Cell Systems

Transparent Peer Review Record

Improved predictions of antigen presentation and TCR recognition with MixMHCpred2.2 and PRIME2.0 reveal potent SARS-CoV-2 CD8+ T-cell epitopes

David Gfeller, Julien Schmidt, Giancarlo Croce, Philippe Guillaume, Sara Bobisse, Raphael Genolet, Lise Queiroz, Julien Cesbron, Julien Racle, Alexandre Harari

Summary

Initial Submission: Received: June 23, 2022
Preprint: doi.org/10.1101/2022.05.23.492800
Scientific editor: Bernadett Gaal, DPhil

First round of review: Number of reviewers: 2
2 confidential, 0 signed
Revision invited July 20, 2022
Major changes anticipated
Revision received Oct 12, 2022

Second round of review: Number of reviewers: 1
1 original, 0 new
1 confidential, 0 signed
Accepted Dec 8, 2022

Cell Systems Transparent Peer Review Record

Editorial decision letter with reviewers' comments, first round of review

Dear David,

I hope this email finds you well. The reviews of your manuscript are back and I've appended them below. On balance, the reviewers appreciate the goals of the work presented here; they've provided constructive comments that are aligned with our hopes for the paper. Accordingly, we're happy to invite a revision.

You will see that among other concerns, both Reviewers point out that it is not clearly demonstrated that the methodological improvements were needed to enable identification of the SARS-CoV-2 epitopes, and Reviewer 1 raises concerns about the strength of the experimental validation of the SARS-CoV-2 epitopes. These concerns would be particularly important to address to help demonstrate the utility of the methodological headway described in the manuscript and in light of the emphasis placed on the findings. That said, in line with Reviewer 1's advice, which echoes my own from our first discussion, I would advisee you to ensure that the methodological advances are better emphasised.

In addition to these points, I've highlighted points that seem to warrant special attention. I'd also like to be explicit about an almost philosophical stance that we take at Cell Systems.

I hope you find this feedback helpful. If you have any questions or concerns, I'm always happy to talk, either over email or Zoom. More technical information and advice about resubmission can be found below my signature. Please read it carefully, as it can save substantial time and effort later.

I look forward to seeing your revised manuscript.

All the best,
Bernadett

Bernadett Gaal, DPhil
Editor-in-Chief, Cell Systems

Reviewers' comments:

Reviewer #1: In this work, Gfeller, Schmidt et al report the results of a series of in silico and in vitro experiments. First, the authors curated different datasets of HLA-I ligands and neo-epitopes. Second, they used this data to train two algorithms: MixMHCpred2.2, which predicts antigen presentation, and PRIME2.0, which models TCR recognition propensity. The authors provide comparisons with state-of-the-art methods and use MixMHCpred to find motifs that are experimental contaminants. Finally, the authors used these tools to map new epitopes in the SARS-COV-2 genome and present results of their experimental validation.

The manuscript presents algorithmic improvements, shows encouraging results by outperforming state-of-the-art methods, and provides some experimental evidence for real case use. The authors also went to great lengths to add some insights into the underlying motifs in epitopes and I commend their transparency in sharing their algorithms and data.

1. I do not think the title reflects the full scope of the paper. The title ignores they gathered a dataset of experimentally verified neo-epitopes and presented HLA-I ligands, the algorithm to predict antigen presentation MixMHCpred2.2) and the one to predict T-cell activation (PRIME2.0), which, combined, improved the prediction of HLA-I ligands and neo-epitopes.

2. The section about experimental validation using SARS-COV-2 epitope identification is rather weak, with only 6 individuals included and only 213 peptides tested. The results look anecdotal, and there is little in terms of statistical analysis of their validity in terms of supporting the claims in the previous section about the algorithms being better than what would have been achieved with other versions. Overall, these data distract from the more important methodological improvement in the current versions of MixMHCpred2.2 and PRIME2.0 and highlight these tools in the paper at greater length that are better shown on the much larger datasets used in those sections.

3. The part of the contaminants could be explained in more detail: what constitutes a contaminant, ie, how do you know a given motif is a contaminant and not part of the pattern of the amino acids that bind to the HLA protein but that MixMHCpred misses? The authors say "Peptides assigned to the flat motif (trash) in MixMHCp or to motifs corresponding to alleles not supposed to be in the sample were considered as predicted contaminants." (Isn't that a Catch-22?). I fail to understand how a motif can be classified as a contaminant without further experimental validation. The paper should clarify this apparent contradiction.

4. I find the section on examining the difference in peptide length modeling of the different methods rather confusing. A straightforward comparison would be to run just the 9-mers in the dataset on all the different prediction methods; presumably that would show that HLA-Athena performs more comparable on this basis, which would directly make the point that the 'motif' extracted is fine, but comparing motif scores across different lengths is done poorly for HLA-Athena. This importance of providing an *overall rank* of predicted peptides, and not just comparing peptides of the same length is still often lost on people. In that context, please also cite the Trolle paper when discussing this here, which was the first to explicitly model peptide length available for binding and the different allele binding specificities to predict allele specific ligand elution length profiles, giving a mechanistic explanation rather than the black-box-neural network approaches.

5. 'The comparison of the results of PRIME2.0 with those obtained when using a logistic regression trained on the same data instead of a neural network (Figure 3C) highlights the improvements obtained with neural networks.' I am not sure this entirely explains the data observed. First, the other methods outperformed by PRIME2 also use neural networks. Second, the use of logistic regression also led to the second-best result in Figure 3D and does not significantly differ from PRIME2.0. Doesn't that mean the training set used is better, rather than the algorithm itself?

6. Methods section, comparison with other tools (p.14-15)
How many peptides does the dataset used for comparison contain? It references another paper, but it would be useful to include it as a Supplementary Material as the authors have it.

7. Page 6, How did they retrieve motifs from binding predictions? Some readers won't know this

Reviewer #2: This paper presents an updated version of the authors' MHC binding/presentation predictor (MixMHCPred 2.2, trained on MHC ligandome data) as well as a significant improvement of their T-cell recognition predictor (PRIME2.0, trained on a curated dataset of neoantigen immunogenicity). These two predictors combine nicely to more robustly embody the two important factors of immunogenicity laid out in the TESLA paper (presentation and recognition). The paper is well done and presents tools which could potentially be useful to the community. They demonstrate some known properties of immunogenic peptides (preference for larger amino acids with hydrophobic side chains) and an interesting trade-off between MHC affinity and important of immunogenic sequence features. They then apply the PRIME model to select candidate T-cell epitopes from the SARS-CoV-2 proteome, creating a candidate set of 213 peptides which were tested using ELIspot against T-cells from three convalescent donors and three immunologically naive donors. This led to the discovery of 19 SARS-CoV-2 T-cell epitopes, of which several were recurrent, such as the A24 restricted epitope YFPLQSYGF. TCR sequencing revealed a memory phenotype enriched TCR which recognizes the SARS-CoV-2 epitope A24:02/QYIKWPWYIW and whose TCR sequence also occurred repeatedly in an external dataset.

My main concern about this paper is that its benefit to the community is primarily through having a validating T-cell epitope predictor to help narrow down the set of candidate peptides in studies like the one you performed. However, it's not clear if 19/213 hits from 3 patients is better than chance or how replacing PRIME with NetMHCpan, MHCflurry, or any other method might have effected the yield of true epitopes. I hope that the authors can better connect the methods development and SARS-CoV-2 explorations by showing that there's some advantage to using PRIME 2.0 as the basis of constructing the ELISpot peptide set.

Another concern I have or at least a source of confusion is the relevance of QYIKWPWYIW as a SARS-CoV-1 epitope. I'm guessing that Leu184 almost certainly has *not* been exposed to 2003 SARS. If the explanation for Leu184 having a SARS-CoV-2 recognizing TCR population is cross-reactivity with other CoVs then it only makes sense to look at 229E, OC43, NL63, and HKU1.

Less significantly, one thing I'm curious about is the compose-ability of PRIME with other MHC presentation predictors. How would PRIME perform if given % rank values from NetMHCpan or MHCflurry?

I'm not entirely certain where on the border between major and minor revision these concerns fall. I think the paper as presented would be ready for publication if it at least conveyed some information about (a) how well detected SARS-CoV-2 epitopes would have been predicted by other computational screens (b) how well QYIKWPWYIW is conserved in common CoVs.

Authors' response to the reviewers' first round comments

Attached.

Editorial decision letter with reviewers' comments, second round of review

Dear David,

I'm very pleased to let you know that your manuscript is now "accepted in principle," that is, pending our receipt of your final files in their proper format. Congratulations! We have unfortunately not heard back from Reviewer 2, but having considered your responses and the overlap between the two Reviewers' concerns, we are happy to proceed.

This email contains a lot of detailed information. All of it is important, so please read it very carefully. The bulleted list below highlights information relevant to our formatting checks, and, after acceptance, your manuscript's transfer to our production department for typesetting and publication. Please review this information, ask any questions you may have, make any changes necessary to your manuscript files, and then upload your final files into Editorial Manager. Once we receive your final files, we can go through our formatting checks.

COVID-19 PAPERS

Cell Press is committed to publishing all COVID-19 papers as soon as possible, and to making that content freely available. To meet this commitment, we will post a pre-proof of your paper to our Coronavirus Hub a few days after acceptance. This pre-proof will include the Supplemental Figures, but it will not include any additional Supplemental Items such as spreadsheets, movies, and datasets. If your paper includes these additional types of supplemental items, we ask that you do a few things to ensure that readers have access to those items.

1. Upload the items to Mendeley Data. When you create the Mendeley dataset, we suggest that you make the title of the dataset the title of your paper followed by the first author's last name (eg. "A study of XYZ. Smith et al"). We also suggest that you make the dataset publicly available at the time that you upload it.
2. Include the DOI (Digital Object Identifier) of your Mendeley dataset in the "Data and Software Availability" section of the STAR methods. Please use this template for the text that you include in the STAR methods: "Additional Supplemental Items are available from Mendeley Data at http://dx.doi.org/[your DOI number from Mendeley Data]".

Please note that Mendeley Data will prompt you to choose a Creative Commons License for your publicly shared datasets. We suggest that you choose the license that is best for you. CC BY, CC BY-ND, and CCO licenses are appropriate for most datasets. You can learn more about Creative Commons Licenses at http://creativecommons.org/licenses/. The final published version of your paper will include all of the supplemental files that you provide with the accepted paper. This Mendeley Data step is an interim solution to accompany release of the pre-proof version of your paper for rapid communication.

Additional editorial requests:

- *If you are using GitHub, please follow the instructions here to archive a "version of record" of your GitHub repo at Zenodo, then report the resulting DOI. Please also ensure that you report in the Key Resources Table the specific versions of PRIME and MixMHCpred that you use in this manuscript.*
- Bar graphs are not acceptable because they obscure important information about the distributions of the underlying data. Please display individual points within your graphs unless their large number obscures the graph's interpretation. In that case, box-and-whisker plots are a good alternative.

We would like to publish your paper as a Report. This will not require any additional formatting as your manuscript is already in the Report format.

It's been a pleasure working with you, please feel free to contact our journal team with questions.

All the best,

Bernadett

Bernadett Gaal, DPhil
Editor-in-Chief, Cell Systems

Reviewer comments:

Reviewer 1

The authors have appropriately addressed my previous concerns.

Reviewer #1: In this work, Gfeller, Schmidt et al report the results of a series of in silico and in vitro experiments. First, the authors curated different datasets of HLA-I ligands and neo-epitopes. Second, they used this data to train two algorithms: MixMHCpred2.2, which predicts antigen presentation, and PRIME2.0, which models TCR recognition propensity. The authors provide comparisons with state-of-the-art methods and use MixMHCpred to find motifs that are experimental contaminants. Finally, the authors used these tools to map new epitopes in the SARS-COV-2 genome and present results of their experimental validation.

The manuscript presents algorithmic improvements, shows encouraging results by outperforming state-of-the-art methods, and provides some experimental evidence for real case use. The authors also went to great lengths to add some insights into the underlying motifs in epitopes and I commend their transparency in sharing their algorithms and data.

- We thank the reviewer for accurately summarizing our manuscript and for recognizing the different strengths of our work.

1. I do not think the title reflects the full scope of the paper. The title ignores they gathered a dataset of experimentally verified neo-epitopes and presented HLA-I ligands, the algorithm to predict antigen presentation MixMHCpred2.2) and the one to predict T-cell activation (PRIME2.0), which, combined, improved the prediction of HLA-I ligands and neo-epitopes.

- We thank you for this feedback. After discussing with the editor, we decided to change the title to better reflect the scope of the manuscript and specifically mention the two predictors (MixMHCpred2.2 and PRIME2.0). The new title that we propose reads: “Improved predictions of antigen presentation and TCR recognition with MixMHCpred2.2 and PRIME2.0 reveal potent SARS-CoV-2 CD8+ T-cell epitopes”

2. The section about experimental validation using SARS-COV-2 epitope identification is rather weak, with only 6 individuals included and only 213 peptides tested. The results look anecdotal, and there is little in terms of statistical analysis of their validity in terms of supporting the claims in the previous section about the algorithms being better than what would have been achieved with other versions. Overall, these data distract from the more important methodological improvement in the current versions of MixMHCpred2.2 and PRIME2.0 and highlight these tools in the paper at greater length that are better shown on the much larger datasets used in those sections.

- We thank you for this important feedback and we tried to put more emphasis on the methodological aspects of our work, including several new analyses based on the different comments of the reviewers. We agree that the prospective application to SARS-CoV-2 does not provide evidence of the improvements over other methods. This part was covered in Figure 3 and Figure S3, and we have clarified this point in the manuscript. However, we still see some value in this work, in the sense that it shows that PRIME2.0 could be applied prospectively to new data, and not only to the re-predictions of known epitopes in a cross-validation setting. The fact that we could validate some epitopes demonstrates the possibility to use PRIME2.0 for practical applications. We also envision that the identification of SARS-CoV-2 epitopes (especially the one in effector/memory cells) may be of interest to several people. Screening more donors and more epitopes is unfortunately beyond the scope of this manuscript.

3. The part of the contaminants could be explained in more detail: what constitutes a contaminant, ie, how do you know a given motif is a contaminant and not part of the pattern of the amino acids that bind to the HLA protein but that MixMHCpred misses? The authors say "Peptides assigned to the flat motif (trash) in MixMHCp or to motifs corresponding to alleles not supposed to be in the sample were considered as predicted contaminants." (Isn't that a Catch-22?). I fail to understand how a motif can be classified as a contaminant without further experimental validation. The paper should clarify this apparent contradiction.

- We apologize for not providing more details on this part of the manuscript and we have expanded this analysis. As with all high-throughput technologies, we expect some contaminants in HLA-I peptidomics data. Contaminants can consist of wrongly identified spectra (due to the 1%FDR used in MS analysis), contaminants due to bad washing of the instrument, to some residual unspecific binding of the antibody used in the pull-down, or other unidentified causes. For many of these contaminants (especially those that do not bind to HLA-I alleles), we do not expect to see any clear motif. This can be captured by the motif deconvolution approach of MixMHCp, where an unspecific (flat) motif is included by construction to handle peptides that do not match any of the learned motif (similar concept as the trash cluster in GibbsCluster). In case of contaminations from previous standard proteomics (e.g., bad washing of the instruments), contaminants are expected to show a trypsin-like motif at the C-terminus (R/K). In case of contaminations from previous HLA-I peptidomics, contaminants are expected to show motifs of other HLA-I alleles. In addition, for engineered or transfected mono-allelic cells, there is always the risk that the cells are actually not fully mono-allelic and that there is some residual expression of the endogenous HLA-I.

All three types of contaminations (i.e., unspecific, trypsin-like motifs and motifs from other HLA-I alleles) were observed in our data. For instance, in the mono-allelic HLA-B*37:01 samples, we observed a motif resembling strongly the expected trypsin-like motif (Figure 1A and Figure S1B). For the mono-allelic HLA-B*27:05, HLA-C*07:01, HLA-C*07:02 and HLA-C*07:04 samples of Sarkizova et al. NBT 2020, our motif deconvolution identified a second motif which is identical to the one of HLA-C*01:02 and which does not resemble any HLA-B*27:05 or HLA-C*07 motifs (Figure S1C). Moreover, this secondary motif was never observed in any other HLA-C*01:02^{neg} sample that contains any of these four allele (data not shown). This is a strong indication that there may be some contaminations, likely from the mono-allelic HLA-C*01:02 sample which was profiled in the same study (Sarkizova et al. NBT 2020).

Considering that mono-allelic data have often been used ‘as is’ (i.e., without QC) to train predictors, we believe this is something important to bring to the attention of the community (see also Sricharoensuk et al., Front Immunol 2022 and Fritsche et al., MCP 2021 for similar observations). Of course, despite quite strong evidences, we can never be 100% sure that all peptides assigned to the flat motif or to motifs corresponding to other alleles are real contaminants. This is the reason why we always spoke about ‘predicted contaminants’.

4. I find the section on examining the difference in peptide length modeling of the different methods rather confusing. A straightforward comparison would be to run just the 9-mers in the dataset on all the different prediction methods; presumably that would show that HLA-Athena performs more comparable on this basis, which would directly make the point that the 'motif' extracted is fine, but comparing motif scores across different lengths is done poorly for HLA-Athena. This importance of providing an *overall rank* of predicted peptides, and not just comparing peptides of the same length is still often lost on people. In that context, please also cite the Trolle paper when discussing this here, which was the first to explicitly model peptide length available for binding and the different allele binding specificities to predict allele specific ligand elution length profiles, giving a mechanistic explanation rather than the black-box-neural network approaches.

- We thank the reviewer for this insightful comment and we performed the proposed analyses (Figure 2C-D). As anticipated by the reviewer, we observed very few statistically significant differences in AUC values when considering separately each peptide length. We agree that this is a much stronger argument to show that most predictors correctly capture HLA-I motifs, and that important differences are mainly due to differences in

modelling peptide length distributions. For this reason, we have removed the previous Figure 2C and replaced it with the analysis of AUC values at different lengths (new Figure 2C-D). We have also mentioned the important results of Trolle et al. (previously only cited in the introduction) and how this provides a mechanistic explanation to the discrepancies between peptide length distribution observed in naturally presented HLA-I ligands and peptide length distributions inferred from binding assays. We are thankful to the reviewer for proposing this alternative analysis.

5. 'The comparison of the results of PRIME2.0 with those obtained when using a logistic regression trained on the same data instead of a neural network (Figure 3C) highlights the improvements obtained with neural networks.' I am not sure this entirely explains the data observed. First, the other methods outperformed by PRIME2 also use neural networks. Second, the use of logistic regression also led to the second-best result in Figure 3D and does not significantly differ from PRIME2.0. Doesn't that mean the training set used is better, rather than the algorithm itself?

- We apologize for the confusion. This sentence was meant to apply only on the comparison between PRIME2.0 trained based on neural networks and PRIME2.0 trained with a logistic regression. For the other methods, we do not think that we can draw robust conclusion regarding the impact of neural networks since these other methods are trained on HLA-I ligands and therefore predict antigen presentation, not TCR recognition. For the results in Figure 3D, we actually believe that the 'similar' results of PRIME2.0 trained with neural network or logisitic regression are consistent with the interpretation provided in Figure 3E: namely, that there are correlations between immunogenicity propensity and binding affinity to HLA-I. So when considering the full spectrum of possible binding affinities, using models that capture these correlations is useful. Reversely, when restricting to data (both positives and negatives) that have high predicted binding affinities, these correlations are less important, and therefore PRIME2.0 trained with a neural network or a logistic regression perform similarly.

6. Methods section, comparison with other tools (p.14-15)
How many peptides does the dataset used for comparison contain? It references another paper, but it would be useful to include it as a Supplementary Material as the authors have it.

- We have included the different test sets in Dataset S3. In total, this represents 78,011 HLA-I ligands (sum over all ligands in each sample), and 312,044 random negatives.

7. Page 6, How did they retrieve motifs from binding predictions? Some readers won't know this

- This part was removed based on one of the previous comments.

Reviewer #2: This paper presents an updated version of the authors' MHC binding/presentation predictor (MixMHCPred 2.2, trained on MHC ligandome data) as well as a significant improvement of their T-cell recognition predictor (PRIME2.0, trained on a curated dataset of neoantigen immunogenicity). These two predictors combine nicely to more robustly embody the two important factors of immunogenicity laid out in the TESLA paper (presentation and recognition). The paper is well done and presents tools which could potentially be useful to the community. They demonstrate some known properties of immunogenic peptides (preference for larger amino acids with hydrophobic side chains) and an interesting trade-off between MHC affinity and important of immunogenic sequence features. They then apply the PRIME model to select candidate T-cell epitopes from the SARS-CoV-2 proteome, creating a candidate set of 213 peptides which were tested using ELIspot against T-cells from three convalescent donors and three immunologically naive donors. This led to the discovery of 19 SARS-CoV-2 T-cell epitopes, of which several were recurrent, such as the A24 restricted epitope YFPLQSYGF. TCR sequencing revealed a memory phenotype enriched TCR which recognizes the SARS-CoV-2 epitope A24:02/QYIKWPWYIW and whose TCR sequence also occurred repeatedly in an external dataset.

- We thank the reviewer for this accurate summary of our work.

My main concern about this paper is that its benefit to the community is primarily through having a validating T-cell epitope predictor to help narrow down the set of candidate peptides in studies like the one you performed. However, it's not clear if 19/213 hits from 3 patients is better than chance or how replacing PRIME with NetMHCpan, MHCflurry, or any other method might have effected the yield of true epitopes. I hope that the authors can better connect the methods development and

SARS-CoV-2 explorations by showing that there's some advantage to using PRIME 2.0 as the basis of constructing the ELISpot peptide set.

- We apologize for the lack of clarity. The data supporting the improvement of PRIME2.0 compared to existing tools are shown in Figure 3. This is based on a large set of more than 500 validated immunogenic peptides, and hundred times more negatives. The data shown in Figure 4 were meant to provide evidences that PRIME2.0 can be applied in a prospective way to identify epitopes that are biologically relevant. We agree that the results in Figure 4 do not provide direct evidence of improvements over other methods, since we only made predictions with PRIME2.0, so it's unclear what would have been the results of applying NetMHCpan, MHCflurry or any other tool instead of PRIME2.0. This part should therefore be considered as a prospective analysis in which we were completely blind to the output. We have clarified these points in the manuscript. When computing the scores of the immunogenic peptides identified in this work with the alleles of their respective samples, we observed that PRIME2.0 had the best scores (Figure S4A). However, this analysis is obviously biased, because the peptides were selected with PRIME2.0. When computing AUC values within the set of 213 peptides that were experimentally tested, we observed that all tools had AUC larger than 0.5 (Figure S4B), with not tool showing statistically significantly better AUC. However, here again, it is hard to draw firm conclusions since we had not tested peptides predicted by other tools, and the number of immunogenic peptides (19) is small. Unfortunately, we do not have T-cells remaining for the donors, so we could not run new experiments with peptides pre-selected with other tools.

Another concern I have or at least a source of confusion is the relevance of QYIKWPWYIW as a SARS-CoV-1 epitope. I'm guessing that Leu184 almost certainly has *not* been exposed to 2003 SARS. If the explanation for Leu184 having a SARS-CoV-2 recognizing TCR population is cross-reactivity with other CoVs then it only makes sense to look at 229E, OC43, NL63, and HKU1.

- We are very thankful to the reviewer for pointing out this issue and agree that cross-reactivity only with SARS-CoV-1 is unlikely to explain the immunogenicity of the QYIKWPWYIW epitope. We therefore analyzed the other coronaviruses mentioned by the reviewer. Strikingly, we observed that the SARS-CoV-1 epitope is very well conserved across all 4 species (see below and Table 2):

  229E: TYIKWPWWVW
  OC43: YYVKWPWYVW

NL63: NYIKWPWWVW
HKU1: MYVKWPWYVW

To demonstrate the cross-reactivity hypothesis, we tested these four peptides with cells transfected with the TCR recognizing both QYIKWPWYIW and QYIKWPWYVW. We observed cross-reactivity with all four peptides (Figure 4E).

Less significantly, one thing I'm curious about is the compose-ability of PRIME with other MHC presentation predictors. How would PRIME perform if given % rank values from NetMHCpan or MHCflurry?

- We included this analysis in the revised version. Overall, we observed similar improvements when using NetMHCpan, MHCflurry or HLAthena to train PRIME2.0 compared to these HLA-I ligand predictors used on their own (Figure S3C-D). This demonstrates the robustness of the PRIME2.0 framework to other predictions of HLA-I binding. Moreover, the analysis of the importance for immunogenicity of different amino acids at different values of binding affinity led to similar results when stratifying the data with NetMHCpan, MHCfluury or HLAthena (Figure S3E). These results demonstrate the robustness of our observations and interpretation.