A021 · Advancing Peptide-HLA Class I Prediction with Active Learning Frameworks for Improved Cancer Vaccine Design

Presenter: Jessika Baral — Harvard Medical School, Broad Institute, MGH

Authors: Jessika Baral, Luis Correa-Medero, Marta Wilbrink, Cleo Forman, Timothy Zhu, Kasidet Manakongtreecheep, Emma C. Duggan, Carl R. Klauser, Sisi Sarkizova, Matthew Bakalar, Steven A. Carr, Luca Pinello, Jennifer G. Abelin, Wengong Jin, Catherine Wu, Nir Hacohen

Session: Day 2 · Fri Oct 2, 2026 · 10:00–10:45 AM

Despite the success of immune checkpoint blockade therapies, many patients relapse due to insufficient repertoires of tumor-reactive T cells, highlighting the need for immunogenic targets. Cancer vaccines offer a promising personalized approach for treatment by targeting neoantigens, tumor-specific peptides presented by patient-specific human leukocyte antigen (HLA) complexes. Existing peptide-HLA (pHLA) prediction algorithms have enabled cancer vaccine development, but recent studies show that most peptides failed to elicit effective T-cell responses, underscoring the need for more accurate binding and immunogenicity predictors. Current approaches remain constrained by predictive power, incomplete peptide representation, and noisy labels of negatives. To address these challenges, we developed HLAGaia, a deep learning pHLA binding predictor trained on over 16 million peptides with a 50:1 nonbinder:binder ratio. HLAGaia outperformed five state-of-the-art classifiers, including BigMHC and NetMHCPan, on held-out test data with median cumulative PPV (mPPV) of 0.88. We then used an active learning framework to iteratively prioritize pHLA pairs with the highest predictive uncertainty for experimental testing using a high-throughput E. coli-based binding assay developed in our lab, HLAPlex. This enabled efficient exploration and ground-truth labeling of a combinatorially large interaction space. Retraining with relabeled peptides improved detection of allele-specific binding motifs and increased mPPV by 12.4% on a held-out test dataset. Through integrating model training with large-scale experimental validation, we provide a more effective tool for neoantigen selection and a framework for how AI-driven approaches can improve precision medicine through continuous, data-guided algorithm development.

Advancing Peptide-HLA Class I Prediction with Active Learning Frameworks for Improved Cancer Vaccine Design

Antigen-targeted therapies, such as cancer vaccines, offer a promising strategy to direct tumor-specific T cells against neoantigens, tumor-specific proteins derived from somatic mutations. Neoantigens are processed, cleaved into peptides, and presented on the surface of malignant cells and antigen-presenting cells as part of peptide-HLA class I (pHLA) complexes. T cell receptors can recognize these complexes and cause T cell expansion, leading to an anti-tumor host response. Given the vast space of possible neoantigen-HLA pairs, only a small fraction of which is actually presented on cells, there is a pressing need to (1) accurately identify antigenic targets for existing and new immunotherapies and (2) develop highly specific antigen-targeting therapies. Over the past decade, the Hacohen lab and others developed neoantigen-targeting personalized cancer vaccines (NeoVax clinical trials). Each patient received a computationally designed vaccine of 20 tumor peptide antigens selected based on predicted presentation by patient-specific HLA-I alleles. Yet only about 16% of designed peptides induced a CD8+ T cell response, underscoring the need for more robust models of both peptide presentation and downstream T cell response.

To address this, we built **HLAGaia**, a deep learning pHLA binding model that learns binding motif-specific embeddings and outperforms existing state-of-the-art classifiers. We then evaluated two complementary uncertainty-quantification strategies to guide active learning and improve HLAGaia iteratively: (1) intermodel uncertainty across existing classifiers, and (2) intramodel uncertainty, identifying pHLA pairs with the highest entropy in HLAGaia prediction. All pHLA pairs were then experimentally validated using **HLAPlex**, a novel high-throughput *E. coli*-based pHLA production system developed in the Hacohen lab. A single library can screen ~10,000 peptides and achieves 85–90% identification of positive controls. To ensure robustness, all library peptides were additionally screened for reliable *E. coli* production and mass spectrometry detectability; peptides failing either criterion were excluded from analysis.

HLA-A*11:01 is a high-value allele for this effort, carried by 50-60% of some indigenous populations. Of the experimental peptides tested for binding to HLA-A*11:01 across multiple acquisition strategies and experimental rounds, 8,256 were identified as novel binders, including peptides of motifs never previously reported for this allele. Retraining HLAGaia with a subset of all the newly labeled data improved mPPV on held-out test data by 12.4%. The retrained model also offered interpretability; counterfactual mutagenesis provided mechanistic evidence: across HLA-A*11:01 binders, mutating P2 or the peptide C-terminus produced the largest drops in predicted binding probability. A C-terminus mutation caused a mean 79.2% loss and strongly favored K/R, consistent with previous pHLA complex stability data as well.

This work also addresses a fundamental limitation of pHLA modeling: negative labels are typically assigned to peptides that have not been observed as binders, rather than those that have been experimentally confirmed to be nonbinders. However, exhaustively testing all pHLA pairs is prohibitively expensive and time-consuming. This work establishes a scalable, uncertainty-guided framework for computationally identifying the highest-value pHLA pairs to validate, then pairs that with HLAPlex's high-throughput system to generate ground-truth labels.

Significance. This approach offers a generalizable template for efficiently closing the loop between computational prediction and experimental validation in vaccine design. By targeting model uncertainty rather than randomly or exhaustively sampling peptide space, active learning yields disproportionately informative training data per experiment performed, thus improving mPPV, while reducing experimental cost and time. The advantages extend beyond performance: we identified previously unknown HLA-A*11:01-binding motifs and highlighted biochemical reasoning with model interpretability. Together, these results demonstrate that iterative learning frameworks can simultaneously improve predictive accuracy and deepen biological understanding of pHLA presentation, advancing the broader goal of designing more effective, precisely targeted cancer vaccines.