Beyond natural amino acids: extending immunogenicity risk assessment to non-canonical peptide drugs through chemical feature encoding

Maria Cairoli[1], Catherine Betts[2], Morten Nielsen[3], Olga Obrezanova[4], **Leonardo De Maria**[1]

1 Hit Discovery, Discovery Computation & AI, Discovery Sciences, Biopharmaceuticals R&D, AstraZeneca, Gothenburg, Sweden
2 Immune Safety, Clinical Pharmacology and Safety Sciences, Biopharmaceuticals R&D, AstraZeneca, Cambridge, UK
3 Department of Health Technology, Technical University of Denmark, Lyngby, Denmark
4 Biologics Engineering, Oncology R&D, AstraZeneca, Cambridge UK

https://doi.org/10.17952/38EPS.2026.P257

Abstract

Peptide therapeutics increasingly incorporate non-natural amino acids (NNAAs) to improve their properties, yet current in silico MHC class II binding predictors, central to immunogenicity risk assessment, do not account for their chemical diversity. We present a machine learning framework combining chemical fingerprints with sequence information to predict MHC-II binding for peptides containing both natural (NAA) and non-natural amino acids. Evaluating 31 molecular representations, we show that peptide-level fingerprints fail due to loss of positional information, while residue-level fingerprints (direct-encoding and similarity-based) match BLOSUM62 and one-hot encodings and correctly recover binding cores and canonical motifs. In a citrullination case study, we observed similar linear correlation across encoding strategies, with residue-level fingerprints showing marginal improvements in quantitative accuracy. This is consistent with citrulline rarely occupying anchor positions. The framework is compatible with pan-allele architectures (NNAlign / NetMHCIIpan) and extensible to diverse synthetic modifications, supporting immunogenicity risk assessment for emerging peptide therapeutics.

Introduction

Therapeutic peptides incorporate NNAAs to improve half-life, permeability, and stability. Yet the impact of NNAAs on immunogenicity, driven largely by peptide–MHC class II (pMHC-II) binding and downstream T-cell activation, remains poorly characterized. Current *in silico* pMHC-II predictors are trained almost exclusively on the 20 natural amino acids. Existing NNAA extensions use sparse encodings or neutral placeholders (e.g., 'X'), which cannot represent the chemical diversity of synthetic residues used in modern peptide design.

Gap: No framework simultaneously (i) encodes chemically diverse NNAAs, (ii) preserves the positional information needed for binding core identification, and (iii) is compatible with existing pan-allele architectures.

Objective

Evaluate if chemical fingerprints, combined with sequence positional information, provide effective molecular representations for MHC-II binding prediction across peptides containing both natural and non-natural amino acids.

Methods

Data

- MHC-II binding affinity (IEDB): 14 HLA-DR alleles
- Citrullinated peptides: 225 (DRB1*01:01) and 283 (DRB1*04:01).
- MHC-I benchmark: 5 HLA-A/B alleles, 9-mers.
- Human proteome sampling (UNIPROT) for motif validation.
- Hobohm clustering to minimize train/test redundancy.

Molecular Representations (31)

- **Sequence-based:** BLOSUM62, one-hot, NNAAIndex.
- **Peptide-level fingerprints:** ECFP4/16, FCFP, RDKit, AVALON, TT, AP, MACCS, MAPC, PepFuNN, AmorProt.
- **Direct-encoding residue fingerprints (F):** per-residue chemical fingerprint matrices.
- **Similarity-based residue fingerprints (S):** pairwise Tanimoto matrices, BLOSUM-like.

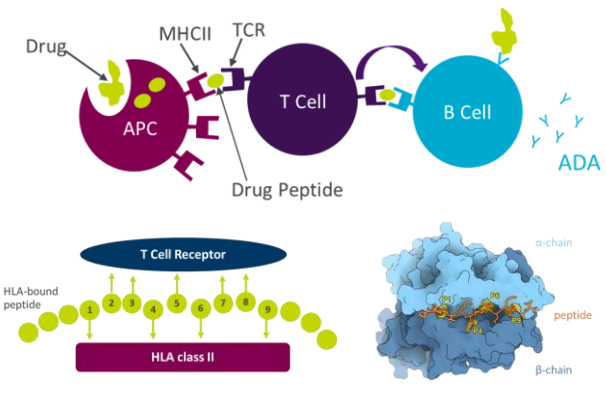
Model

NNAlign-based MLP, with nested 5-fold CV, Optuna hyperparameter tuning, 15-epoch hydrophobic-anchor burn-in and ensembling.

Evaluation

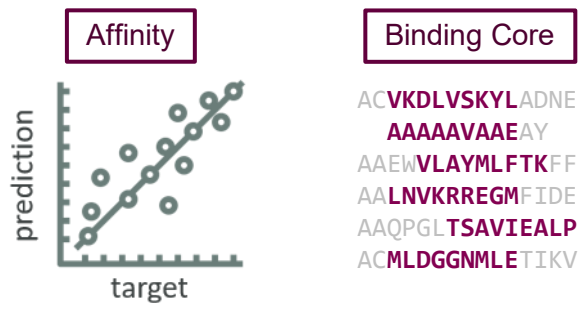
- Out-of-fold PCC, R^2, RMSE;
- paired Wilcoxon vs. BLOSUM62 with Benjamini-Hochberg FDR correction;
- bootstrap 95% CIs;
- binding-core validation against 15 HLA-DR peptides with PDB-derived cores;
- sequence logos via Seq2Logo.

The Immune Response



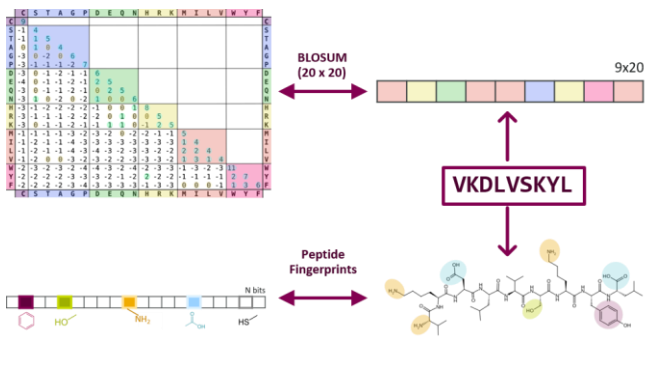
Schematics of Anti-Drug Antibodies (ADAs) generation. A stretch of nine amino acids drives peptide-MHC II presentation that can lead to helper T cells activation, generating ADAs.

In-silico predictions

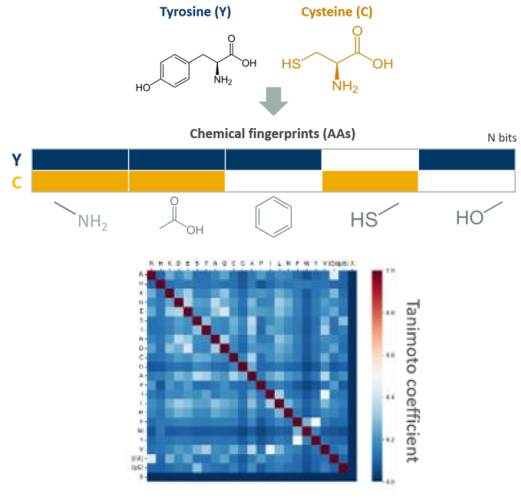


MHC II binding prediction methods should be able to predict *both* the peptide binding affinity and its binding core.

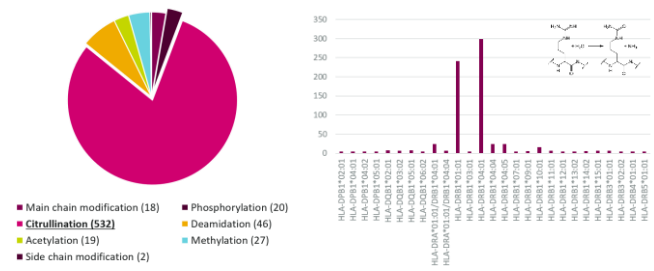
Encodings for Model Building



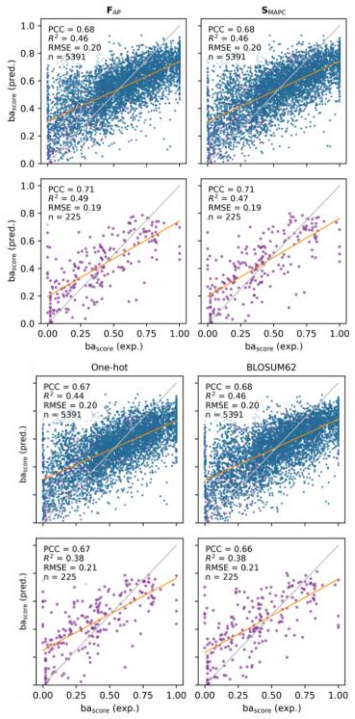
Sequence-based encodings capture AA-level peptide information using substitution scores such as BLOSUM but are limited mainly to the 20 natural amino acids and a few non-natural amino acids. SMILES-based encodings add chemical, structural, and compositional information, but initially lose positional information. This can be fixed with residue-level fingerprints and similarity (Tanimoto).



Results: citrullination case study



Citrullination provided the largest dataset for model building as per data availability in IEDB.



Prediction performance Binding affinity predictions for DRB1*01:01 and using models trained using both NAA-based (blue) and citrullinated (purple) peptides. In each subplot, the first row shows out-of-fold predictions on both NAA-based and citrullinated peptides, while the second row shows predictions on citrullinated peptides only. Blue dots: NAA-based peptides; purple dots: citrullinated peptides. Yellow line: fitted regression; gray line: identity line.

Conclusions

Novel framework: A machine learning approach combining chemical fingerprints with sequence encodings for MHCII binding prediction, where residue-level fingerprints match traditional encodings (BLOSUM62, one-hot) and accurately identify binding cores.

Promising for NNAAs: Chemical encoding offered only marginal improvements for citrullinated peptides but is expected to be more valuable for non-natural amino acids without natural analogues, though data scarcity limits full evaluation.

Extensible and application-ready: Compatible with pan-allele architectures, this framework could support early-stage immunogenicity risk assessment for therapeutic peptides, vaccines, and antibody-drug conjugates as more NNAAs data becomes available.

References

Cellular Immunology. 2015. 295: 118-126
Frontiers in Immunology. 2023. 14: 1285899
Annu. Rev. Biomed. Data Sci. 2020. 3:191–215

Acknowledgements

Maria Cairoli is supported by the AstraZeneca Postdoc Programme. The authors thank W. Czechtizky, C. Tyrchan, E. Nittinger, G. Geylan, and A. Tibo for helpful discussions.
Data & code: Training data available at *Zenodo: 10.5281/zenodo.19692117* upon publication.