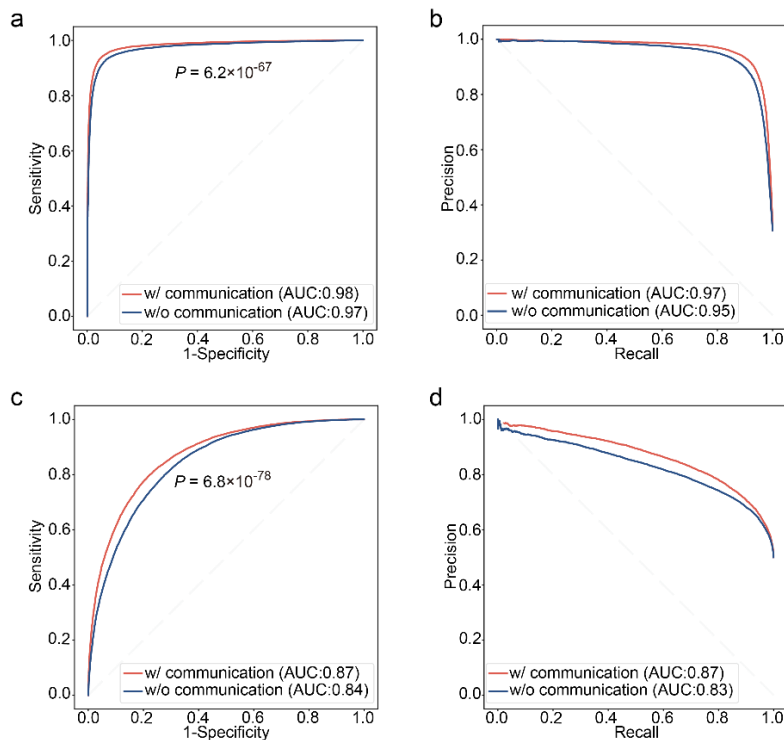
To address this problem, we developed an atomic-level communication mechanism between the antigen and the HLA (**Supplementary Fig. 1**). The beneficial effect of this mechanism is that it allows the atomic-level information of the antigen and HLA to mutually

adjust during the early stages of graph convolution, thereby enhancing the fitting capability of deepAntigen. Specifically, we introduced a virtual super node for the graph and utilized gating-based attention to enable both intra-graph and inter-graph information transfer. In each iteration of graph convolution, atomic nodes first exchange information with the super node, followed by information exchange between the super node and another super node. Consequently, the embedding of an HLA allele can be adaptively adjusted according to the antigen, and vice versa.



Supplementary Fig.1 | Communication mechanism.

To demonstrate the effectiveness of the atomic-level communication mechanism, we conducted ablation experiments on antigen-HLA I and antigen-HLA II binding prediction tasks. Our results showed that this mechanism significantly improves the performance of deepAntigen (Delong test $P = 6.2 \times 10^{-67}$ for antigen-HLA I binding prediction; $P = 6.8 \times 10^{-78}$ for antigen-HLA II binding prediction) (**Supplementary Fig.2**).



Supplementary Fig.2 | Performance of deepAntigen with/without communication mechanism. a-b, The receiver operating characteristic (ROC) curves and Precision-Recall (PR) curves of deepAntigen on the antigen-HLA I benchmark test dataset (Dataset 1). **c-d,** The ROC curves and PR curves of deepAntigen on the antigen-HLA II benchmark test dataset (Dataset 5). ‘w/’ is abbreviation of ‘with’ and ‘w/o’ is abbreviation of ‘without’.

We have updated the results comparing deepAntigen with current state-of-the-art methods. The overall performance of deepAntigen significantly outperforms other algorithms, and its sensitivity at high specificity is also comparable (**Fig.2a,d**).

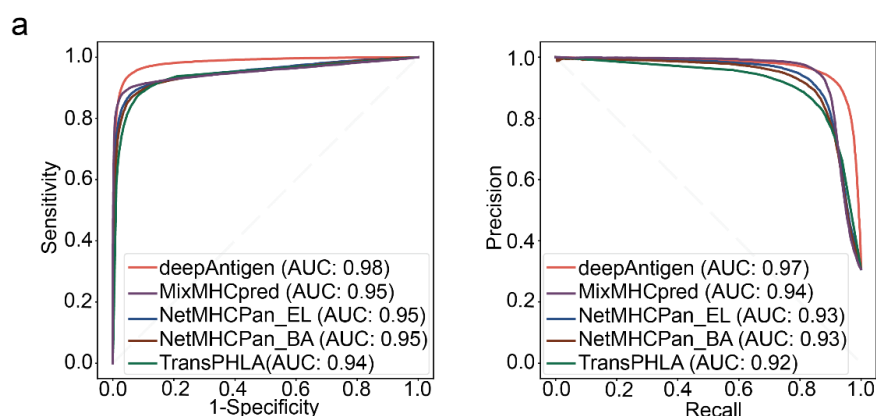


Fig.2 | Performance evaluation of deepAntigen in predicting antigen-HLA binding. a, The ROC curves, PR curves of deepAntigen and state-of-the-art methods on the antigen-HLA I benchmark test dataset (Dataset 1).

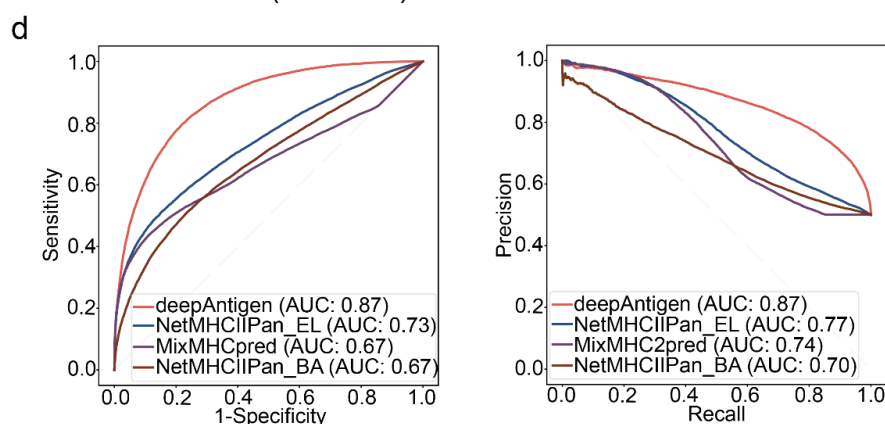


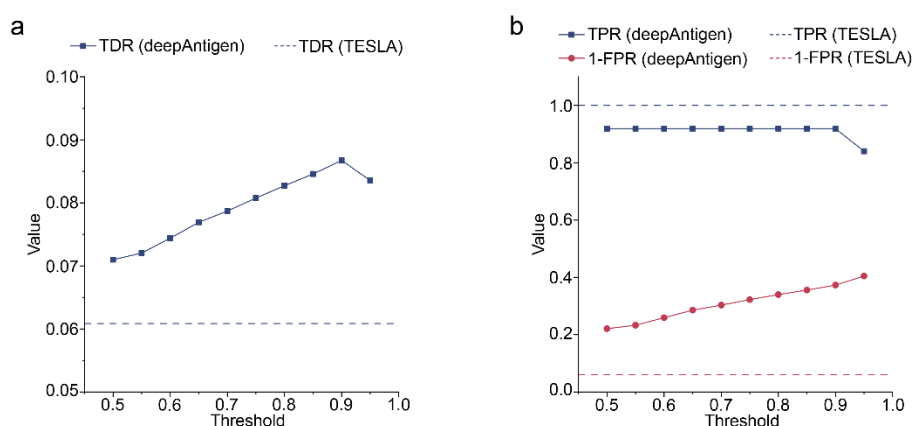
Fig.2 | Performance evaluation of deepAntigen in predicting antigen-HLA binding. d, The ROC curves, PR curves of deepAntigen and state-of-the-art methods on the antigen-HLA II benchmark test dataset (Dataset 5).

2.Clinical verification is based on a very limited scale: One of the applaudable parts of this study is the proposed method has been used in practical clinical experiments, but only two patients' tissue is used to report the performance, which cannot generate trustworthy conclusions for me. The reason is simple, it is not convincing enough to avoid any coincidental biased reported results in neoantigens effectiveness, which is not also comparable to the figure reported by the tumor neoantigen selection alliance.

Response: Thank you very much to the reviewer for recognizing our application of deepAntigen in practical clinical experiments. We understand your concern regarding the fact that tissue samples from only two patients were used to report the performance of

deepAntigen, and we appreciate you pointing out that this performance is not directly comparable to the results reported by the tumor neoantigen selection alliance (TESLA) global consortium (**PMID: 33038342**).

To address your concern regarding the comparison with TESLA, we collected 608 antigens screened by the research from TESLA, of which 37 antigens were immunogenic, as validated by biological experiments. We carefully verified that none of these 608 antigens overlapped with any antigens in the training dataset of deepAntigen. We then used deepAntigen with different threshold to predict the binding of these 608 antigen-HLA I pairs. Because, the number of negative antigens far exceeds that of positive antigens in practical clinical settings, we increased the threshold for predicting positive antigens beyond the default value of 0.5. Our results indicated that deepAntigen achieve an average true discovery rate (TDR) of 0.08, which is over the 0.06 (37/608) reported by TESLA (**Extended Fig.3a**). deepAntigen successfully recalled 34 out of 37 immunogenic neoantigens, while achieving a lower false positive rate compared to TESLA (**Extended Fig.3b**).



Extended Fig.3 | Identification of neoantigens by deepAntigen from 608 candidate antigens provided by TESLA. **a**, The comparison of true discovery rate (TDR) between deepAntigen and TESLA. **b**, The comparison of true positive rate (TPR) and false positive rate (FPR) between deepAntigen and TESLA.

To address your concern regarding the limited number of cancer patients, we included three additional patients (a lung cancer patient #P1057556, a breast cancer #P1060513, and a pancreatic cancer #P9280216). We followed the same procedures to collect tissue and peripheral blood, perform WES-seq, RNA-seq, and TCR-seq, call and filter mutations, and apply deepAntigen to identify potential neoantigens. Our results showed that out of the 100 neoantigens identified by deepAntigen from the total five cancer patients, 15 were confirmed through ELISPOT assays to significantly activate T cell immune responses (**Fig.5b,c,d and Extended Fig.4**).

Thank you once again for your thoughtful insights and valuable suggestions. We have worked diligently to revise our manuscript in response to your concerns. We believe these improvements significantly enhance the quality of our paper.

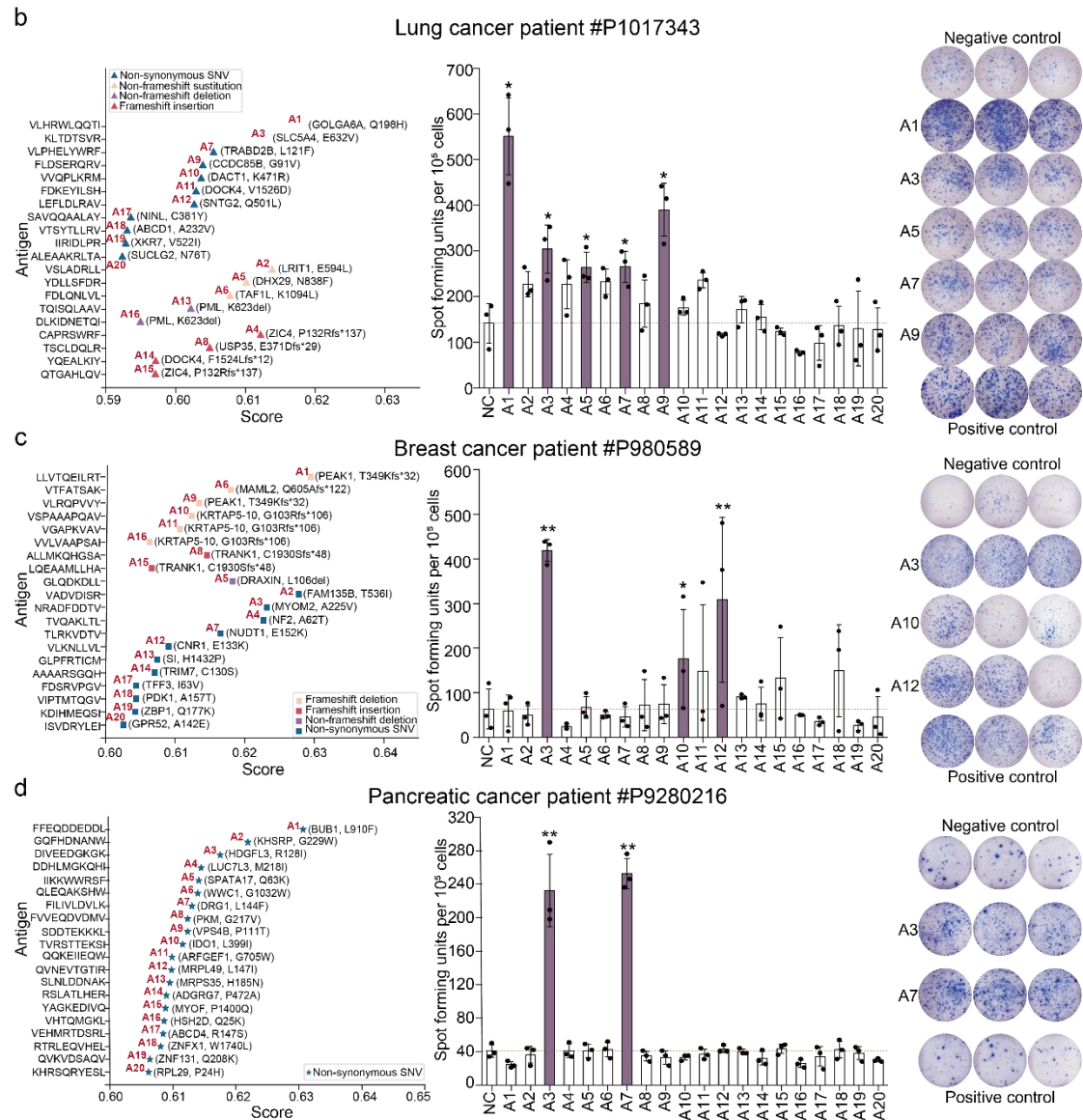
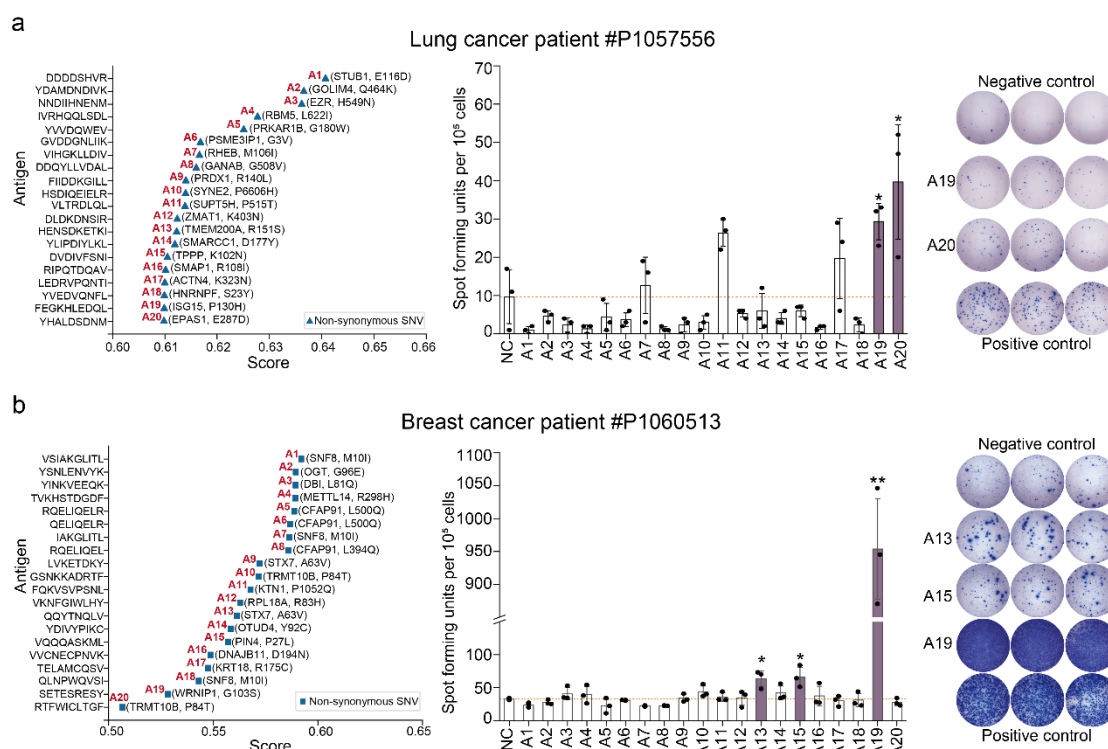


Fig.5 | Predicted tumor neoantigens were experimentally validated by ex vivo ELISPOT assays. **b**, Among the top 20 neoantigens with the highest predicted immunogenicity scores in the lung cancer patient #P1017343, five neoantigens activated strong immune responses in vitro. **c**, Among the top 20 neoantigens with the highest predicted immunogenicity scores in the breast cancer patient #P980589, three neoantigens activated strong immune responses in vitro. **d**, Among the top 20 neoantigens with the highest predicted immunogenicity scores in the pancreatic cancer patient #P9280216, two neoantigens activated strong immune responses in vitro. Bar plots show the mean values with 1.5 times standard errors. Statistical significance was determined based on distribution-free resampling (DFR) test. ‘*’ represents $P < 0.05$. ‘***’ represents $P < 0.01$. ‘NC’ is the abbreviation for ‘Negative control’.



Extended Fig.4 | Predicted tumor neoantigens were experimentally validated by ex vivo ELISPOT assays. **a**, Among the top 20 neoantigens with the highest predicted immunogenicity scores in the lung cancer patient #P1057556, two neoantigens activated strong immune responses in vitro. **b**, Among the top 20 neoantigens with the highest predicted immunogenicity scores in the breast cancer patient #P1060513, three neoantigens activated strong immune responses in vitro. Bar plots show the mean values with 1.5 times standard errors. Statistical significance was determined based on distribution-free resampling (DFR) test. ‘*’ represents $P < 0.05$. ‘***’ represents $P < 0.01$. ‘NC’ is the abbreviation for ‘Negative control’.

#Minor concerns:

1. Some of the figures are quite confusing and even misleading: For instance, Fig.2 a) compares the ROC curves across different models, but they are so close to each other, and the claimed best one - deepAntigen even performs worse in some of the regions (e.g. When 1-specificity is small in the left subfigure, and when the recall is small in the right subfigure).

Response: We thank the reviewer for raising these issues, and we sincerely apologize for the inaccurate description. Following your guidance, we adjusted the model architecture and retrained the model. The new test results show that the overall performance of deepAntigen is optimal, with comparable sensitivity at high specificity or precision at low recall. We have revised the corresponding description in the manuscript to make it more rigorous.

2. Another misleading example is Fig. 2 b), the gap between the deepAntigen performance for the antigen length 8, 10 and 11 seems to significantly differ from the others, but this effect is obvious just because the y-axis scale is too small roughly from 0.9 to 1. As a result, it may not be justice to claim that the deepAntigen really generates a stable prediction

accuracy but others do not. I would recommend, if possible, testing a wider range of antigen length to observe if this decreasing trend is more significant somehow.

Response: Thank you for your suggestion, and we sincerely apologize for any confusion caused. We have adjusted the scale of the y-axis and emphasized that deepAntigen is relatively more stable compared to other methods, even though all methods achieved fairly high predictive performance in prediction of antigen-HLA I binding. In this dataset, the length of all antigens is within the range of [8, 11], which covers the length of most HLA I-restricted antigens (PMID: 22076556).

3. Moreover, Fig.2 c), most of the scatters concentrate in the top-right region, leading to badly distinguishable demonstration. I will be happy to check a more detailed plot about the data concentrating region for better comparison across the results.

Response: Thank you for your careful review. We have provided the original data in the **Source Data** and tried to label the alleles where deepAntigen performs worse than other methods (Fig.2c,f). In fact, points near the diagonal indicate that the performance of the two methods is comparable. For the convenience of the reviewer, we also provided and statistical information in Table 1,2.

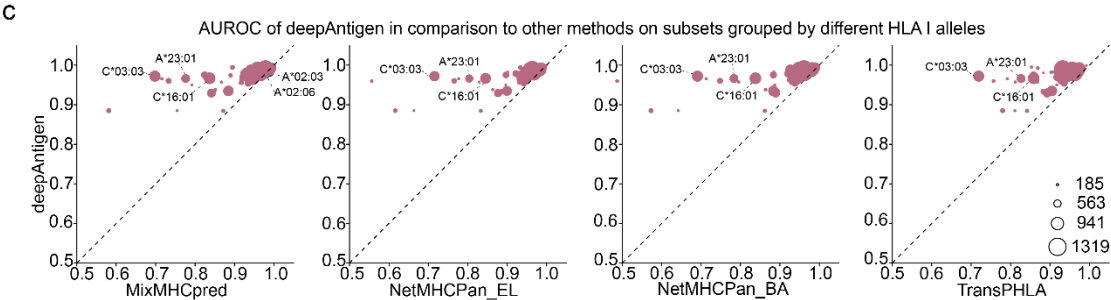


Fig.1 | Performance evaluation of deepAntigen in predicting antigen-HLA binding. c, The AUROC of deepAntigen and state-of-the-art methods on individual subsets grouped by HLA I allele type from the antigen-HLA I benchmark test dataset. Circles represent different alleles, and the size of each circle indicates the number of antigens. Magenta circles indicate deepAntigen outperforms others, while dark blue circles signify inferiority.

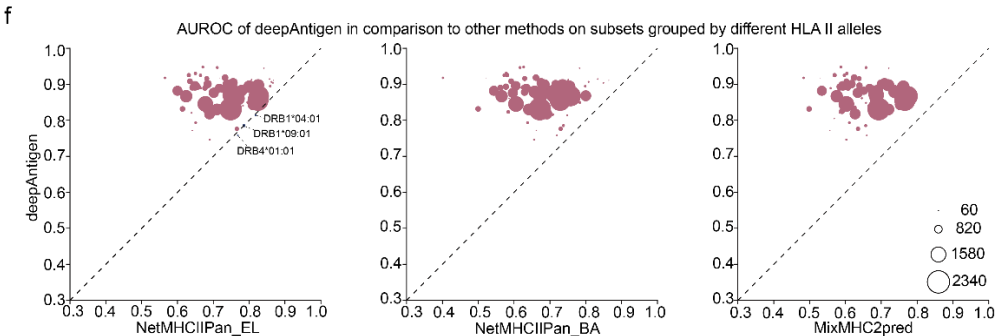


Fig.2 | Performance evaluation of deepAntigen in predicting antigen-HLA binding. f, The AUROC of deepAntigen and state-of-the-art methods on individual subsets grouped by HLA II allele type from the antigen-HLA II benchmark test dataset. Circles represent different alleles, and the size of each circle indicates the number of antigens. Magenta circles indicate deepAntigen outperforms others, while dark blue circles signify inferiority.

Table 1. Details about performance comparison on 95 HLA I alleles

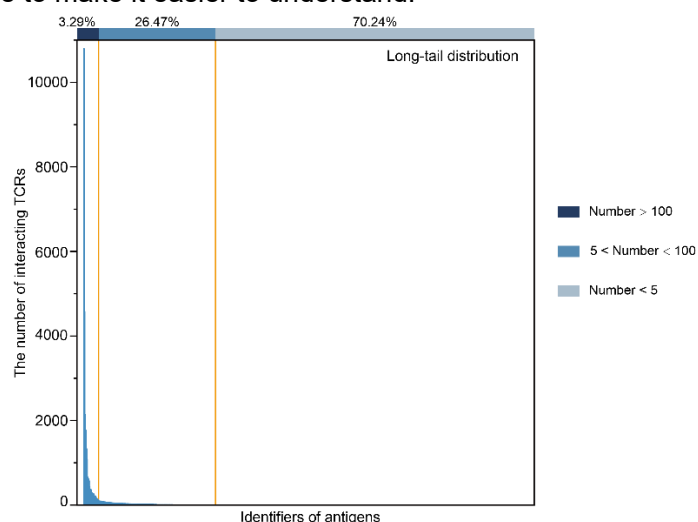
Methods	Total alleles	deepAntigen better	deepAntigen worse
deepAntigen	95	/	/
MixMHCpred	95	89	6
NetMHCpan_EL	95	93	2
NetMHCpan_BA	95	93	2
TransPHLA	95	95	0

Table 2. Details about performance comparison on 83 HLA II alleles

Methods	Total alleles	deepAntigen better	deepAntigen worse
deepAntigen	83	/	/
NetMHCIIpan_EL	83	80	3
NetMHCIIpan_BA	83	83	0
MixMHC2pred	83	83	0

4.A problematic figure for me is Supplementary Fig.4. I get confused by the annotations shown in the figure, I suppose the right black vertical line should be the higher quantile at 96.71% the left one should be a smaller quantile at 70.24%, but I am really not sure if the annotations have been mis-located.

Response: We sincerely apologize for the confusion caused by the lack of clarity. We have revised the figure to make it easier to understand.

**Supplementary Fig.11 | The number of known interacting TCRs for each antigen.**

5.Some equations are not well explained and justified: For example, I guess the equation (10)-(12) demonstrates the core of the positional information for atom features encoding, but what do the symbols i , d , h , and PE actually mean? These equations really need more explanation and justification. For example, why this encoding approach is informative to include the positional information, and why sinusoidal functions are chosen?

Response: Thank you very much for the reviewer's suggestions. We have added explanations for the specific meanings of these symbols of **equations (27) (28) (29)** in the manuscript in lines 672-674.

$$\mathbf{v}_i^L = \mathbf{v}_i^L + PE(i) \quad (27)$$

$$PE(i, 2d) = \sin\left(\frac{i}{10000^{(2d/h)}}\right), d = 0, 1, 2, \dots, \frac{h}{2} - 1 \quad (28)$$

$$PE(i, 2d + 1) = \cos\left(\frac{i}{10000^{(2d/h)}}\right), d = 0, 1, 2, \dots, \frac{h}{2} - 1 \quad (29)$$

“ **PE** is positional encoding matrix. i is the position index of atoms in the graph. h is the dimension of positional encoding, which is equal to the number of hidden units in the GCN. d is the index used to traverse the dimension of each positional encoding vector.”

This positional encoding scheme is inspired by the Transformer encoding approach, with the distinction that each atom is regarded as a token. We have added the reason why sinusoidal functions are chosen in lines 676-680.

“The sinusoidal positional encoding provides a smooth, continuous, and unique encoding for each atom in the graph. In cases where the number of atoms differs across graphs, the relative positions of atoms in different graphs can still remain consistent. Additionally, this encoding can also represent atom counts that the model has never encountered during training.”

6. Similar to the equation (14), I cannot find an explanation for the two quantities e^s_i and e^s_j .

Response: We apologize for this lack of clarity and appreciate the reviewer’s suggestion. We have revised this equation and added explanations for these two quantities.

$$S_{ij} = e^{\text{score}_i} e^{\text{score}_j} \quad (31)$$

“ score_i is the score of the i -th atom of the antigen. score_j is the score of the j -th atom of the TCR.”

Responses to Reviewer #4

#Overview

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Response: Thank you for your valuable insights and helpful suggestions for our research. They have been instrumental in improving the quality and clarity of our paper. We sincerely appreciate the time and effort you have dedicated to this.

Dear Referees,

Thank you very much for the constructive comments and the opportunity to resubmit the revised manuscript to **Nature Communications**. Enclosed are the point-by-point responses to the comments from four reviewers. In the point-by-point responses, the original comments from reviewers are marked *blue*, and our responses in *black*.

Responses to Reviewer #2

Reviewer #2 (Remarks to the Author):

Overview

I have reviewed the revised manuscript and I am satisfied with all the corrections made. The authors have adequately addressed all my previous comments and concerns or indicated explicitly in the text the connections I have pointed in the first review. I believe that the paper can be accepted in its current form.

Response: Thank you for your thorough review and positive feedback on the revised manuscript. We greatly appreciate your time and effort in evaluating our work, as well as your valuable comments during the review process. We are pleased to hear that the corrections and clarifications adequately addressed your concerns and that the manuscript is now suitable for acceptance. Thank you again for your support and contribution to improving our paper.

Responses to Reviewer #4

Reviewer #4 (Remarks to the Author):

#Overview

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Response: We greatly appreciate your insightful feedback and thoughtful comments for our study. Your input has significantly enhanced the quality and coherence of our manuscript. Thank you for the time and dedication you have invested in this process.

Reviewer #4 (Remarks on code availability):

I have been able to login on CodeOcean and get access to documentation.

I created a non-identifiable email for that.

I also tested the package installation on a Google Colab environment.

GPUs had to be used on the runtime.

I don't know if same can be done under a free account.

Response: Thank you for reviewing our code, which helps improve its reproducibility. Our program automatically detects whether a GPU is available. If available, it will use the GPU to accelerate computations; if not, it can still be performed normally using the CPU. The

code previously on CodeOcean was not formally released during the peer review stage. We have now released the code, and it is accessible and runnable at <https://codeocean.com/capsule/6907488/tree/v1> for free accounts as well.