nature portfolio

Corresponding author(s): Rachel Karchin

Last updated by author(s): Jun 14, 2023

Reporting Summary

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

Statistics

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

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

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

Software and code

Policy information about availability of computer code

Data collection	Predictions were generated using the following methods: MHCnuggets 2.4.0, NetMHCpan-4.1, MHCflurry-2.0, MixMHCpred-2.1, PRIME-1.0, PRIME-2.0, TransPHLA (not versioned), and HLAthena (not versioned). The backend libraries included TensorFlow 2.9.1, and PyTorch 1.13.0 with Python 3.9.12. Multiple sequence alignment was performed with the buildmodel and align2model tools from SAM suite v3.5.
Data analysis	BigMHC 1.0 (https://github.com/karchinlab/bigmhc), scikit-learn 1.0.2, pandas 1.4.2, numpy 1.21.5, matplotlib 3.5.1, seaborn 0.12.2, py3Dmol 2.0.1, AlphaFold2, ImmunoSELECT-R (not versioned), and FEST (not versioned)

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

Data

Policy information about availability of data

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

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All data, including model outputs and MANAFEST data, are provided in our public Mendeley repository: https://data.mendeley.com/datasets/dvmz6pkzvb.

All data except MANAFEST data were collected from publicly available sources:
- MHCflurry-2.0 (https://doi.org/10.1016/j.cels.2020.06.010)
- NetMHCpan-4.1 (https://doi.org/10.1093/nar/gkaa379)
- PRIME-1.0 (https://doi.org/10.1016/j.xcrm.2021.100194)
- PRIME-2.0 (https://doi.org/10.1016/j.cels.2022.12.002)
- TESLA (https://doi.org/10.1016/j.cell.2020.09.015)
- IEDB (https://doi.org/10.1093/nar/gky1006)
- NEPdb (https://doi.org/10.3389/fimmu.2021.644637)
- Neopepsee (https://doi.org/10.1093/annonc/mdy022)
- IPD-IMGT/HLA (https://doi.org/10.1093/nar/gkz950)
- IPD-MHC 2.0 (https://doi.org/10.1093/nar/gkw1050)
- UniProt (https://doi.org/10.1093/nar/gkac1052) accession codes: P01899, P01900, P14427, P14426, Q31145, P01901, P01902, P04223, P14428, P01897, Q31151.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	Not applicable
Population characteristics	Not applicable
Recruitment	Not applicable
Ethics oversight	Not applicable

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

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

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

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

Life sciences study design

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

Sample size	Sample sizes were determined by data availability from open sources as well as previously acquired MANAFEST data.
Data exclusions	Epitopes that could not be evaluated by other methods were removed. These include epitopes with dummy amino acids (e.g. 'X'), epitopes longer than 11 amino acids for comparisons with HLAthena, and instances with alleles that prior methods cannot evaluate (MixMHCpred, PRIME, and HLAthena). Instances that appear in training data are excluded from testing sets.
Replication	All code used in this study is open-sourced for reproducing results. Detailed instructions and scripts are provided for data preparation, model evaluation, and figure generation: https://github.com/KarchinLab/bigmhc
Randomization	Not applicable
Blinding	Not applicable

Reporting for specific materials, systems and methods

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

Materials & experimental systems

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

Methods

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