Resolution of recursive data corruption to transform T-cell epitope discovery

Grzegorz Preibisch[*,1,2], Michał Tyrolski[*,1#], Piotr Kucharski[1],
Stanislaw Giziński[1], Piotr Grzegorczyk[1]
Sungho Moon [4], Sangwoo Kim [4],
Balyn Zaro[3], Anna Gambin[2]

Abstract

Accurate prediction of MHC class I-presented peptides is essential for any vaccine or T-cell therapy design, yet reported gains on in silico benchmarks have not translated into clinical successes. Here we show that this discrepancy may come from a common methodological error: immunopeptidomics datasets are fundamentally contaminated by existing prediction models through prediction-based deconvolution and filtering, resulting in an iterative confirmation bias. An audit of the IEDB, the biggest database in the field, reveals that as of January 2025, 55.8% of assessable data are labeled by computational models rather than verified experimentally. This inflates in silico benchmarks while degrading real-world applicability on new data, effectively making it impossible to objectively test model performance, which can lead to choosing suboptimal solutions and decreasing the chance of any therapy's clinical success. In silico simulation shows that iterative data corruption maintains high AUROC while top-of-list retrieval collapses. We reframe epitope discovery as a protein-centric learning-to-rank task and introduce deepMHCflare, a model evaluated exclusively on clean data. deepMHCflare achieves 0.80 Precision@4 on mono-allelic benchmarks versus 0.55-0.65 for gold-standard prediction models. A preclinical cancer vaccine study validated that 2 of the 4 deepMHCflare-nominated peptides were immunogenic, with a third independently confirmed in the literature.

Selecting peptide candidates for cancer vaccines and T-cell therapies remains a principal bottleneck in translational immunology [1–3]. Computational predictors are routinely used to rank peptides for synthesis and immunogenicity testing, yet improvements reported on public benchmarks have not consistently translated into higher prospective yields [2, 3]. This discrepancy is commonly attributed to biological complexity, but it also depends on how training and evaluation data are generated and curated.

Most immunopeptidomics experiments are performed on cells expressing multiple HLA alleles, and assigning each eluted peptide to its presenting allele requires computational deconvolution [4]. In practice, this deconvolution frequently relies on predictions from existing models, either explicitly, through prediction-guided pseudo-labelling [5], or indirectly, through iterative motif convergence [4]. As a result, peptide labels in public repositories become correlated with prior model outputs, as researchers were using those methods to label the data. Consequently, peptides matching existing expectations are preferentially retained, while discordant peptides are depleted. We term this recursive process *systematic confirmation bias* (Figure 1). Among current tools, the

[1]Deepflare, Warsaw, Poland.
[2]University of Warsaw, Department of Mathematics, Informatics and Mechanics, Warsaw, Poland.
[3]Department of Pharmaceutical Chemistry, University of California San Francisco, San Francisco, United States.
[4]Progeneer, Seoul, South Korea.
[*]Equal contribution.
[#]Currently an independent researcher. Contributions to this work were made while affiliated with Deepflare.

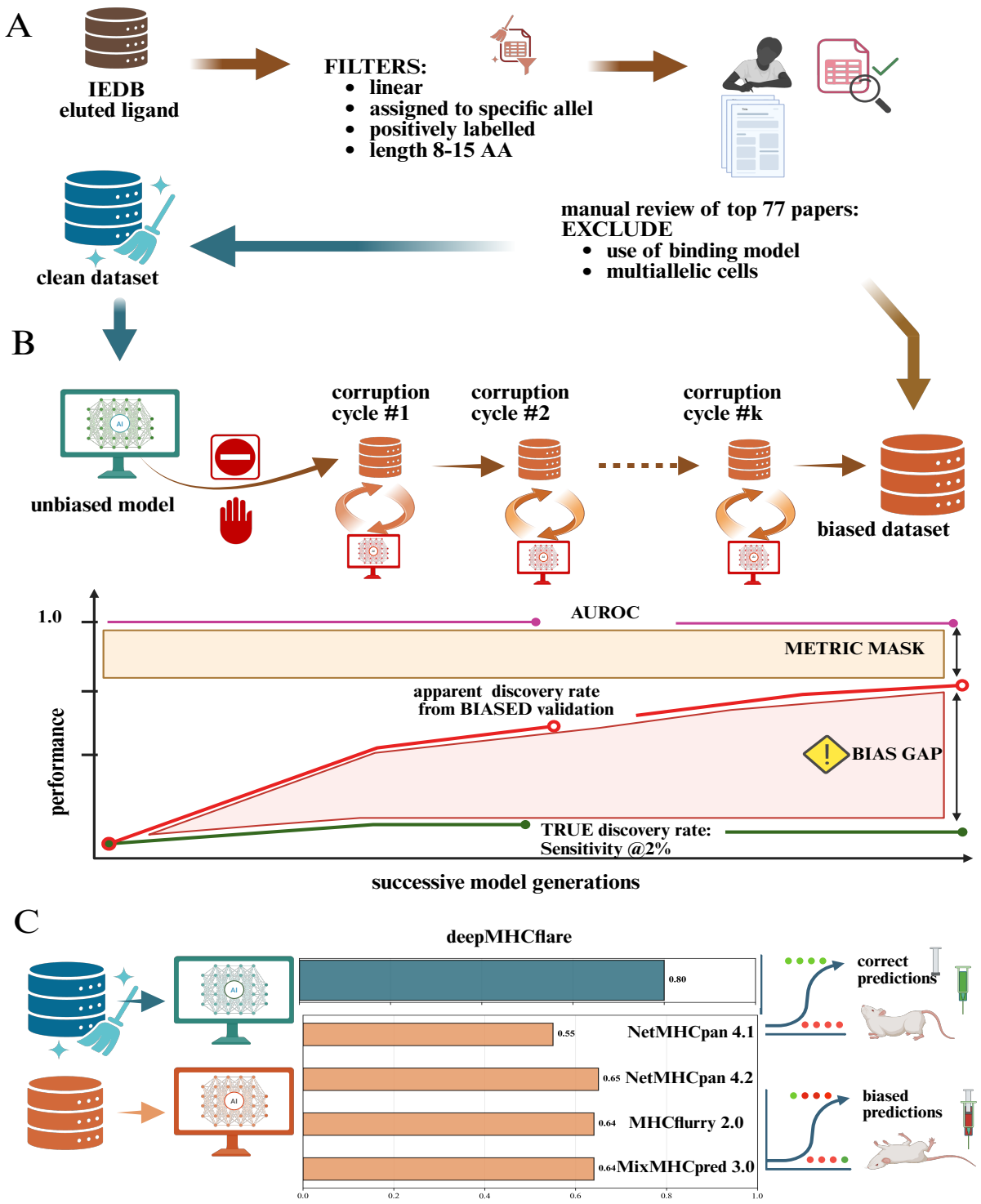


Figure 1: Systematic confirmation bias causes hidden performance stagnation. (A) The IEDB database was filtered according to the given criteria, and then, based on manual review of the publications that provided the largest amount of data, a clean dataset was constructed. The remaining part of the database is referred to as the biased dataset. (B) Consecutive cycles of model and data corruption and their performance. True Discovery Rate (green) stagnates at approximately 0.23 while Apparent Discovery Rate (red) rises to 0.58, creating a performance illusion (pink shaded area, called bisd gap). AUROC (magenta) remains above 0.89 throughout, masking the stagnation (yellow area). (C) The R-precision@4 achieved by DeepMHCflare and other (biased) models on on mono-allelic immunopeptidomics data. Created in https://BioRender.com

neural-network-based predictors NetMHCpan 4.1 [6, 7] and MHCflurry 2.0 [8, 9] exert influence through supervised deconvolution, while the motif-based MixMHCpred 3.0 [10, 11] operates through unsupervised motif convergence.

This problem is compounded by evaluation practice. Epitope discovery is a ranking problem under tight experimental budgets, where only a handful of top-ranked candidates per antigen can be tested [12]. Yet model assessment has relied predominantly on AUROC (area under the receiver operating characteristic curve), a global discrimination metric equivalent to the probability that a randomly chosen positive is scored higher than a randomly chosen negative. On a balanced dataset this is a natural measure of quality. However, in immunopeptidomics, where a single true binder must be found among 500 or more decoys, an AUROC of 0.95 means the true binder sits on average at position 25 – well outside the handful of candidates an experimentalist can afford to synthesise and test. AUROC is therefore insensitive to changes in top-of-list ordering on highly imbalanced datasets [13]. This effect is visible in the cancer vaccine case study (Figure 2C). Both models achieve similar AUROC, yet deepMHCflare concentrates true binders at the top of the ranking while NetMHCpan does not. Under predictor-dependent validation, AUROC can remain high even as enrichment among actionable candidates deteriorates (Figure 1B).

Here, we present deepMHCflare, a protein-centric learning-to-rank predictor for MHC class I presentation, building on our earlier work [14]. We first audit the Immune Epitope Database (IEDB) [15] to quantify the prevalence of predictor dependence in public immunopeptidomics data. We then use in silico simulation to demonstrate how recursive data corruption decouples standard metrics from real predictive performance. Finally, we evaluate deepMHCflare exclusively on predictor-independent benchmarks and in a prospective cancer vaccine study, showing that the combination of predictor-independent training data, a protein-language-model backbone, and a ranking-aware training objective yields substantial improvements over established methods.

To quantify the extent of predictor dependence in public immunopeptidomics data, we audited all mass spectrometry-based eluted-ligand entries in the Immune Epitope Database (IEDB; January 2025 snapshot) [15], totalling 3,971,663 records. A two-stage protocol, programmatic SQL classification followed by systematic review of 37 publications, assigned each entry to one of four categories: *Clean* (experimentally resolved allele assignments from mono-allelic cell lines or allele-specific antibody pull-downs), *Biased* (allele labels assigned or confirmed by computational predictors), *Multi-allelic* (valid peptides from multi-allelic sources lacking allele-level resolution), and *Insufficient metadata* (entries whose provenance could not be determined). In total, only **44.2%** of assessable entries (1,511,463) were verified to be *clean* (Extended Data Tables 1a–c). Among the 3,417,839 entries where provenance could be determined, **55.8%** (1,906,376) carry predictor-dependent labels. The remaining *multi-allelic* and *insufficient*-metadata fractions, over one million entries each, cannot be used for allele-specific training without computational deconvolution, which reintroduces the same predictor dependence. These categories reflect fundamentally different degrees of predictor dependence in the provenance chain of each measurement. Based on this audit, we constructed a predictor-independent evaluation benchmark from the resolved mono-allelic subset [16, 17], enabling model selection and validation free from recursive contamination, cf. Figure 1A.

We designed an in silico simulation to quantify the effect of recursive data corruption on model performance (see Online Methods). Starting from a baseline model trained on clean mono-allelic data [16], we initiated an iterative cycle. Each generation's model was used to filter an unlabeled dataset, and the filtered data, enriched with the model's own biases, were used to train the next generation. This process was replicated five times with independent model initialisations to isolate the effect of data corruption from model training stochasticity, confirming that the observed performance degradation is driven by systematic bias rather than random variation in model training.

The results revealed a striking divergence as illustrated in Figure 1B. When evaluated on the

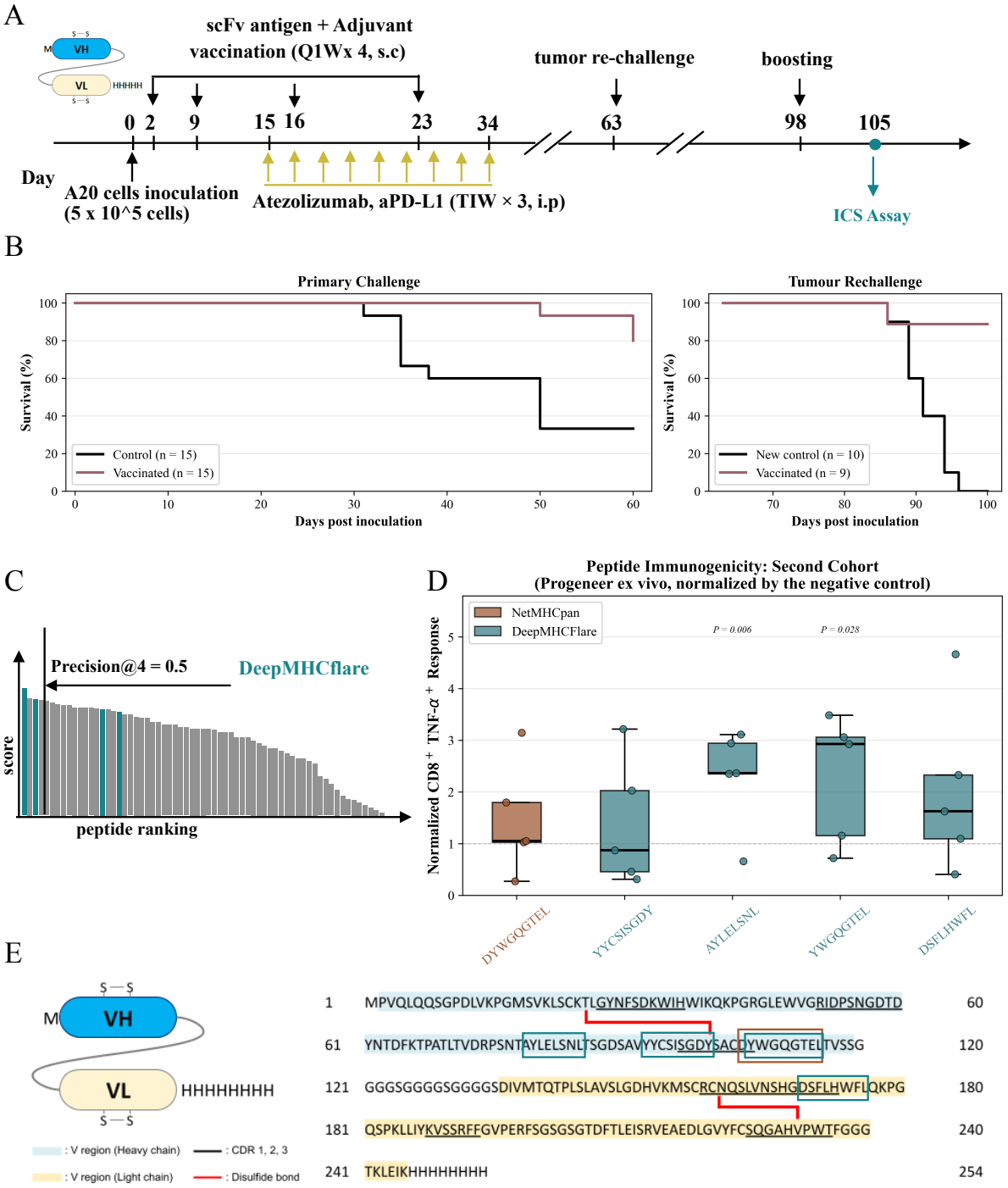


Figure 2: **Preclinical cancer vaccine study.** (A) Experimental timeline. BALB/c mice ($n = 15$ per arm) were inoculated with A20 cells, vaccinated with adjuvanted A20-scFv and Atezolizumab. Surviving mice were re-challenged on day 63, boosted on day 98, and splenocytes harvested for ICS on day 105 ($n = 5$). (B) Kaplan–Meier survival curves for primary challenge ($n = 15$ per arm) and tumour rechallenge (9 vaccinated survivors versus 10 naïve controls; log-rank $P < 0.01$ for both comparisons). (C) Illustrative peptide rankings for two models with similar AUROC; coloured bars mark validated epitopes, showing superior top-of-list enrichment by deepMHCflare. (D) Normalised CD8$^+$ TNF-α^+ responses (ICS, fold-change over negative control) for deepMHCflare-selected peptides. Exact p-values shown (one-sided paired t-test). (E) A20-scFv sequence with tested peptides boxed; heavy-chain (cyan) and light-chain (gold) variable domains and CDR loops shown in relevant colours.

corrupted data, models appeared to improve with each generation. AUROC remained above 0.89 across all iterations. To quantify the actual impact on discovery, we measured Sensitivity@Top2% – the fraction of the true peptidome that researchers can discover when applying the standard 2% rank threshold used by widely adopted prediction tools [6, 9] and commonly adopted by experimentalists selecting candidates for validation [18, 19] (see Online Methods). When evaluated against the clean, predictor-independent validation set, Sensitivity@Top2% declined substantially with each corruption iteration. Apparent discovery rate rose to 0.58 even as the true rate stagnated at 0.23, close to the 0.11 sensitivity expected from random selection on this validation set, cf. Figure 1B. High AUROC thus masked a near-complete collapse in the ability to retrieve actionable candidates. Together, biased data and a misleading metric create an illusion of progress. Models increasingly predict what was previously known rather than enabling new discoveries, and AUROC masks this stagnation [13].

We evaluated deepMHCflare against the most widely used prediction tools in the field, and the primary contributors to the data bias described above, NetMHCpan 4.1 [6], NetMHCpan 4.2 [7], MHCflurry 2.0 [9], and MixMHCpred 3.0 [11], on a held-out set of mono-allelic cell lines [16]. Specifically, we measured performance using Precision@4, the fraction of true positives among the top four ranked candidates per protein. Such a measurement reflects the common experimental budget of selecting a small number of high-confidence candidates for validation. For deepMHCflare, training and evaluation used non-overlapping partitions of the same mono-allelic dataset, split by source antigen to prevent train–test leakage (see Online Methods). For competing models, such a guarantee cannot be made, as these datasets have been publicly available for several years and overlap with data sources used in their training pipelines. Therefore, any such overlap would favour the competing models, making our results a conservative estimate of the true performance difference. On this clean benchmark, deepMHCflare achieved 0.80 Precision@4, a 23-45% improvement over all established methods (Figure 1C).

We conducted additional ablation studies to assess generalisation. First, we held out 21 alleles entirely from training. The model retained meaningful ranking ability on these unseen alleles, confirming that it captures transferable binding patterns rather than memorising allele-specific features. Second, we tested generalisation on the multi-allelic HLA Ligand Atlas [20] , an unseen, out-of-distribution dataset of over 90,000 ligands from 227 patient tissue samples. deepMHCflare generalised to this clinically realistic regime. We then retrained the model on additional clean data to quantify the improvement (Supplementary Note; see Online Methods).

To assess whether benchmark gains translate to a practical setting, we conducted a prospective cancer vaccine study in the A20 BALB/c murine lymphoma model (Figure 2A). Vaccinated mice showed significantly prolonged survival compared with controls in both the primary challenge and tumour rechallenge (Figure 2B). deepMHCflare ranked candidate peptides from an A20 B-cell lymphoma scFv antigen [21, 22] for the H-2K^{d} allele; despite comparable AUROC, deepMHCflare concentrated validated epitopes near the top of the ranked list (Figure 2C). The top four peptides were synthesized and tested for immunogenicity by intracellular cytokine staining (see Online Methods for details).

Two of four deepMHCflare-selected peptides elicited significant CD8^{+} TNF-α^{+} responses ($P = 0.006$ and $P = 0.028$, one-sided paired t-test; Figure 2D), and a third (YYCSISGDY) was independently reported as the only tumour-specific CDR3-derived epitope from A20 [21] (Figure 2E; Extended Data Table 6).

The confirmation bias we describe may extend beyond MHC prediction. Any setting in which model outputs shape subsequent training data is inherently susceptible to recursive contamination [23]. As computational methods become embedded in experimental pipelines across biology, systematic auditing of data provenance should become standard practice. Prospective validation across diverse antigens and HLA haplotypes will be required to establish the generality and practical impact of the effects observed here.

References

[1] Ziyang Zhang, Peter J. Rohweder, Chayanid Ongpipattanakul, Koli Basu, Markus-Frederik Bohn, Elizabeth Dugan, Veronica Steri, Byron C. Hann, Kevan M. Shokat, and Charles S. Craik. A covalent inhibitor of k-ras(g12c) induces mhc-i presentation of haptenated peptide neoepitopes targetable by immunotherapy. *Cancer cell*, 40:1060 – 1069.e7, 2022. URL `https://api.semanticscholar.org/CorpusID:252219723`.

[2] Pramod K Srivastava. Cancer neoepitopes viewed through negative selection and peripheral tolerance: a new path to cancer vaccines. *The Journal of Clinical Investigation*, 134, 2024. URL `https://api.semanticscholar.org/CorpusID:268081279`.

[3] Daniel K Wells, Marit M van Buuren, Kristen K Dang, Vanessa E Huber, David G Bates, et al. Key parameters of tumor epitope immunogenicity revealed through a consortium approach improve neoantigen prediction. *Cell*, 183(3):818–834.e13, 2020. doi: 10.1016/j.cell.2020.09.015.

[4] Michal Bassani-Sternberg and David Gfeller. Deciphering hla-i motifs across hla peptidomes improves neo-antigen predictions and identifies allostery regulating hla specificity. *PLoS Computational Biology*, 13(8):e1005725, 2017. doi: 10.1371/journal.pcbi.1005725.

[5] Bruno Alvarez, Birkir Reynisson, Carolina Barra, Soren Buus, Nicola Ternette, Tim Connelley, Massimo Andreatta, and Morten Nielsen. Nnalign_ma; mhc peptidome deconvolution for accurate mhc binding motif characterization and improved t-cell epitope predictions. *Molecular & Cellular Proteomics*, 18(12):2459–2477, 2019. doi: 10.1074/mcp.TIR119.001658.

[6] Birkir Reynisson, Bruno Alvarez, Sinu Paul, Bjoern Peters, and Morten Nielsen. Netmhcpan-4.1 and netmhciipan-4.0: improved predictions of mhc antigen presentation by concurrent motif deconvolution and integration of ms mhc eluted ligand data. *Nucleic Acids Research*, 48(W1):W449–W454, 2020. doi: 10.1093/nar/gkaa379.

[7] Jonas Birkelund Nilsson, Birkir Reynisson, and Morten Nielsen. Netmhcpan-4.2: improved prediction of cd8+ epitopes by use of transfer learning and structural features. *Frontiers in Immunology*, 16:1616113, 2025. doi: 10.3389/fimmu.2025.1616113.

[8] Timothy J O'Donnell, Alex Rubinsteyn, Maria Bonsack, Angelika B Riemer, Uri Laserson, and Jeff Hammerbacher. Mhcflurry: Open-source class i mhc binding affinity prediction. *Cell Systems*, 7(1):129–132.e4, 2018. doi: 10.1016/j.cels.2018.05.014.

[9] Timothy J O'Donnell, Alex Rubinsteyn, and Uri Laserson. Mhcflurry 2.0: Improved pan-allele prediction of mhc class i-presented peptides by incorporating antigen processing. *Cell Systems*, 11(1):42–48.e7, 2020. doi: 10.1016/j.cels.2020.06.010.

[10] David Gfeller, Philippe Guillaume, Justine Michaux, Julien Racle, et al. The length distribution and multiple specificity of naturally presented hla-i ligands. *The Journal of Immunology*, 201(12):3705–3716, 2018. doi: 10.4049/jimmunol.1800914.

[11] Daniel M Tadros, Julien Racle, and David Gfeller. Predicting mhc-i ligands across alleles and species: how far can we go? *Genome Medicine*, 17:1–15, 2025. doi: 10.1186/s13073-025-01450-8.

[12] Tie-Yan Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331, 2009. doi: 10.1561/1500000016.

[13] Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3):e0118432, 2015. doi: 10.1371/journal.pone.0118432.

[14] Stanisław Giziński, Grzegorz Preibisch, Piotr Kucharski, Michał Tyrolski, Michał Rembalski, Piotr Grzegorczyk, and Anna Gambin. Enhancing antigenic peptide discovery: Improved mhc-i binding prediction and methodology. *Methods*, 224:1–9, 2024.

[15] Randi Vita, Swapnil Mahajan, James A Overton, Sandeep Kumar Dhanda, Sheridan Martini, Jason R Cantrell, Daniel K Wheeler, Alessandro Sette, and Bjoern Peters. The immune epitope database (iedb): 2018 update. *Nucleic Acids Research*, 47(D1):D339–D343, 2019. doi: 10.1093/nar/gky1006.

[16] Siranush Sarkizova, Susan Klaeger, Phuong M Le, Letitia W Li, Giacomo Hartmann, Wandi Zhang, Timothy J O'Donnell, et al. A large peptidome dataset improves hla class i epitope prediction across most of the human population. *Nature Biotechnology*, 38(2):199–209, 2020. doi: 10.1038/s41587-019-0322-9.

[17] Pouya Faridi, Chen Li, Sri H Ramarathinam, Julian P Vivian, Patricia T Illing, Nicole A Mifsud, Rochelle Ayala, Jiangning Song, Stephanie Gras, and Anthony W Purcell. A subset of hla-i peptides are not genomically templated: Evidence for cis-and trans-spliced peptide ligands. *Science Immunology*, 3(28):eaar3947, 2018. doi: 10.1126/sciimmunol.aar3947.

[18] Greta Forlani, Maryam Shallak, Francesca Celesti, and Roberto S Accolla. Ciita-transduced glioblastoma cells uncover a rich repertoire of clinically relevant tumor-associated hla-ii antigens. *Molecular & Cellular Proteomics*, 20:100032, 2021. doi: 10.1074/mcp.RA120.002201.

[19] Gabriel Goncalves, Kerry A Mullan, Divya Duscharla, Rochelle Ayala, Nathan P Croft, Pouya Faridi, and Anthony W Purcell. Ifnγ modulates the immunopeptidome of triple negative breast cancer cells by enhancing and diversifying antigen processing and presentation. *Frontiers in Immunology*, 12:645770, 2021.

[20] Ana Marcu, Leon Bichmann, Leon Kuchenbecker, Daniel Johannes Kowalewski, Lena Katharina Freudenmann, Linus Backert, Lena Mühlenbruch, András Szolek, Maren Lübke, Philipp Wagner, et al. Hla ligand atlas: a benign reference of hla-presented peptides to improve t-cell-based cancer immunotherapy. *Journal for immunotherapy of cancer*, 9(4):e002071, 2021.

[21] Max Warncke, Maike Buchner, G. Thaller, Anna Dodero, Alla Bulashevska, Dietmar Pfeifer, Jens Timmer, and Hendrik Veelken. Control of the specificity of t cell-mediated anti-idiotype immunity by natural regulatory t cells. *Cancer Immunology, Immunotherapy : CII*, 60:49 – 60, 2010. URL `https://api.semanticscholar.org/CorpusID:9937123`.

[22] K Jin Kim, Colette Kanellopoulos-Langevin, Ruth M Merwin, David H Sachs, and Richard Asofsky. Establishment and characterization of balb/c lymphoma lines with b cell properties. *The Journal of Immunology*, 122(2):549–554, 02 1979. ISSN 0022-1767. doi: 10.4049/jimmunol.122.2.549. URL `https://doi.org/10.4049/jimmunol.122.2.549`.

[23] David Graber, Peter Stockinger, Fabian Meyer, Siddhartha Mishra, Claus Horn, and Rebecca Buller. Resolving data bias improves generalization in binding affinity prediction. *Nature Machine Intelligence*, pages 1–13, 2025. doi: 10.1038/s42256-025-01124-5.

[24] P Parham and WF Bodmer. Monoclonal antibody to a human histocompatibility alloantigen, hla-a2. *Nature*, 276(5686):397–399, 1978. doi: 10.1038/276397a0.

[25] Christopher J. C. Burges, Robert J. Ragno, and Quoc V. Le. Learning to rank with nonsmooth cost functions. In *Neural Information Processing Systems*, 2006. URL `https://api.semanticscholar.org/CorpusID:8604596`.

[26] Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large batch optimization for deep learning: Training bert in 76 minutes. *arXiv preprint arXiv:1904.00962*, 2019.

[27] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637): 1123–1130, 2023. doi: 10.1126/science.ade2574.

[28] Patrick Kunzmann and Kay Hamacher. Biotite: a unifying open source computational biology framework in python. *BMC bioinformatics*, 19(1):346, 2018. doi: 10.1186/s12859-018-2367-z.

Acknowledgements

We gratefully acknowledge Poland's high-performance computing infrastructure PLGrid for providing computer facilities and support within computational grant no. PLG/2025/018310. We also acknowledge PCSS for providing access to the Eagle HPC cluster within computational grant no. pl0418-01. AG was supported by the National Science Centre, Poland, project 2024/53/B/ST6/03852.

Author Contributions

G.P. conceived the learning-to-rank methodology, identified and quantified the recursive data bias, designed and executed the in silico bias-simulation experiments, trained production models, performed data analysis, participated in peptide selection for the prospective validation study, and led the writing of the manuscript. M.T. engineered the machine learning pipeline, developed the model architecture and training methodology, ran model training experiments, and contributed to writing and editing the manuscript. P.K. engineered the machine learning pipeline, provided technological oversight, and implemented evaluation metrics. S.Gi. engineered the machine learning pipeline, participated in peptide selection for the prospective validation study, contributed to training early model versions, and was involved in early methodology development. P.G. performed the final quantification of data bias in the IEDB audit. S.M. and S.K. designed and carried out the experimental vaccination and intracellular cytokine staining assays. B.Z. supervised experimental mass spectrometry methodology. A.G. supervised computational and statistical methodology.

Competing Interests

G.P., M.T., P.K., P.G., and S.Gi. are or were employees of Deepflare. B.Z. and A.G. serve as scientific advisors to Deepflare. S.M. and S.K. are employees of Progeneer. Deepflare and Progeneer may benefit commercially from the methods described in this work.

Data and Code Availability

Source code, trained model weights, and raw experimental data will be made available upon publication of the peer-reviewed article.

Online Methods

Data Curation and Quality Audit

The Immune Epitope Database (IEDB) [15] provides the foundational training data for virtually all MHC binding prediction algorithms and, equally importantly, serves as a primary source of evidence about specific epitopes. Any bias or error in the database therefore impacts both prediction tools and researchers who search the database directly, assuming it provides only experimental data. To quantify what fraction of the data may be impacted by bias, we developed a two-stage protocol - programmatic SQL classification followed by systematic paper-level review of 37 publications - that categorised entries into four groups:

- **Clean.** Data in which no prediction was used to assign an allele to a peptide. This group comprises immunopeptidomics data from mono-allelic cell lines and from studies that used allele-specific antibodies on multi-allelic samples (e.g. BB7.2, which precipitates only peptides bound to HLA-A2 [24]), and entries confirmed clean by paper-level review.
- **Biased.** Entries whose allele labels were assigned or confirmed by computational predictors, including multi-allelic data subjected to computational deconvolution, rank-threshold filtering, or prediction-based quality control.
- **Multi-allelic.** Valid peptides from multi-allelic sources labelled only at the class level (e.g., "HLA class I"), which require computational deconvolution for allele-specific use.
- **Insufficient metadata.** Entries whose allele-assignment method could not be determined from the available metadata - typically antibody-based captures where host genotype information is missing.

The systematic paper-level review of 37 publications reclassified 1,038,004 entries as biased and 823,151 as clean beyond what the programmatic SQL classification identified. Among the 3,417,839 assessable entries (Clean + Biased), 55.8% carry predictor-dependent labels (Extended Data Tables 1a,b). Classification verdicts, entry counts, and evidence quotes for each reviewed publication are provided in Extended Data Table 1c.

Dataset description

We constructed a predictor-independent benchmark dataset free from the influence of existing model predictions through a multi-stage curation protocol applied to all eluted-ligand records classified as *Clean* in the IEDB as of January 2025. This dataset comprises 300,869 predictor-independent records from studies that passed all criteria, primarily from Sarkizova et al. [16] and Faridi et al. [17]. It was used both for validation of deepMHCflare and to simulate systematic confirmation bias.

In both cases, to avoid train–test leakage we separated data by holding out specific proteins from which peptides were or could have been derived. For deepMHCflare we additionally held out a set of alleles to perform ablation studies of the model's generalisation to unseen alleles. Remaining records were retained as additional training data representative of publicly available, potentially contaminated data, as the clean subset was a small fraction of the total and the rigorous train–test split already held out a substantial number of alleles and peptides.

For multi-allelic validation, we selected patient tissue samples from the HLA Ligand Atlas [20], comprising over 90,000 ligands from 227 patient tissue samples across 51 alleles. These samples provided a realistic multi-allelic evaluation setting in which the model must simultaneously discriminate binders across all alleles expressed by a given individual.

Evaluation metric

To simulate a budget-constrained setting where an experimentalist must select only four peptides, we define **Precision@4**. For each protein in the test set, we generated a candidate pool (1:512 positive-to-negative ratio) and ranked all candidates by predicted binding score on the mono-allelic benchmark. Precision@4 measures the fraction of true positives among the top four ranked candidates for each protein. For contexts with fewer than four positives, we use R-Precision (Precision@R, where R is the number of true positives) to avoid penalising the model for an impossible target.

Learning-to-Rank Framework for Protein-Level Epitope Discovery

We reframed MHC class I presentation prediction as a protein-centric learning-to-rank task. For each experimentally validated epitope, we identified its source protein and generated a candidate pool of all possible 8–15-mer peptides using a sliding window. From this pool, 128–256 peptides served as negatives for each positive. This context-aware hard-negative sampling ensured that negatives included peptides highly similar to the positive-including single-residue extensions, truncations, and overlapping sequences-forcing the model to learn fine-grained features that distinguish true epitopes from near-identical decoys. All sampled negatives were cross-referenced against the full set of known positives to prevent false negatives.

The training objective combines a pairwise margin ranking loss, weighted by LambdaRank importance [25], with weighted binary cross-entropy. For each positive peptide p and negative peptide n from the same source protein, the margin ranking loss is:

$$L_{\text{rank}} = \sum_{n \in N} \max\big(0,\ m - (s(p) - s(n))\big) \cdot |\Delta\text{nDCG@5}_{pn}| \tag{1}$$

where $s(\cdot)$ are model scores, m is a fixed margin, and $\Delta\text{nDCG@5}_{pn}$ is the change in normalised discounted cumulative gain from swapping positions of items p and n, so that correctly ordering the top positions contributes most to the gradient. The combined loss is:

$$L = 0.9 \cdot L_{\text{rank}} + 0.1 \cdot L_{\text{wBCE}} \tag{2}$$

where L_{wBCE} is binary cross-entropy with class weights $[0.1, 0.9]$, providing a stable classification signal under extreme class imbalance ($\sim$0.2% positive rate).

The model was trained with the LAMB optimiser [26] at a static learning rate of 0.003 with weight decay of 0.2. The per-device batch size was 1024 with 64 gradient-accumulation steps, yielding an effective batch size of 65,536. Training used bf16-mixed precision with PyTorch DDP for 100 epochs, with validation every 5 epochs. Full hyperparameters are provided in Extended Data Table 3.

Model architecture

deepMHCflare uses ESM2-t6-8M [27], a 6-layer, 8-million-parameter protein language model, as its backbone encoder. Each HLA allele is represented by a pseudo-sequence of approximately 182 amino acids comprising the alpha-1 and alpha-2 domains of the MHC class I heavy chain. Pseudo-sequences were extracted by pairwise alignment of full-length allele sequences against UniProt reference heavy chains using the biotite library [28]; the resulting database covers 14,856 HLA class I alleles. The MHC pseudo-sequence and candidate peptide (8–15 amino acids) are concatenated into a single input token sequence and processed jointly by the transformer encoder, so that the self-attention mechanism operates over both MHC and peptide positions.

Three pooling strategies are applied to the encoder's final-layer hidden states: mean pooling with square-root length normalisation, element-wise max pooling, and the `[CLS]` token embedding. The resulting vectors are concatenated into a 960-dimensional representation and passed through

a linear classification head (960 → 2) with sigmoid activation. Hidden dropout is set to 0.2; attention dropout is set to 0.0. The architecture is depicted in Extended Data Figure 3.

Simulation of systematic confirmation bias

The predictor-independent dataset was partitioned into K=4 non-overlapping folds based on MHC allele (23 alleles per fold), with no allele overlap between folds. Within each fold, data were split into training (80%) and validation (20%) subsets by holding out separate protein sources to prevent peptide-level leakage. Negative samples were sampled from the same source antigens as the positives at a fixed positive-to-negative ratio of 1:4 for training and 1:9 for validation, approximating the class balance used by established predictors. The procedure was replicated five times with independent model initialisations to isolate the effect of data corruption from model training stochasticity, confirming that the observed performance degradation is driven by systematic bias rather than random variation in model training.

The simulation proceeded through four iterations ($i \in \{0, 1, 2, 3\}$). Iteration 0 served as the clean baseline: model M_0 was trained on uncorrupted Fold 0. In each subsequent iteration: (1) model M_{i-1} predicted binding scores for Fold i; (2) a positive observation retained its label only if its predicted rank fell within the top 2% of all scores for that fold, and all others were relabelled as negative-this cutoff mirrored the standard rank threshold employed by NetMHCpan 4.1 [6] and reflected filtering criteria documented in several large IEDB datasets; (3) the corrupted training data from Fold i was merged with the complete training set from iteration $i-1$:

$$D_{\text{train}}^{(i)} = D_{\text{train}}^{(i-1)} \cup D_{\text{corrupted}}^{(i)} \tag{3}$$

(4) a new model M_i was trained on the accumulated dataset. Each model was evaluated on both the original predictor-independent validation set and a corrupted validation set created by applying the same corruption rule. To quantify the impact of the bias caused by the 2% threshold present in the vast majority of biased papers, we evaluated what fraction of the peptidome may be discovered by the model at that threshold. We define this metric as **Sensitivity@Top2%**: the fraction of true positives whose predicted rank falls within the top 2% of all candidates.

Prospective Validation Study Design

The prospective cancer vaccine study was conducted by Progeneer Inc. (Seoul, Republic of Korea). Candidate 8–9-mer peptides were predicted from the A20 B-cell lymphoma scFv antigen sequence using deepMHCflare. The top four ranked peptides were selected for testing. DYWGQGTEL, the top-ranked peptide predicted by NetMHCpan 4.1 [6], was additionally included as a reference control; notably, Warncke et al. [21] showed that this peptide induces suppressive $CD4^+$ rather than cytotoxic $CD8^+$ T-cell responses. Eight-week-old female BALB/c mice were inoculated subcutaneously with 5×10^5 A20 cells on day 0. The vaccinated group ($n = 20$) received four weekly subcutaneous doses of A20-scFv protein (50 µg) formulated with ProLNG-002 adjuvant (TLR4a 5 µg + TLR7/8a 70 µg; days 2, 9, 16, 23) combined with Atezolizumab administered intraperitoneally three times per week for three weeks (days 15–34). Vehicle controls received no treatment.

By day 34, tumor volumes in the control and vaccinated groups showed a statistically significant difference (Mann–Whitney test, $p < 0.0001$), demonstrating the vaccine's antitumor efficacy. At day 63, surviving mice were re-challenged with A20 cells in the contralateral flank; all re-challenged mice remained tumour-free, with an additional control group showing no survival. This confirmed a durable antitumour memory response. To identify which peptides drive this T-cell response, surviving mice received a single vaccine boost on day 98. One week later (day 105), splenocytes were harvested and stimulated with the four deepMHCflare-selected peptides and DYWGQGTEL as a reference control.

Spleens were dissociated using the Spleen Dissociation Kit (Miltenyi Biotec, 130-095-926) following the manufacturer's protocol. Tissues stored in RPMI-1640 were transferred to C-tubes and mechanically dissociated into single-cell suspensions using the GentleMACS Dissociator (Miltenyi Biotec). Cell suspensions were filtered to remove debris and treated with RBC lysis buffer at room temperature for 30 minutes. After washing, cells were resuspended in RPMI-1640 supplemented with 10% FBS, counted using a 3-FL counter (ThermoFisher).

Single-cell suspensions (1×10^6 cells/100 µL) were seeded into 96-well plates and stimulated with specific peptides (1 µg/mL each; Genscript) in the presence of Monensin (1:1000; Invitrogen, 00-4505-51) and Brefeldin A (1:1000; BioLegend, 420601) in RPMI-1640 medium (Welgene, LM011-03) supplemented with 25 mM HEPES, 50 µM 2-mercaptoethanol, and 10% fetal bovine serum (FBS; Gibco, 16000-044). Cells were incubated for 6 hours at 37°C in a 5% CO_2 incubator. After stimulation, cells were stained with surface markers CD3 (FITC; ThermoFisher, 11-0032-82), CD4 (PE-Cy7; ThermoFisher, 25-0042-82), and CD8 (PerCP-Cy5.5; ThermoFisher, 45-0081-82) in Flow Cytometry Staining Buffer (Invitrogen, 00-4222-26) for 30 minutes at 4°C in the dark. Cells were then washed twice and fixed/permeabilized using the Intracellular Fixation & Permeabilization Buffer (Invitrogen, 88-8824-00) for 30 minutes. Intracellular cytokine staining was performed with antibodies against IFN-γ (PE; ThermoFisher, 12-5773-82), IL-2 (APC; ThermoFisher, 17-7021-82), and TNF- (APC; ThermoFisher, 17-7321-82) for 20 minutes at room temperature in the dark. After washing, cells were resuspended in permeabilization buffer and analyzed using a flow cytometer (Agilent, Novocyte3000).

Immunogenicity analysis

Immunogenicity was assessed per peptide by one-sided paired t-test comparing the percentage of TNF-α-producing $CD8^+$ T cells in peptide-stimulated versus matched negative-control (DMSO) samples from the same mouse ($p < 0.05$). This was the single pre-specified primary readout. No correction for multiple comparisons was applied; uncorrected p-values are reported for each peptide independently, consistent with standard practice in small-cohort epitope immunogenicity studies.

Immunogenicity was read out on day 105, after tumour re-challenge at day 63 and vaccine boost at day 98 ($n = 5$ vaccinated tumour-free survivors). Because these mice were selected as tumour-free survivors, they may represent strong immune responders; however, the within-mouse paired design controls for this enrichment. Four deepMHCflare-selected peptides and one reference peptide (DYWGQGTEL) were tested. Shapiro–Wilk tests confirmed that paired differences were consistent with normality for all peptides and controls (DMSO control: $p = 0.52$; formic acid control: $p = 0.46$), supporting the use of the paired t-test:

- **AYLELSNL** - 2.3×, $p = 0.006$. Significant.
- **YWGQGTEL** - 2.3×, $p = 0.028$. Significant.
- **YYCSISGDY** - 1.4×, not significant. Warncke et al. [21] independently identified YYCSISGDY as the only genuinely tumour-specific CDR3 peptide from A20 with no normal immunoglobulin homology, and reported that it induced antigen-specific effector T cells. Despite this literature support, the peptide did not reach significance in our assay.
- **DSFLHWFL** - inadvertently reconstituted in formic acid instead of DMSO. Against the matched formic-acid negative control: 2.0×, $p = 0.170$ (not significant). Against the DMSO negative control used for the other peptides: $p = 0.053$ (borderline, but this cross-vehicle comparison was not pre-specified). We conservatively classify DSFLHWFL as a weak positive given the confounded vehicle.

Secondary readouts ($CD4^+/CD8^+ \times$ IFN-γ, IL-2, Granzyme B) are reported in Extended Data Table 6. For bias-simulation experiments, standard errors were computed across five independent replications.

Extended Data

Extended Data Table 1a: Headline classification of IEDB eluted-ligand entries (assessable subset).

Category	Count	% of Assessable
CLEAN	1,511,463	44.2%
BIASED	1,906,376	55.8%
Assessable total	**3,417,839**	**100.0%**
Multi-allelic unassigned	1,351,486	-
Insufficient metadata	1,078,637	-
Grand total	**3,971,663**	

Extended Data Table 1b: Detailed subcategory definitions and counts for each data usability classification.*

Category	Reasoning	Count	%
1. CLEAN (1,511,463 - 44.2% of assessable)			
1.1 Mono-allelic	Cell line expresses a single HLA allele. Evidence code: "Single allele present". Allele label is experimental; no inference.	572,451	14.4%
1.2 Allele-specific antibody	Immunopurification with an allele-specific antibody (e.g., BB7.2 for HLA-A2). Only peptides from the target allele are captured; label is experimentally resolved.	100,717	2.5%
1.3 Paper-review reclassified	Entries originally classified as biased or unverifiable by the programmatic SQL stage, but confirmed clean upon systematic paper-level review.	823,151	20.7%
2. BIASED (1,906,376 - 55.8% of assessable)			
2.1 Inferred by motif	Evidence code: "Inferred by motif or alleles present". Allele label assigned computationally; introduces predictor dependence.	809,580	20.4%
2.2 Deconvoluted	Multi-allelic source with pan-specific IP (e.g., W6/32) but stored as allele-specific. Requires computational deconvolution.	58,792	1.5%
2.4 Paper-review reclassified	Entries originally classified as clean or unverifiable by the programmatic SQL stage, but identified as biased upon systematic paper-level review.	1,038,004	26.1%
3. MULTI-ALLELIC UNASSIGNED (1,351,486)			
3.1 Multi-allelic	Multi-allelic source labelled only as a class (e.g., "HLA class I"). Valid peptides, but unusable for allele-specific training without deconvolution.	1,351,486	34.0%

Category	Reasoning	Count	%
4. INSUFFICIENT METADATA (1,078,637)			
4.1 Insufficient metadata	Antibody-based capture, but `h_mhc_types_present` is missing or empty. Cannot determine whether source is mono-allelic.	1,078,637	27.2%
Total		**3,971,663**	**100.0%**

*Subcategories 1.3 and 2.4 represent entries reclassified during the systematic paper-level review of 37 publications, beyond the initial programmatic SQL classification.

Antibody	Specificity	Count
Pan-HLA Class I (Requires Deconvolution)		
W6/32	Pan-HLA-I (A, B, C)	844,216
Pan-HLA Class II (Requires Deconvolution)		
Tu39	Pan-HLA-II (DR, DP, DQ)	447,539
B7/21	HLA-DP	86,438
LB3.1	HLA-DR	71,172
L243	HLA-DR	59,897
Allele-Specific (Clean Data)		
BB7.2	HLA-A2 (A*02 family)	52,688

Extended Data Table 1d: Specificity and IEDB counts for common immunopurification antibodies.

Extended Data Table 1c: Systematic review of 37 publications with classification verdicts, entry counts, and evidence quotes.

Title	PMID	Verdict	Type	Count	Quote
HLA Ligand Atlas: a benign reference of HLA-presented peptides to improve T-cell-based cancer immunotherapy.	33858848	BIASED	Mono-allelic	365,152	Peptides were categorized as strong binders against a given HLA allotype if any tool reported it as such (netMHCpan-4.1 s rank $\leq$ 0.5, netMHCIIpan-4.0 s rank $\leq$ 1.0, MixMHC2pred s rank $\leq$ 1.0).
		CLEAN	Multi-allelic	591,784	Multi-allelic patient tissue data stored at class level without predictor-based allele assignment; classified as clean upon manual review.
A large peptidome dataset improves HLA class I epitope prediction across most of the human population.	31844290	CLEAN	Mono-allelic	219,893	No single passage captures the full methodology; manual review confirms predictor-independent allele assignment.
		CLEAN	Multi-allelic	36,353	
The natural HLA ligandome of glioblastoma stem-like cells: antigen discovery for T cell-based immunotherapy.	29557506	BIASED	Multi-allelic	218,122	HLA annotation was performed using the prediction tool SYFPEITHI or an extended in-house database. Experimental validation of the predicted HLA restriction was obtained for a subset of ligands.
Immunopeptidomic analysis of influenza A virus infected human tissues identifies internal proteins as a rich source of HLA ligands.	35051231	BIASED	Mono-allelic	85	Class I HLA allotype 9-mer binding motifs derived from host immunopeptides using unsupervised clustering using MixMHCp.
		BIASED	Multi-allelic	104,486	
High-throughput and Sensitive Immunopeptidomics Platform Reveals Profound Interferonγ-Mediated Remodeling of the Human Leukocyte Antigen (HLA) Ligandome.	29242379	BIASED	Multi-allelic	97,086	Binding predictions were assigned to peptides only if they were predicted to bind to only one HLA allotype. The threshold for binding was set to rank <2% ... using NetMHC4.0.
IFNγ Modulates the Immunopeptidome of Triple Negative Breast Cancer Cells by Enhancing and Diversifying Antigen Processing and Presentation.	33968037	BIASED	Multi-allelic	94,325	Filtering on 8-12 mer peptides, we analyzed these peptides according to their HLA allotype predicted binding affinities using NetMHC pan 4.0.
Spliced Peptides and Cytokine-Driven Changes in the Immunopeptidome of Melanoma.	32938616	BIASED	Mono-allelic	76,884	Binding prediction We used the NetMHC4 (29, 30) algorithms for binding predictions for both spliced and linear peptides. A default rank cutoff of 2 was implemented as a binder peptide (Supplementary Table S2). Immunogenicity prediction We used the Immune Epitope Database and Analysis Resource (www.iedb.org), Class I Immunogenicity algorithm for immunogenicity predictions of linear peptides (31). The peptides used in functional assays indicated as "predicted immunogenic" had an immunogenicity score between 0.37 and 0.55 (Supplementary Table S3).
		BIASED	Multi-allelic	15,285	
HLA class II immunopeptidomics reveals that co-inherited HLA-allotypes within an extended haplotype can improve proteome coverage for immunosurveillance.	34357683	BIASED	Multi-allelic	84,754	pHLA-II complexes were immunoaffinity purified using SPV-L3 (anti-DQ), LB3.1 (anti-DR), B721 (anti-DP). Motif deconvolution performed using GibbsCluster.
Immunopeptidome Analysis of HLA-DPB1 Allelic Variants Reveals New Functional Hierarchies.	32350084	BIASED	Mono-allelic	81,042	HLA-DP molecules were isolated by immunoaffinity chromatography using anti-HLA-DP antibody B7/21. Gibbs clustering was used for motif identification.
Integrated proteogenomic deep sequencing and analytics accurately identify non-canonical peptides in tumor immunopeptidomes.	32157095	BIASED	Multi-allelic	75,412	To evaluate the binding affinity of HLAIp, MixMHCpred.v2 prediction software was run on all HLAIp ranging in length from 8 to 14 amino acids. Peptides with a p-value$\leq$0.05 were considered binders.

Title	PMID	Verdict	Type	Count	Quote
The self-peptide repertoire plays a critical role in transplant tolerance induction.	34428180	CLEAN	Mono-allelic	70,168	MHC complexes were isolated by immunoaffinity purification using solid phase-bound monoclonal antibodies SF1-1.1.10 (anti-H-2Kd), K9-178 (anti-H-2Kb), Y3 (anti-H-2Kb/Kk), and 28.14.8s (anti-H-2Db).
Identification of Tumor Antigens Among the HLA Peptidomes of Glioblastoma Tumors and Plasma.	31154438	BIASED	Multi-allelic	56,128	HLA annotation was performed using the prediction tool SYFPEITHI or an extended in-house database. Only unique protein IDs represented by peptides with predicted HLA binding (NetMHC 3.4 percentile rank < 2%).
CIITA-Transduced Glioblastoma Cells Uncover a Rich Repertoire of Clinically Relevant Tumor-Associated HLA-II Antigens.	33592498	BIASED	Mono-allelic	55,721	MixMHCpred (v.2.0.2) (49,50) was used to predict binding of peptides on patients' HLA class I alleles, and MixMHC2pred (v.1.0) (51) was used to predict binding of peptides on patients' HLA class II alleles (supplemental Table S5). HLA-DPA1/DPB1 and HLA-DQA1/DQB1 pairs were assigned for each patient based on their availability in MixMHC2pred. For each peptide, the HLA allele with the lowest rank was assigned as binder. Peptides with multiple binding alleles assigned were removed from the analysis. Some of the HLA class II alleles were not available for the prediction analysis, leading to a miss assignment of HLA class II alleles to some of the HLA class II peptides. Those peptides will therefore have an artificially higher binding rank. For the analysis of the distribution of HLA binding peptides, only peptides with rank $\leq$2 were selected.
Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allostery regulating HLA specificity.	28832583	CLEAN	Multi-allelic	53,689	No single passage captures the full methodology; manual review confirms predictor-independent allele assignment.
A subset of HLA-I peptides are not genomically templated: Evidence for cis- and trans-spliced peptide ligands.	30315122	CLEAN	Mono-allelic	50,222	No single passage captures the full methodology; manual review confirms predictor-independent allele assignment.
Promiscuity of Peptides Presented in HLA-DP Molecules from Different Immunogenicity Groups Is Associated With T-Cell Cross-Reactivity.	35251023	CLEAN	Mono-allelic	47,042	No single passage captures the full methodology; manual review confirms predictor-independent allele assignment.
The Peptide Repertoire of HLA-B27 may include Ligands with Lysine at P2 Anchor Position.	29393594	BIASED	Mono-allelic	42,419	Quite complicated: "The entire peptidome was subjected to binding prediction analysis using NetMHCpan 4.0. Peptides with lowest predicted Rank of less than 2 by NetMHCpan were assumed to be ligands of the endogenous alleles of each cell line." "The B1.23.2 and HB-119 hybridoma cell lines were used for the production of the B1.23.2 and ME1 mAbs, respectively, and used for the immunoaffinity purification of the HLA molecules. B1.23.2 is a high affinity mouse IgG2a mAb with specificity for HLA-B and HLA-C (also for some HLA-A allotypes), primarily associated with β2-microglobulin (β2m). The B1.23.2 mAb does not react with the rat MHC even in the presence of human β2m. The HB-119 hybridoma produces ME1, a mouse IgG1 mAb with specificity for HLA B07, B22, B27, B14 and B*46." "The cervix epithelial HeLa (HLA-A03, A68, B15:03 and C16:04) and the HLA-class I defective lymphoblastoid C1R (HLA-A02:01, B35:03 and C04:01) cell lines were previously transfected with either HLA-B27:05 or HLA-B*27:09 or their corresponding C67S mutation"
In-depth mining of the immunopeptidome of an acute myeloid leukemia cell line using complementary ligand enrichment and data acquisition strategies.	32387766	BIASED	Mono-allelic	22,063	No single passage captures the full methodology; manual review confirms reliance on predictor-based allele assignment.
		BIASED	Multi-allelic	19,295	
Immunopeptidomic Analysis Reveals That Deamidated HLA-bound Peptides Arise Predominantly from Deglycosylated Precursors.	32357974	CLEAN	Mono-allelic	38,754	pHLA complexes were immunoaffinity purified using W6/32 monoclonal antibody. MS/MS data were searched by PEAKS Studio 8.5. A 1% FDR threshold was applied.
Biogenesis of HLA Ligand Presentation in Immune Cells Upon Activation Reveals Changes in Peptide Length Preference.	32983136	BIASED	Multi-allelic	35,075	top10 method containing an inclusion list of precursors which comprised 9-12 mer peptides that were predicted by NetMHC pan4.0 (<2% rank).

Title	PMID	Verdict	Type	Count	Quote
High-sensitivity HLA class I peptidome analysis enables a precise definition of peptide motifs and the identification of peptides from cell lines and patients' sera.	26992070	CLEAN	Multi-allelic	34,797	1847 of the identified peptides were predicted to bind to B*27:04 and GibbsCluster-1.0 server analysis demonstrated excellent agreement.
A tissue-based draft map of the murine MHC class I immunopeptidome.	30084848	BIASED	Mono-allelic	34,498	"H2Db- and H2Kb-associated peptides were isolated by a conventional immunoaffinity purification method using the monoclonal antibodies B22-249.R1 and Y-3, respectively22" "o select the final list of high-confidence H2Db- and H2Kb-associated peptides, strict cut-off criteria were applied: FDR 1% (peptide-spectrum match level); 8–9 and 9–11 amino acids in length for H2Kb and H2Db peptides, respectively; and IC50<500nM (NetMHC v4.0). "
The immunopeptidomes of two transmissible cancers and their host have a common, dominant peptide motif.	33460454	CLEAN	Multi-allelic	33,959	The alpha chains of MHC-I molecules isolated during immunoprecipitation were utilized for allotype identification using mass spectrometry.
Thermostability profiling of MHC-bound peptides: a new dimension in immunopeptidomics and aid for immunotherapy design.	33298915	BIASED	Mono-allelic	33,468	All post-processed peptides from the HLA-A*02:01 and HLA-B*07:02 were compiled, and the sequences were clustered48 to identify motifs characteristic of the HLA-C*04:01 allele, which is expressed at relatively low levels, and the HLA-B*35:03 allele, expressed at residual levels, by C1R cells18. Only ligands identified in the C1R background dataset, comprising ligands in the work by Schittenhelm et al.18 and in-house identified C1R ligands, were included. HLA-B*35:03 peptides were subsequently removed from further analysis, as these represented just 73 peptides.
Carbamazepine Induces Focused T Cell Responses in Resolved Stevens-Johnson Syndrome and Toxic Epidermal Necrolysis Cases But Does Not Perturb the Immunopeptidome for T Cell Recognition.	33912179	CLEAN	Mono-allelic	32,848	To characterize the HLA-B*15:02 binding motif, peptide sequences identified at a 5% peptide FDR in control data sets representing the endogenous HLA class I and II of C1R cells were removed. So too were peptides assigned without a protein accession or mapping exclusively to proteins in the contaminant database.
The Length Distribution and Multiple Specificity of Naturally Presented HLA-I Ligands.	30429286	BIASED	Multi-allelic	32,682	Peptides were assigned to specific alleles based on the highest responsibility value observed for the corresponding motif.
Mass Spectrometry Based Immunopeptidomics Leads to Robust Predictions of Phosphorylated HLA Class I Ligands.	31848261	BIASED	Multi-allelic	32,506	No single passage captures the full methodology; manual review confirms reliance on predictor-based allele assignment.
Solving an MHC allele-specific bias in the reported immunopeptidome.	32897882	BIASED	Mono-allelic	29,154	To assign peptides that passed the MS quality filters described above to their HLA complexes that they most likely bind to, we used the netMHCpan 4.0 algorithm (24) with default settings. No binding affinity predictions were enabled. Therefore, all peptides with affinity percentage ranks below 2 were considered binders.
		BIASED	Multi-allelic	154	
Unveiling the Peptide Motifs of HLA-C and HLA-G from Naturally Presented Peptides and Generation of Binding Prediction Matrices.	28904123	BIASED	Mono-allelic	27,795	ue to endogenous expression of HLA-B*35:03 and HLA-C*04:01 in C1R cells, isolated HLA ligands of these allotypes had to be excluded from further analysis to allow for identification of HLA ligands of the transfected allele. GibbsCluster 1.1 (39) is an unsupervised way to cluster peptides according to their sequence similarity.
Mass Spectrometry Profiling of HLA-Associated Peptidomes in Mono-allelic Cells Enables More Accurate Epitope Prediction.	28228285	CLEAN	Mono-allelic	27,102	No single passage captures the full methodology; manual review confirms predictor-independent allele assignment.
Accurate MHC Motif Deconvolution of Immunopeptidomics Data Reveals a Significant Contribution of DRB3, 4 and 5 to the Total DR Immunopeptidome.	35154160	BIASED	Multi-allelic	27,015	GibbsCluster deconvolution for MHC class I artificial dataset was performed using default parameters as described earlier (8), and including 1-6 clusters, motif length 9 amino acids, using "trash cluster" with a threshold=2, performing single sequence moves at every iteration, max length deletions = 5, max length insertions=1, number of seeds for initial starting conditions=5, Number of iterations x sequence x temperature step = 50.

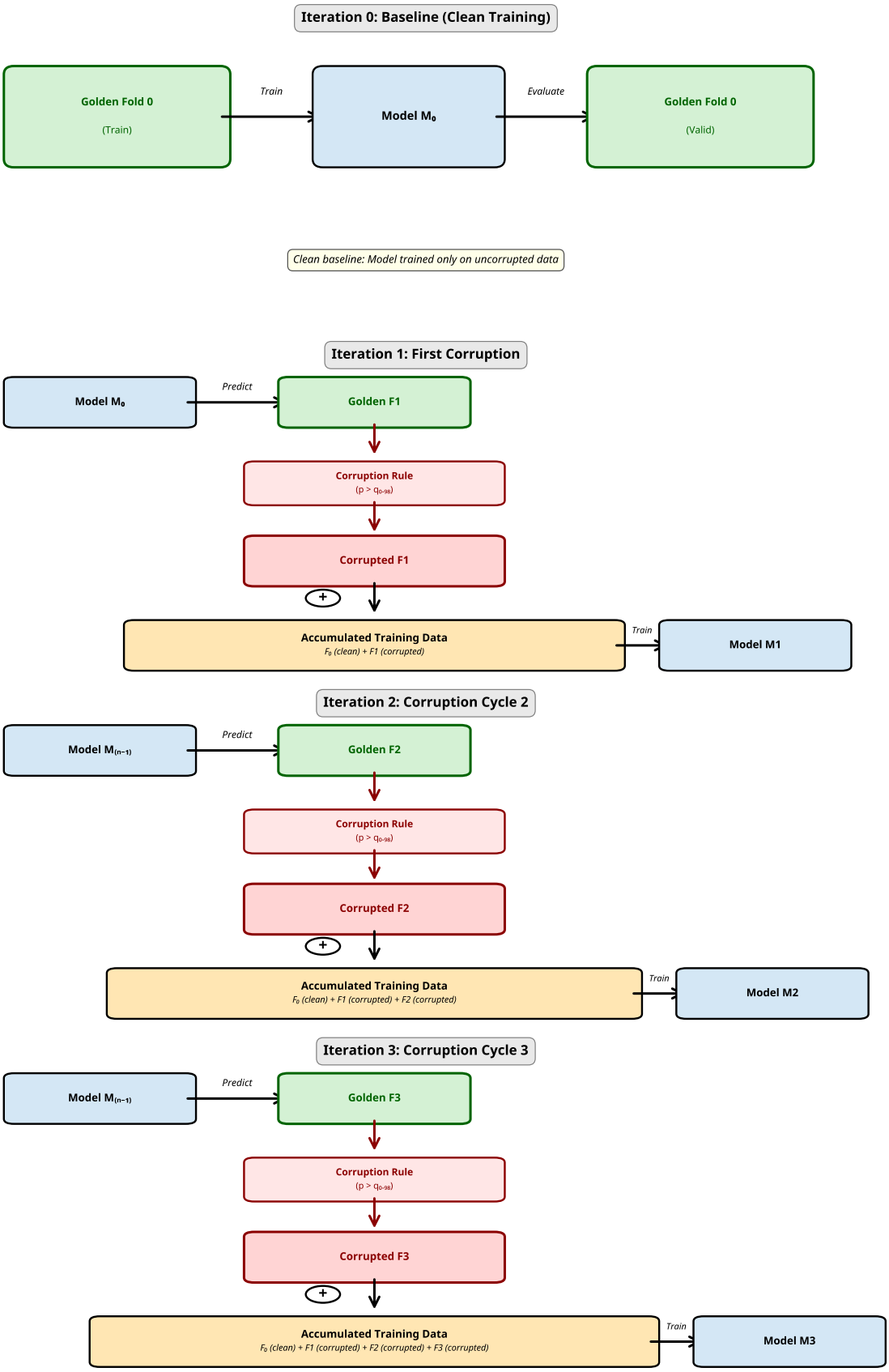
Title	PMID	Verdict	Type	Count	Quote
An Integrated Genomic, Proteomic, and Immunopeptidomic Approach to Discover Treatment-Induced Neoantigens.	33936100	BIASED	Multi-allelic	24,319	Immunopeptidome datasets were evaluated with the PLAtEAU script (44) to identify core binding epitopes. Only peptides reported in all three biological replicates (with the criteria of at least observed in one of the three technical replicate injections) were considered for further PLAtEAU analysis. The minimum epitope length was set to 9 for the MHC-I and both 9 and 13 for the MHC-II. The default option of impute with lowest measured value in each run was enabled. Both the entire set of peptides from the ligandome analysis and the subsequently defined core epitope output from PLAtEAU were used as input for affinity prediction using NetMHCpan 4.1 (45) for MHC-I data and NetMHCIIpan 4.0 (10) for the MHC-II data set. The predicted rank threshold for strong binding peptides were 0.500 or a rank threshold for weak binding peptides of 2.000 for MHC-I and a rank threshold of 10 for MHC-II were included for analysis. The highest affinity allele across the six GRANTA-519 alleles which met the above criteria was considered for further analysis.
Immunopeptidomic Profiling of HLA-A2-Positive Triple Negative Breast Cancer Identifies Potential Immunotherapy Target Antigens.	29786170	BIASED	Mono-allelic	269	Peptides with a length between 8 and 14 aa had attributes of HLA-A*0201 binding with a NetMHCpan rank score equal or smaller than 2.
		BIASED	Multi-allelic	23,749	
Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation.	25576301	BIASED	Mono-allelic	79	No single passage captures the full methodology; manual review confirms reliance on predictor-based allele assignment.
		BIASED	Multi-allelic	22,919	
Exploration of peptides bound to MHC class I molecules in melanoma.	25645385	BIASED	Mono-allelic	69	The HLA-binding of each peptide was predicted using the ANN algorithm in IEDB and a binding affinity cut-off of 500nM.
		BIASED	Multi-allelic	12,623	
HLA-B and cysteinylated ligands distinguish the antigen presentation landscape of extracellular vesicles.	34211107	BIASED	Mono-allelic	3,246	Only the NetMHC 4.0 calculated 9-mer cores were used for these plots.
		BIASED	Multi-allelic	651	
Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry.	27869121	BIASED	Mono-allelic	160	Algorithms used for prediction of peptide ligands. Affinity to the corresponding allotypes predicted using NetMHC 4.0.
		BIASED	Multi-allelic	616	

Extended Data Table 2: Allele-to-fold assignment for the in silico bias simulation. The partition is fixed across all five replications ($n = 3 \ldots 7$). Each fold contains 23 alleles.

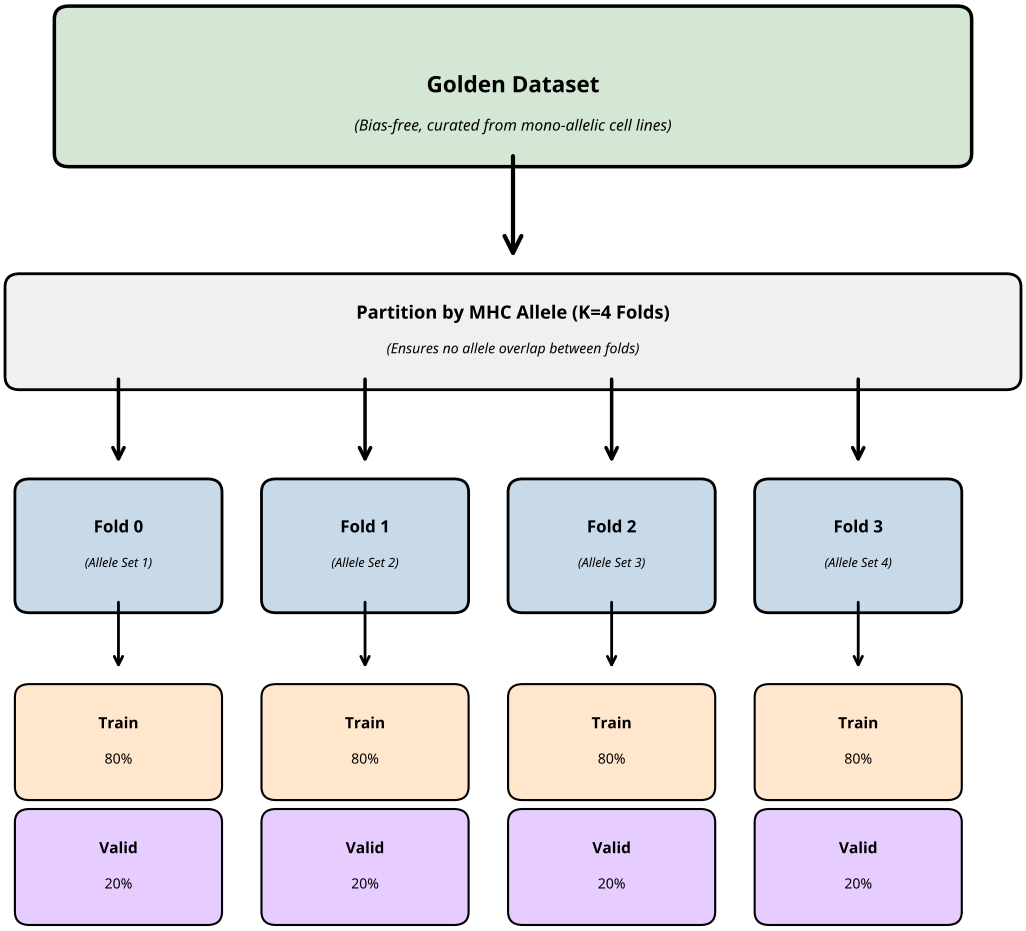
Fold	Alleles
0	A*02:04, A*30:01, A*30:02, A*33:01, A*33:03, B*08:01, B*18:01, B*35:01, B*38:01, B*40:06, B*42:01, B*45:01, B*49:01, B*50:01, B*52:01, B*58:02, C*02:02, C*03:03, C*04:01, C*04:03, C*05:01, C*07:04, C*08:01
1	A*02:02, A*02:03, A*02:07, A*02:11, A*11:01, A*25:01, A*29:02, A*34:02, A*36:01, A*66:01, A*68:02, B*13:01, B*15:01, B*15:02, B*15:03, B*27:05, B*38:02, B*40:01, B*46:01, B*53:01, B*58:01, C*14:02, G*01:04
2	A*02:01, A*02:05, A*03:01, A*11:02, A*23:01, A*24:02, A*24:07, A*26:01, B*07:02, B*13:02, B*15:10, B*15:17, B*40:02, B*44:02, B*44:03, B*51:01, B*55:02, B*56:01, B*57:03, C*03:02, C*07:02, C*08:02, C*17:01
3	A*01:01, A*02:06, A*31:01, A*32:01, A*34:01, A*68:01, A*74:01, B*07:04, B*14:02, B*35:03, B*35:07, B*37:01, B*54:01, B*55:01, B*57:01, C*01:02, C*12:02, C*12:03, C*14:03, C*15:02, C*16:01, G*01:01, G*01:03

Extended Data Table 3: Hyperparameters for the final deepMHCflare model.

Parameter	Value
Backbone	ESM2-t6-8M (6 layers, 8M parameters)
MHC representation	Alpha-1 + alpha-2 pseudo-sequence (∼182 AA)
Input encoding	Concatenation: MHC pseudo-seq + peptide
Max sequence length	1536 tokens
Pooling	mean_sqrt + max + CLS (concatenated)
Head	Linear, 2 outputs, sigmoid
Loss	90% LambdaRank NDCG@5 + 10% weighted CE
Optimizer	LAMB, lr = 0.003, weight_decay = 0.2
Effective batch size	1024 × 64 grad. accum. = 65,536
Precision	bf16-mixed
Strategy	DDP
Epochs	100 (validation every 5)
Hidden dropout	0.2
Attention dropout	0.0
Sampler	ContextualSampler (3000 ctx/MHC, 16 ctx/batch)
Seed	42



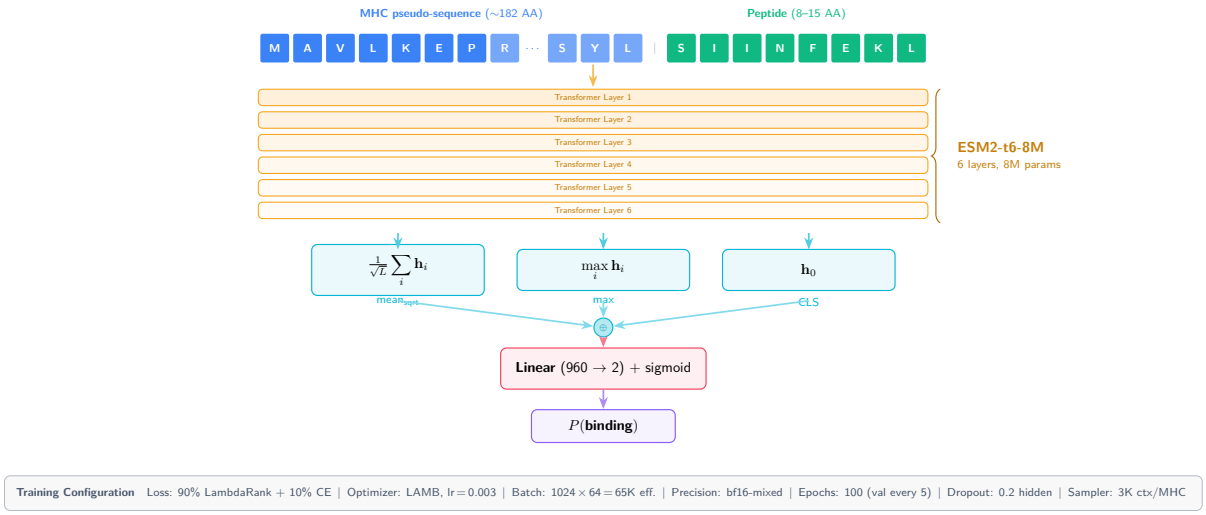
Extended Data Figure 1: Iterative simulation of systematic confirmation bias. The four-iteration process used to simulate bias accumulation. In each iteration $i > 0$, the model from the previous step (M_{i-1}) predicts on and subsequently corrupts the next clean fold, which is then added to the growing training dataset.



Extended Data Figure 2: Predictor-independent dataset partitioning strategy. The predictor-independent dataset was partitioned into four non-overlapping folds defined by MHC allele. Within each fold, data were further split into 80% training and 20% validation sets.

Extended Data Table 4: Design dimensions explored during model development.

Dimension	Range explored	Final choice
Architecture	Single encoder vs. dual (separate MHC/peptide) encoders, incl. funnel variant	Single encoder (concat)
Backbone	ESM2 variants (8M–650M)	ESM2-t6-8M
Loss function	Pure CE, LambdaRank (Precision@k, NDCG@k for k=5,10), combinations at ratios 0.1–0.9	90% LambdaRank NDCG@5 + 10% weighted CE
Learning rate	10^{-6} to 10^{-3}	3×10^{-3}
Batch size	64–1024, grad. accum. 1–128	$1024 \times 64 = 65$K effective
Regularisation	Hidden dropout 0.0–0.2, weight decay 0.05–0.3, layer freezing 0–5 layers	dropout = 0.2, weight_decay = 0.2, no freezing
Sampling	Contextual (50–3000 ctx/MHC), parent-protein, mixed-batch	Contextual, 3000 ctx/MHC, 16 per batch
Scheduler	None, CosineAnnealing, custom TwoLRs	None (static LR)
Datasets	7+ dataset generations, interspecies, golden subsets	Predictor-independent mono-allelic + curated multi-allelic



Extended Data Figure 3: deepMHCflare architecture. The MHC alpha-1/alpha-2 pseudo-sequence (∼182 AA) is concatenated with the candidate peptide (8–15 AA) and processed by a pretrained ESM2-t6-8M encoder (6 transformer layers, 8M parameters). Three pooling strategies - mean with square-root normalisation, max, and CLS token - produce a 960-dimensional representation passed through a linear head to yield a binding probability.

Extended Data Table 5: Summary of datasets used for the final deepMHCflare model.

Dataset	Role	Observations	Positives	Pos. rate	Alleles
Mono- & multi-allelic training pool	Train	167.0M	370,879	0.22%	72
In-distribution validation (mono-allelic held-out)	Val (ID)	4.0M	6,259	0.16%	23
Out-of-distribution validation (21 unseen alleles)	Val (OOD)	5.4M	6,539	0.12%	21
HLA Ligand Atlas (multi-allelic patient tissues)	Val (multi)	282.5M	388,262	0.14%	51

Extended Data Table 5b: Allele composition of each dataset split. Each row lists the alleles for one locus within one split.

Split	Locus	n	Alleles
Train (72)	H-2	2	D^b, K^b
	HLA-A	18	02:01, 02:03, 02:04, 02:05, 02:07, 03:01, 03:02, 03:07, 11:01, 11:10, 24:02, 24:06, 24:13, 25:01, 29:02, 29:13, 32:01, 32:02
	HLA-B	36	07:02, 15:01, 15:02, 15:11, 15:15, 15:18, 18:03, 27:01, 27:02, 27:03, 27:04, 27:06, 27:07, 27:08, 27:09, 27:10, 27:16, 39:01, 39:05, 39:06, 39:09, 39:24, 41:01, 41:02, 41:03, 41:04, 41:05, 41:06, 44:05, 44:08, 44:09, 44:27, 44:28, 47:01, 51:08, 73:01
	HLA-C	9	02:02, 03:03, 03:13, 05:01, 07:02, 07:04, 07:06, 12:02, 12:04
	HLA-E	1	01:03
	HLA-F	3	01:01, 01:03, 01:04
	HLA-G	3	01:01, 01:03, 01:04
Val ID (23)	HLA-A	11	02:01, 02:03, 02:04, 02:05, 02:07, 03:01, 11:01, 24:02, 25:01, 29:02, 32:01
	HLA-B	3	07:02, 15:01, 15:02
	HLA-C	6	02:02, 03:03, 05:01, 07:02, 07:04, 12:02
	HLA-G	3	01:01, 01:03, 01:04
Val OOD (21)	HLA-A	6	02:02, 02:06, 02:11, 11:02, 23:01, 24:07
	HLA-B	9	07:04, 15:03, 15:10, 15:17, 18:01, 27:05, 44:02, 44:03, 51:01
	HLA-C	6	03:02, 08:01, 08:02, 12:03, 15:02, 17:01
Val multi (51)	HLA-A	16	01:01, 02:01, 02:05, 03:01, 11:01, 23:01, 24:02, 25:01, 26:01, 26:08, 29:02, 30:01, 32:01, 68:01, 68:02, 69:01
	HLA-B	21	07:02, 07:06, 08:01, 13:02, 14:02, 15:01, 18:01, 27:05, 35:01, 35:03, 35:08, 37:01, 39:01, 40:02, 41:01, 44:02, 44:03, 45:01, 49:01, 50:01, 58:01
	HLA-C	14	02:02, 03:03, 03:04, 04:01, 05:01, 06:02, 07:01, 07:02, 07:04, 08:02, 12:03, 15:05, 16:01, 17:01

Extended Data Table 6: Peptide sequences tested in the prospective validation study. Four peptides were selected by deepMHCflare; DYWGQGTEL was included as a reference control (top-ranked NetMHCpan 4.1 prediction). Immunogenicity was defined as a statistically significant increase in $CD8^+$ TNF-α^+ T-cells (one-sided paired t-test, $p < 0.05$).

Sequence	**BCR Region**	**Prediction Tool**	**Immunogenic**
YYCSISGDY	VH_CDR3	deepMHCflare	No[‡]
AYLELSNL	VH_FR3	deepMHCflare	Yes
DSFLHWFL	VL_FR2	deepMHCflare	Weak[*]
YWGQGTEL	VH_CDR3, FR4	deepMHCflare	Yes
DYWGQGTEL[†]	VH_CDR3, FR4	NetMHCpan 4.1	No

[*]Inadvertently reconstituted in formic acid instead of DMSO; immunogenicity was assessed against a matched formic acid vehicle control (2.0×, not significant). Classified as weak positive due to the confounded vehicle.
[†]Reference control; top-ranked NetMHCpan 4.1 prediction. Warncke et al. [21] showed that this peptide induces suppressive $CD4^+$ rather than cytotoxic $CD8^+$ T-cell responses.
[‡]Independently identified by Warncke et al. [21] as the only genuinely tumour-specific CDR3 peptide from A20 that induced antigen-specific effector T cells.