https://doi.org/10.1038/s43018-023-00675-z

Breaking the performance ceiling for neoantigen immunogenicity prediction

Hugh O'Brien, Max Salm, Laura Morton, Maciej Szukszto, Felix O'Farrell, Charlotte Boulton, Pablo D. Becker, Yardena Samuels, Charles Swanton, Marc R. Mansour, Sine Reker Hadrup & Sergio A. Quezada

Neoantigen immunogenicity prediction is a burgeoning field with vast potential; however, the shortage of high-quality data and biases in current datasets limit model generalizability. Here we discuss some of the pitfalls that may underly this limited performance and propose a path forward.

Neoantigen-based personalized tumor therapies are emerging as a promising treatment modality. Although efficient and accurate selection of immunogenic neoantigens is a critical determinant of therapy success, the number of candidates to select from frequently exceeds the therapy's payload. For example, in melanoma there are typically around ten times more candidate mutations than can be included in a neoantigen vaccine. This selection issue is further compounded by a low base rate of neoantigen immunogenicity (~2–6%)[1], which we define here as the ability of a neoantigen to elicit a T cell response, and by the fact that most neoantigens are unique to a tumor.

Although in vitro neoantigen immunogenicity screening approaches are improving in accuracy and throughput, they remain resource intensive. Computational immunogenicity-prediction models therefore offer higher efficiency and are frequently relied upon for target selection. Despite recent improvements in models assessing immunogenicity, however, their performance is limited by a shortage of diverse, high-quality data. In particular, the generalizability of models to a highly variable domain, such as neoantigens presented on a diverse set of human leukocyte (HLA) molecules, remains challenging.

In this Comment, we focus on approaches for predicting antigen-specific $CD8^+$ T cell responses; the prediction of $CD4^+$ T cell responses is less well developed, and the issues discussed here are even more pronounced in that context.

Publicly available datasets are limited and skewed

A general rule in machine learning is that more data are often better. Although high data quality is also essential, highly variable and irregular distributions require substantial data volumes for accuracy and generalizability. Neoantigens are both naturally rare and hard to identify; for example, gastrointestinal tumors typically harbor one to three immunogenic neoantigens[1]. As such, neoantigen training data from primary tumors are slow to accrue and very scarce, with 2,067 human neoantigens currently deposited in the CEDAR database[2] (accessed in September 2023). In practice, models are trained on datasets from more accessible proxy sources, such as viral and wild-type self-peptides. This is a considerably smaller pool of source epitopes compared to the space of possible neoantigens, which can arise from even a single-amino-acid change to a self-peptide. The degree of change required for a self-peptide to escape central tolerance remains to be established. Existing data sources therefore do not capture the type of variance seen in mutated self-neoantigens, a gap that is reflected in the limitations of existing models.

Additional biases in the existing datasets of immunogenic neoantigens affect their usefulness in training prediction algorithms. These biases include survivorship bias due to immunoediting, ascertainment bias of major histocompatibility class (MHC) alleles due to an over-representation of samples with European ancestry, and tumor type bias due to tumors with a high mutational burden. Moreover, to accommodate the limited throughput of screening methods, it is common to pre-select targets using a peptide–MHC binding predictor, which propagates bias from the predictor training set into new training data. For example, models trained using immunopeptidomics data systematically under-represent cysteine-containing peptides[3]. Finally, immunogenicity screening assays yield a variety of outputs related to T cell activity or neoepitope–HLA recognition, leading to variable definitions of immunogenicity that could introduce observer bias.

Classification-based supervised learning requires negative training data. However, technical factors such as assay sensitivity and biological factors such as T cell state may contribute to 'false negative' data, whereby a neoantigen may be immunogenic in another context (Fig. 1a). For instance, differences have been observed in neoantigen reactivity in patient samples before and after immunotherapy[4]. This can be problematic, especially for models that consider only the peptide–MHC pair as inputs. Neoantigens may elicit a response in some patients but not others, which may be predictable from their T cell repertoire, expression profile or other patient-specific features.

Although unbiased immunogenicity datasets are beginning to emerge in increasing volumes, additional steps are likely to be required to achieve robust neoantigen prediction.

Going beyond peptide–MHC presentation

The formation of peptide–MHC (pMHC) complexes is a prerequisite for T cell immunogenicity; however, given the numerous factors that affect neoantigen immunogenicity beyond pMHC complex formation, there is a need for a holistic, multimodal approach to immunogenicity modeling that goes beyond peptide–MHC pairings (Fig. 1b).

Improved immunogenicity prediction of antigens has been demonstrated in models utilizing existing datasets of known immunogenic pMHC combinations alone. Additional features such as the stability of a pMHC complex, physiochemical properties of immunogenic epitopes and HLA-specific anchor locations make it possible to learn from either feature engineering[5] or transfer learning techniques[6]. Features such as pMHC complex stability can be predicted from pMHCs and included

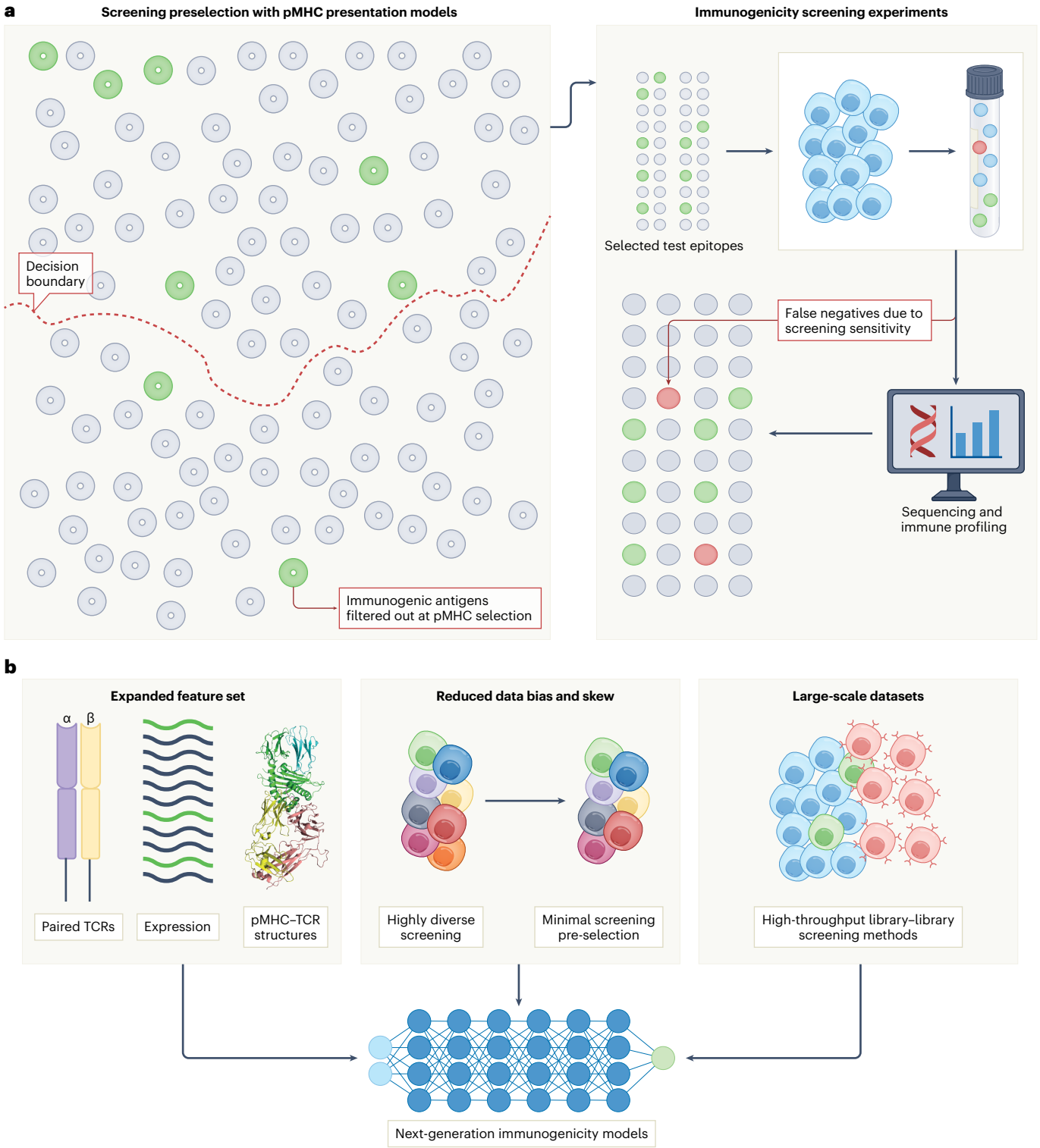

Fig. 1 | Challenges and future directions in cancer neoantigen-prediction methods. a, Representation of two categories of issues with current data sources. Pre-selection with pMHC models before immunogenicity screening removes many possible variants before testing. This creates a carried-over bias from existing models due to untested targets. Immunogenicity screening methods currently have low sensitivity, resulting in a high possibility of false negatives from screening experiments. **b**, Factors contributing to improved neoantigen prediction in future models. Expanded feature sets are necessary to give a full picture of immunogenicity, providing important context and functional information beyond pMHC inputs. New methods of immunogenicity screening can reduce bias in future datasets by screening a broader set of targets. New screening technologies make large-scale datasets possible.

in downstream models as additional inputs[7], but this requires specific training datasets. With larger volumes of data, these methods are likely to provide improved immunogenicity prediction compared to pMHC binding models; however, inputs beyond the peptide–MHC pairing are required to get a full picture of antigen immunogenicity.

Structural modeling of pMHC class I complexes interacting with their cognate T cell receptors (TCRs) can complement immunogenicity-prediction efforts. Although the number of experimentally derived structures available for training is limited, this may be addressed by improvements in the accuracy of models that predict the 3D structure of a protein. This strategy has yielded some early successes, with adaptions of AlphaFold being used to predict pMHC binding[8]. Going beyond pMHC binding with existing datasets remains challenging because of the variability of neoantigens — especially if TCR binding is being modeled. Moreover, it is noteworthy that current structural models generate a fixed structure, which may not provide the types of features necessary for a functional classification such as immunogenicity.

To improve the practical usefulness of prediction models in a cancer neoantigen setting, it may be necessary to include patient-specific antigenicity features in future models alongside global immunogenicity features. If the goal is to predict neoantigens that could serve as viable treatment targets, there are numerous tumor-intrinsic features that could influence target selection, including antigen-presentation capability, sufficient neoantigen expression levels and clonality within the tumor. Neoantigen expression is required for T cell recognition, with several models demonstrating the value of including expression levels[7]. However, immunogenicity assays do not directly measure antigen expression features, and most publicly available datasets do not include it, complicating training and benchmarking efforts. Similarly, the prevalence of a neoantigen within the tumor cell population (that is, antigen clonality) also contributes to the ability to mount antigen-specific T cell responses[9], but this information is not routinely incorporated into existing modeling efforts. T cell immune evasion, a feature of many advanced tumors, is also relevant to future neoantigen-prediction models. Direct loss of neoantigen recognition can occur via neoantigen loss or via compromised antigen presentation[10]. Interestingly, HLA loss is frequently subclonal[11], which would render clonal neoantigens presented by the lost HLA allele effectively subclonal. HLA loss is frequently observed following resistance to personalized immunotherapy. This selective pressure may be mitigated by simultaneously targeting multiple neoantigens presented by diverse HLA alleles, and by accounting for compromised antigen presentation as a component in target selection. In this regard, one could favor neoantigens predicted to be presented by extant HLA alleles and redundantly across multiple HLA alleles. Moreover, a strategy should not ignore the potential for MHC class II presentation.

Finally, improved prediction of immunogenic neoantigens can only be achieved by factoring in donor-specific TCR repertoires. It would also be beneficial to include additional related patient-specific information, such as the full HLA background, in future datasets to give a broader description of the TCR landscape, informing patient responses[12]. Predicting which TCRs' post-thymic selection are likely to elicit immune responses against neoantigens defines true immunogenicity. There are currently very limited data available that incorporate fully paired TCR chains, mostly focused on a handful of viral epitopes. As a result, TCR specificity models generalize poorly to unseen epitopes, confirmed in benchmarking studies[13]. Benchmarking with additional inputs, such as TCRs, highlights the issue of data leakage between training and test sets, with cross-contamination by similar epitopes between sets resulting in artificially inflated performance metrics. There are 1,409 TCRs with known reactivity to HLA class I-presented neoantigens in the CEDAR database[2] (accessed in September 2023), limiting any neoepitope-specific predictions. Generalization to unseen epitopes would be required for mostly private neoantigens. Without a TCR component, only part of the picture is being considered by the model, limiting potential accuracy. With new single-cell screening techniques, the volume and quality of paired TCR data will continue to increase — allowing paired TCR–peptide–HLA triplet data to be more easily included in antigen-immunogenicity-prediction models. Importantly, this will help to assign orphan TCRs to their cognate antigens, which could advance tumor-infiltrating lymphocyte (TIL)-based adoptive cell therapies.

Improved screening methods to generate large datasets

Rapid innovation in immunogenicity screening technology is essential for enabling the production of data at higher throughput. To date, improved patient-specific screening techniques such as HANSolo[14] allow faster and lower-bias neoantigen screening by reducing the need for pre-selection. Next-generation multimer assays permit high-throughput screening with high specificity[4]. Library-on-library screening techniques aim to perform screening on much larger scales than previously seen, with libraries of sizes up to 10^5 targets being reported. Methods such as ENTER-seq[15] allow additional features such as cell state to be measured during single-cell screening. This is beginning to shift the data-production bottleneck from assay screening capacity to sample acquisition.

Multimodal screening technologies will be important to both increase the volume of neoantigen datasets and improve the quality of data. Several newer screening methods will allow large increases in the number of known immunogenic pMHC complexes, the main inputs used by existing immunogenicity models. Increased data volume alone has been shown to improve performance and allow improved generalization to unseen targets, such as SARS-CoV-2[5]. Increases in data volumes will continue to improve models, but these improvements will become marginal as more immunogenic neoantigens are identified. Other screening methods can expand the depth of available features, for example allowing larger-scale datasets that include TCR specificities. New tools promise to give us a refined view of immunogenicity, factoring in additional features such as cell state[15], which will be important to future modeling efforts.

In building robust neoantigen immunogenicity-prediction models, data from both healthy donor sources and cancer-specific data sources will be useful. In vitro-stimulated T cells from healthy donors provide an abundant, controlled source of functional T cells, avoiding the patient-specific confounders seen in cancer patients. Datasets from patients with cancer, especially when accompanied by trial outcome information, are equally important in predictive model development. Although likely to be available in smaller numbers long term, they are required to properly test model efficacy and explore cancer-specific variance not explained by future models to identify additional features.

Conclusions and the future of immunogenicity prediction

Large datasets with low bias in generation streams will be necessary to build robust prediction algorithms. Though there have been many improvements in modeling methodologies in recent years, the limited distribution and volume of data points means that we expect currently available models to underperform on new datasets.

The current reliance on pMHC inputs in most modeling approaches reflects the available data. Other input sources are being investigated, but data sizes remain very small. It is our view that other inputs, such as TCR specificity and expression information, are required to come close to a full picture of the immunogenic potential of a neoantigen. Increases in the availability of immunogenic pMHC for neoantigens will undoubtedly improve performance in the short term, but there are limits to what can be learned from the pMHC alone. Future reliable prediction algorithms will require a wider variety of features in high volume to be effective.

Finally, it is essential that the community continues its commitment to sharing and harmonizing these critical datasets. A combined input approach to immunogenicity modeling is necessary, along with large-scale data gathering and sharing.

Hugh O'Brien[1], Max Salm[1] ✉, Laura Morton[1], Maciej Szukszto[2], Felix O'Farrell[1], Charlotte Boulton[2], Pablo D. Becker[1], Yardena Samuels[3], Charles Swanton[4], Marc R. Mansour[2], Sine Reker Hadrup[5] ✉ & Sergio A. Quezada[1,2] ✉

[1]Achilles Therapeutics Ltd, London, UK. [2]Cancer Immunology Unit, Research Department of Haematology, University College London Cancer Institute, London, UK. [3]Weizmann Institute of Science, Rehovot, Israel. [4]The Francis Crick Institute, London, UK. [5]Technical University of Denmark, Lyngby, Denmark.

✉e-mail: m.salm@achillestx.com; sirha@dtu.dk; s.quezada@ucl.ac.uk

Published online: 15 December 2023

References

1. Rosenberg, S. A., Parkhurst, M. R. & Robbins, P. F. *Cancer Cell* **41**, 646–648 (2023).
2. Koşaloğlu-Yalçın, Z. et al. *Nucleic Acids Res.* **51**, D845–D852 (2022).
3. Bruno, P. M. et al. *Nat. Biotechnol.* **1**, 980–992 (2023).
4. Holm, J. S. et al. *Nat. Commun.* **13**, 1935 (2022).
5. Gfeller, D. et al. *Cell Syst.* **14**, 72–83.e5 (2023).
6. Albert, B. A. et al. *Nat. Mach. Intell.* **5**, 861–872 (2023).
7. Gartner, J. J. et al. *Nat. Cancer* **2**, 563–574 (2021).
8. Motmaen, A. et al. *Proc. Natl Acad. Sci. USA* **120**, e2216697120 (2023).
9. Wolf, Y. et al. *Cell* **179**, 219–235.e21 (2019).
10. Martínez-Jiménez, F. et al. *Nat. Genet.* **55**, 820–831 (2023).
11. McGranahan, N. et al. *Cell* **171**, 1259–1271.e11 (2017).
12. Chowell, D. et al. *Nat. Med.* **25**, 1715–1720 (2019).
13. Deng, L. et al. *Front. Immunol.* https://doi.org/10.3389/fimmu.2023.1128326 (2023).
14. Cattaneo, C. M. et al. *Nat. Biotechnol.* **41**, 783–787 (2023).
15. Yu, B. et al. *Cell* **185**, 4904–4920.e22 (2022).

Competing interests

S.A.Q. is co-founder and chief scientific officer and own shares in Achilles Therapeutics. C.S. acknowledges grants from AstraZeneca, Boehringer-Ingelheim, Bristol Myers Squibb, Pfizer, Roche-Ventana, Invitae (previously Archer Dx Inc., a collaboration in minimal residual disease sequencing technologies), Ono Pharmaceutical and Personalis. He is chief investigator for the AZ MeRmaiD 1 and 2 clinical trials and is the steering committee chair. He is also co-chief investigator of the NHS Galleri trial funded by GRAIL and a paid member of GRAIL's scientific advisory board (SAB). He receives consultant fees from Achilles Therapeutics (also an SAB member); Bicycle Therapeutics (also an SAB member); Genentech; Medicxi; the China Innovation Centre of Roche (CICoR), formerly Roche Innovation Centre Shanghai; Metabomed (until July 2022); Relay Therapeutics; and the Sarah Cannon Research Institute. C.S has received honoraria from Amgen, AstraZeneca, Bristol Myers Squibb, GlaxoSmithKline, Illumina, MSD, Novartis, Pfizer and Roche-Ventana; previously held stock options in Apogen Biotechnologies and GRAIL; currently has stock options in Epic Bioscience and Bicycle Therapeutics; and has stock options and is co-founder of Achilles Therapeutics. S.R.H. is the cofounder of PokeAcell and is co-inventor of licensed patents related to T cell detection. H.O'B., M.S., L.M. and F.O'F. are employees of Achilles Therapeutics.