ImmunoGeNN: Accelerating Early Immunogenicity Assessment for Generative Design of Biologics

Magnus Haraldson Høie[1] **Birkir Reynisson**[1] **Paolo Marcatili**[1]

Jesper Ferkinghoff-Borg[1] **Peter Clark**[1] **Kasper Lamberth**[2] **Katharina L. Kopp**[2]

Morten Nielsen[4] **Vanessa Jurtz**[1]

[1] **Novo Nordisk A/S, AI & Digital Innovation, Denmark**
[2] Novo Nordisk A/S, Global Research, Denmark
[3] Novo Nordisk A/S, Therapeutics Discovery, Denmark
[4] Technical University of Denmark, Department of Health Technology, Denmark

Abstract

A critical step in the design of biological drugs is assessment of T cell immunogenicity risk, which may cause loss of therapeutic effect and costly, late-stage clinical trial failures. Computational tools such as NetMHCIIpan are widely used for assessing risk, but are computationally expensive, limiting common drug design tasks like screening large variant libraries and exhaustively screening for deimmunizing variants.

Here, we present ImmunoGeNN, a blazingly fast, distilled neural network predicting peptide-DRB1 risk scores in the North American population at over 300,000 times the rate of NetMHCIIpan, while maintaining a >95% Spearman correlation in risk scores. ImmunoGeNN rapidly identifies dominant peptide binding cores and checks their presence in the human proteome, sharing a >99% agreement with NetMHCIIpan. Furthermore, it maintains NetMHCIIpan's performance in experimental validation against MHC-associated peptide proteomics (MAPPs) of a failed clinical phase II drug, vatreptacog alfa, successfully identifying newly introduced epitopes relative to its endogenous human protein counterpart.

With this speed-up ImmunoGeNN enables screening across millions of designs in a reasonable timeframe, removing a potential bottleneck in de novo design of biologics.

Availability: ImmunoGeNN's source code is freely available under the MIT license with no commercial restrictions. A web-server and downloadable package is made freely available for ease of use, at DTU Healthtech (`https://services.healthtech.dtu.dk/services/ImmunoGeNN`) and BioLib (`https://biolib.com/DTU/ImmunoGeNN`)

1 Motivation

Immune reactions to biological drugs are a major concern in drug development, potentially causing therapeutic failures and late-stage clinical trial setbacks Lamberth et al. [2017]. A critical step in adverse immune reactions occurs when drug peptide fragments are processed and presented by Major Histocompability Complex II (MHC-II) molecules to T cells. These MHC-II presented peptides, typically 13-17 amino acids with a 9-amino acid binding core Chang et al. [2006], can trigger T cell

receptor recognition and lead to anti-drug antibody (ADA) formation in 1-60% of patients Howard et al. [2025], Gehin et al. [2022].

Peptide-MHC-II binding serves as a strong predictor of immunogenicity Karle [2020], Jankowski et al. [2019], Jawa et al. [2020], with in-silico tools like NetMHCIIpan Reynisson et al. [2020] widely used for early risk assessment. Immunogenicity risk scores are calculated by combining MHC-II binding predictions with population allele frequencies Yip et al. [2015], Gonzalez-Galarza et al. [2019].

However, existing tools are computationally expensive at scale. Screening millions of compounds can take days to weeks, making tasks like variant library screening or deimmunizing mutation testing prohibitively costly. Testing all possible double mutations for a single candidate may involve over 100,000 combinations, creating bottlenecks in generative AI pipelines and constraining comprehensive risk assessment.

To address this challenge, we developed ImmunoGeNN - a model trained via distillation Moslemi et al. [2024] on 75,000 peptides using NetMHCIIpan-4.3 data Nilsson et al. [2023]. ImmunoGeNN predicts DRB1 immunogenicity risk scores for North American populations with 3-5 orders of magnitude speed improvement while maintaining >95% Spearman correlation with original algorithms. This enables comprehensive variant space scanning in under 1 minute, making it applicable across biological therapeutic development projects.

2 Results

For proof of concept, we focused on DRB1 gene class MHC-II alleles and their population frequencies in North America, resulting in 97 alleles with relative frequency data. We used NetMHCIIpan-4.3 training, validation and test sets for model development Nilsson et al. [2023].

Given that MHC-II peptides are typically 13-17 residues (most commonly 15) Chang et al. [2006], we standardized on 15-mer peptides as input. We randomly sampled 75,000 15-mers from the NetMHCIIpan-4.3 training set (partitions 0-3), reserved 4,500 peptides from partition 4 for validation, and extracted all unique 15-mers from the evaluation set as our test set.

We calculated each peptide's immunogenicity risk score (IRS) using an internal algorithm based on the sum of its NetMHCIIpan-4.3 peptide-allele eluted ligand rank scores, weighted by each allele's relative frequency in the North American population (see Methods). For ease of interpretability, we then rank-normalized the IRS scores with a Scikit-learn QuantileTransformer Pedregosa et al. [2011], fitted on the training set IRS scores. We term these IRS percentile scores, where an IRS of 99% means a peptide in the top 1% of risk scores, and an IRS of 50% means a peptide with median risk (for details refer to Methods). While the NetMHCIIpan-4.3 dataset additionally contains labels for positives (~17% of peptides, binding to any MHC-II allele) and negatives (~83% of peptides), we only considered the regression task of predicting the IRS score assigned to each peptide in this work.

Since NetMHCIIpan computes scores for all 97 alleles per peptide and processes MHC pseudo sequences (~4x input expansion), we reasoned a distilled model predicting final risk scores directly should be at least 2 orders of magnitude faster.

ImmunoGeNN used a multi-layer perceptron (MLP) regressor Pedregosa et al. [2011] with one-hot encoded peptide sequences. To recognize peptide-allele binding cores, we calculated each peptide's dominant 9-mer binding core and aligned peptides so binding cores occupied consistent positions, padding with "X" for missing residues. At test time, we test all possible alignments and select the highest scoring one. For an overview of relative binding core contribution and ImmunoGeNN's DRB1 preference LOGO-plot, see figures S4 and S5 respectively).

Table 1: Model ablations effect on validation set performance. Performance on the left-out validation set, using the default Scikit learn MLPRegressor (baseline), optimized MLPRegressor with 2 hidden layers, and addition of peptide binding core alignment (see methods). PR-AUC are calculated using peptide IRS scores above the 95th or 99th percentile as positives, and below as negatives respectively, inspired by NetMHCIIpan's thresholds of 5% or 1% for classifying weak or strong peptide-MHC binders.

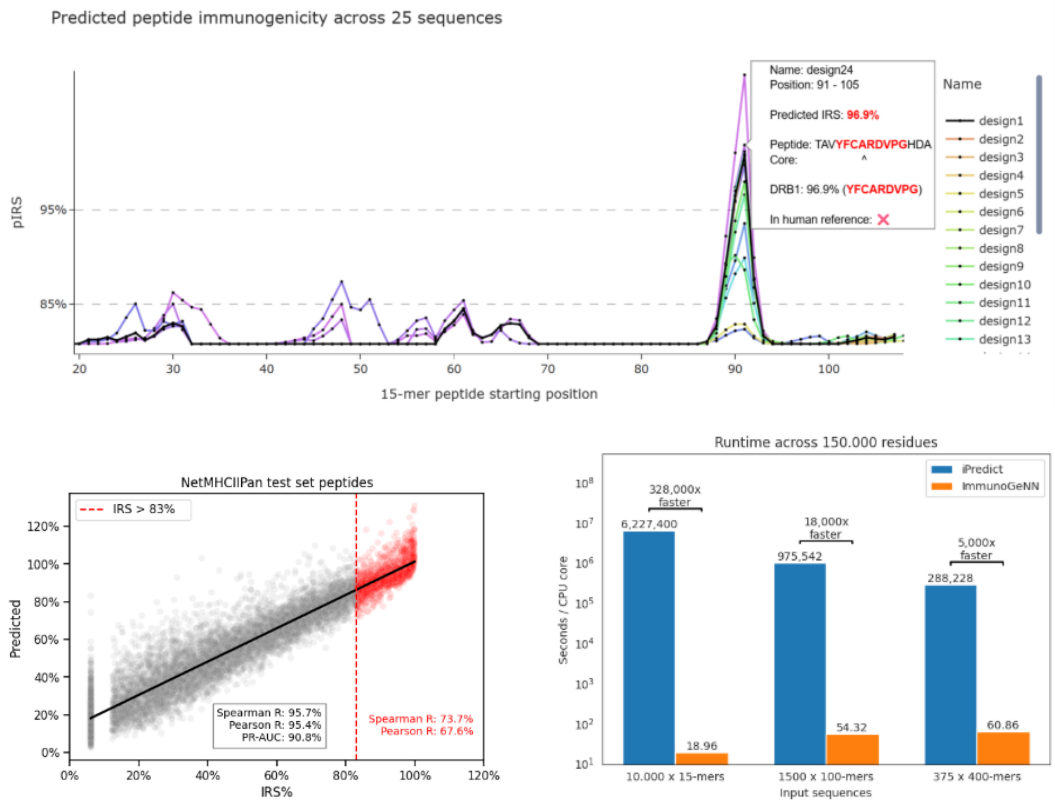

Figure 1: **ImmunoGeNN predicts peptide early immunogenicity risk scores for input protein sequences in FASTA format.** A) Predicted peptide immunogenicity risk scores (pIRS) across 25 antibody sequences from an internal design campaign (parent compound is shown in black). Each sequence is represented with a colored line, where dots represent the score of the 15-mer peptide starting in that position. The tooltip shows the score of the 15-mer in position 75-89 for 'design24' and marks this peptide as not having a binding core in the human reference ("In reference: False"). B) ImmunoGeNN risk scores (pIRS%) correlate >95% with risk estimates calculated using the internal algorithm using NetMHCIIpan (IRS%). Correlations among peptides in the >83% IRS region is shown in red. C) ImmunoGeNN predicts risk scores at between 3 and 5 orders of magnitude the rate of the original algorithm, dependent on the length of input sequences. While ImmunoGeNN supports GPU-acceleration, both methods were compared on CPU only.

Model	*Validation set peptides*	
	Mean absolute error	**Spearman R**
MLPRegressor (baseline)	8.10%	95.0%
MLPRegressor (optimized)	6.20%	96.2%
+ binding core alignment	5.90%	96.3%

Assessing the model's performance on the validation set, comparing with risk scores as calculated with the internal algorithm, the model achieved a Spearman R correlation of 95.0% with a mean absolute error (MAE) of 8.10%. Manually tuning the model's hyperparameters against the validation set (see Methods) improved the performance to a Spearman R of 96.2% (MAE 6.20%), while implementation of dominant binding core alignment increased performance further to 96.3% (MAE 5.90%). Finally, evaluating ImmunoGeNN's performance on the NetMHCIIpan test set (Figure 1B), demonstrated a Spearman R correlation of 95.7% (MAE 8.30%). If choosing an IRS threshold of >83% to designate high-risk peptides, the Spearman R correlation in this region drops to 73.7% (Figure 1B).

Next, we evaluated the speed increase of the model by first generating a dataset of 150.000 random residues, and then randomly splitting it either across 10.000 individual sequences (15-mers), 1500 sequences (100-mers) or 375 sequences (400-mers). While ImmunoGeNN supports GPU-acceleration, these benchmarks were ran on CPU only for fair comparison.

Here, ImmunoGeNN was at least 5.000 times faster for sequences up to 400 residues, and an incredible 18.000 to 330.000 times faster for sequences between 15-100 residues (see Figure 1C). While NetMHCIIpan is a closed-source software making direct architecture comparisons difficult, the speed-up comes at least in part from ImmunoGeNN needing only a single forward pass through a relatively small neural network architecture, and NetMHCIIpan's significant overhead in processing of shorter sequence inputs.

In a design campaign when assessing risk across thousands of designs, it is desirable to compare risk scores on the level of sequences instead of individual peptides. One way to perform this comparison is to simply sum across all peptides in a sequence, to produce summed sequence scores. Performing this analysis across sequences in the NetMHCIIpan test set, we find that ImmunoGeNN maintains a 98.8% Spearman R correlation in rankings with the internal algorithm (Figure S1).

Human reference filtering

When assessing early immunogenicity risk, the goal is to evaluate both peptide presentation likelihood and immune response potential. Peptides in the human proteome are commonly presented on MHCs but cause no immune activation in healthy individuals due to central tolerance Jurewicz and Stern [2019].

The internal risk algorithm de-selects peptides with binding cores present in the human proteome by assigning them a risk score of 0. Applying this approach to the NetMHCIIpan test set, ~33% of peptides receive a score of 0 (see Figure S2 and Methods). Correspondingly, ImmunoGeNN recovers the dominant binding core of each peptide and assigns it a score of 0 if present in the human proteome.

Here, we found >99% agreement between methods in correctly identifying these human binding cores, with only 0.61% false positive rate. With this de-selection enabled, ImmunoGeNN's performance increased to 98.3% Spearman R on the test set and 99.7% Spearman R for sequence-level rankings (Figure S1).

Comparison with Experimental MAPPs Data

To validate ImmunoGeNN's clinical relevance, we compared predictions to experimental MHC-associated peptide proteomics (MAPPs) data for vatreptacog alfa and human factor VIIa (Figure 2). MAPPs is a widely used assay for estimating drug immunogenicity risk, providing empirical evidence of MHC-II presented peptides in human donors Karle [2020].

Vatreptacog alfa, a discontinued factor VIIa analog for major bleeding, failed Phase III trials due to anti-drug antibodies (ADAs) in 11% of patients Lamberth et al. [2017]. Post-hoc analysis linked immunogenicity to two mutations (E296V, M298Q) versus human factor VIIa, despite >99% sequence identity. A third mutation (V158D) did not affect immunogenic risk. The authors also found that peptide-MHC-II binding affinity was an overall good ADA predictor (AUC = 0.71), with 100% of ADA patients also displaying high affinity binders, while none of the low-affinity-only patients developed ADAs Lamberth et al. [2017].

Comparing ImmunoGeNN and the NetMHCIIpan-based algorithm with MAPPs data for both vatreptacog alfa and human factor VIIa, both tools showed a ~45% Spearman correlation with MAPPS presented peptides (see Figure 2 and Methods for more details on this). Here, ImmunoGeNN maintained a 96% correlation with NetMHCIIpan, indicating high agreement in identifying immunogenic regions. If applying the previously suggested IRS threshold of 83%, both tools flag E296V and M298Q mutations as introducing non-self-peptides within a possible high-risk region (ImmunoGeNN max peptide IRS 89.4%, NetMHCIIpan 83.6%, Figure 2C), aligning with clinically observed immunogenic hotspots.

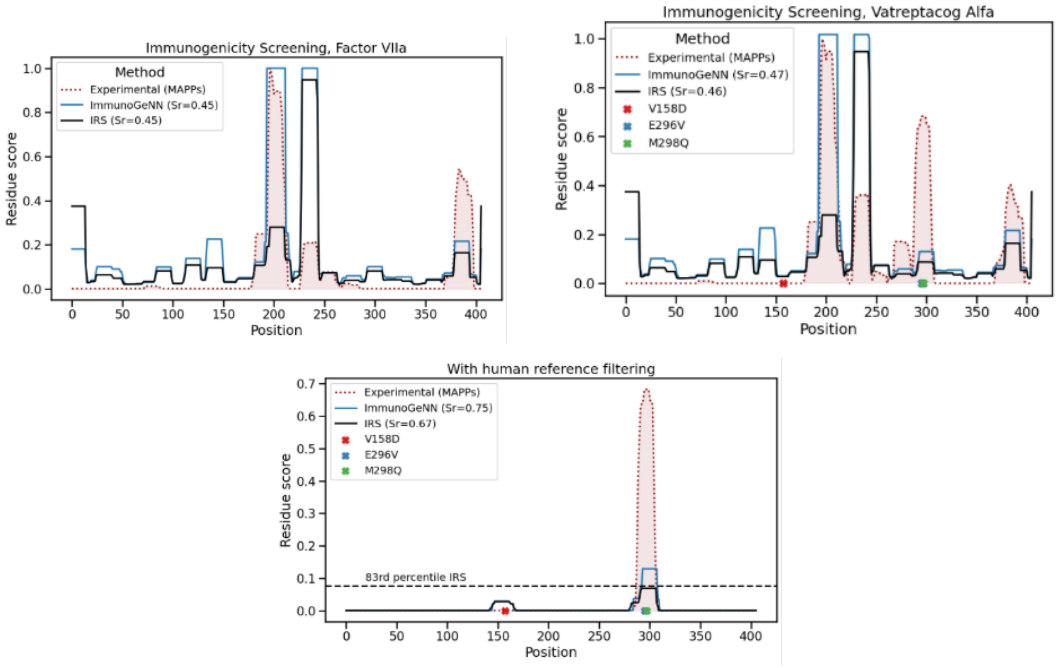

Figure 2: **Case study comparing the IRS profiles of ImmunoGeNN to the internal tool for the discontinued therapeutic vatreptacog alfa and human factor VIIa, overlaid on a MAPPs profile generated from presented peptides in a cohort of 20 and 18 donors respectively.** MAPPs profiles are calculated for each residue by summing the number of overlapping peptides. For ImmunoGeNN and IRS scores, the maximum peptide score for a given position is shown instead (capped at the 99th percentile for visualization purposes). A) Profiles for Factor VIIa, with ImmunoGeNN-MAPPs and IRS-MAPPs Spearman R correlation shown. B) Profiles for vatreptacog alfa, introducing three new mutations and a new MHC-presented region near position 300 (E296V, M298Q). C) Vatreptacog alfa profiles after removal of all peptides present in the human reference. The threshold for 83rd percentile IRS is shown in a dashed line.

3 Conclusion

We present ImmunoGeNN, a tool for in silico early immunogenicity risk assessment designed to support de novo biologics design. ImmunoGeNN runs 3-5 orders of magnitude faster than current state-of-the-art tools with only minor accuracy loss, enabling comprehensive screening across millions of designs and deimmunizing mutation screening in reasonable timeframes. We envision this tool being extremely useful for de novo biologics design, especially integrated with generative tools like RFDiffusion Watson et al. [2023], ProteinMPNN Dauparas et al. [2022] and ESM Lin et al. [2023].

We note some critical limitations of ImmunoGeNN. It only predicts peptide-MHC-II presentation risk - a necessary but insufficient step in anti-drug antibody formation. While maintaining performance with the in-house IRS score for predicting MAPPs-presented peptides, both methods show moderate experimental correlation (Spearman R ~45%). This moderate correlation partly reflects that ImmunoGeNN considers only North American DRB1 alleles, ignoring DRB345, DP and DQ gene classes known to affect immunogenicity. Correlation with IRS scores also drops for high-scoring peptides (IRS >83%) from 96% to 74%. Further clinical work should assess different risk thresholds and when to use more detailed allele-specific predictions like NetMHCIIpan.

ImmunoGeNN is released open-source so users can retrain with different immunogenicity risk functions, allele sets, or populations. It's freely available for commercial use at:

1. DTU Healthtech: `https://services.healthtech.dtu.dk/services/ImmunoGeNN/`
2. BioLib (mirror): `https://biolib.com/DTU/ImmunoGeNN/`

Dataset Construction

The training dataset comprised 75,000 unique 15-mer peptides randomly extracted from NetMHCIIpan-4.3 training partitions 0-3, with 4,500 validation peptides from partition 4. The test set contained all unique 15-mers extracted from 195 sequences in the NetMHCIIpan-4.3 evaluation set. North American DRB1 allele frequencies were obtained from internal data and AlleleFrequencies.net (January 2018), yielding 97 DRB1 alleles with frequency data.

Immunogenicity Risk Score (IRS) Calculation For each peptide-DRB1 allele pair, NetMHCIIpan-4.3 eluted ligand rank scores were inverted and capped at 5.00 (corresponding to 0.2% rank), with stronger binders receiving higher scores. These scores were weighted by DRB1 allele frequency, and the final peptide IRS was calculated as the sum across all peptide-allele scores.

$$\text{Peptide-allele score}_i = \min\left(\frac{1}{\text{Rank}_{\text{EL},i}}, 5.00\right) \times \text{allele frequency}_i \tag{1}$$

$$\text{Peptide IRS} = \sum_{i=1}^{N} \text{Peptide-allele score}_i \tag{2}$$

Formula 1: **Peptide early immunogenicity risk algorithm**. Calculation of peptide IRS, adjusted for allele population frequency numbers and NetMHCIIpan eluted ligand rank. Higher peptide IRS scores indicate stronger peptide to MHC-II binding among the DRB1 gene class, particularly across common DRB1 alleles in the North American population.

To handle outliers, IRS scores were rank-normalized using Scikit-Learn's QuantileTransformer, with models trained to predict IRS percentiles. For visualization purposes, we provide an optional backtransformed score using the fitted QuantileTransformer's inverse_transform method, mapping predicted IRS% scores back to an unnormalized scale (Supplementary figure S3).

Feature Engineering and Model Architecture

Each 15-mer peptide was one-hot encoded using 21 amino acids (20 canonical + 'X' for unknown) into 15×21 arrays. For binding core alignment, peptides were mapped to 21-position sequences with the dominant 9-mer binding core (highest contributing IRS score) aligned to positions 6-14, unfilled positions marked as 'X', then one-hot encoded to 21×21 arrays. During test set evaluation, the dominant binding core was predicted at test-time by screening for which of the 7 possible aligned binding core inputs led to the highest IRS score, and choosing this as the predicted score.

The final MLPRegressor model used two hidden layers (50 and 25 neurons), adaptive learning rate (starting 0.001), ReLU activation, Adam optimizer, mean-squared-error loss, maximum 1000 iterations with early stopping after 10 non-improving iterations (tolerance 0.0001).

Reference Filtering

Human proteome sequences were downloaded from UniProt identifier UP000005640 (downloaded 18th April 2024), before extracting all unique 9-mer peptides Bateman et al. [2023]. We also downloaded all human V and J antibody germlines extracted from IMGT's GENE-DB Giudicelli et al. [2005] using Anarci v1.1 (October 2024), extracting all unique 9-mers Dunbar and Deane [2016].

Reference filtering was performed by screening for matches between the peptide-allele binding core as identified by NetMHCIIpan-4.3, and if a match was found in the human reference, assigning the peptide-allele a score of 0. At test time, we screened all 7 possible aligned binding cores of a peptide. Next, we picked the top scoring binding core as the "consensus" binding core and used this for look for matches in the human reference. If a match was found, we assigned the peptide an IRS score of 0.

MAPPs Analysis

Vatreptacog alfa MAPPs data (611 peptides, 20 donors) came from internal sources; factor VIIa data (347 peptides, 18 donors) from literature Attermann et al. [2021]. Factor VIIa sequence (UniProt P08709) was trimmed by 60 N-terminal residues for alignment. Residue scores were calculated as normalized overlap counts across donors (0-1 range). Both sequences were analyzed using ImmunoGeNN and NetMHCIIpan-based IRS algorithms with/without reference filtering.

MAPPs Generation

Vatreptacog alfa MAPPs followed established protocols Attermann et al. [2021]. For factor VIIa, immature DCs were pulsed with 300nM protein for 4 hours (vs. 100nM in original), matured overnight with LPS (1 µg/ml). LC gradient was optimized: 5-15% over 3 minutes, then 15-35% linear over 57 minutes, maintaining original flow rate.

References

A. Attermann, C. Barra, B. Reynisson, et al. Improved prediction of hla antigen presentation hotspots: Applications for immunogenicity risk assessment of therapeutic proteins. *Immunology*, 162: 208–219, 2021.

A. Bateman, M.-J. Martin, S. Orchard, et al. Uniprot: the universal protein knowledgebase in 2023. *Nucleic Acids Research*, 51:D523–D531, 2023.

S. Chang, D. Ghosh, D. Kirschner, et al. Peptide length-based prediction of peptide-mhc class ii binding. *Bioinformatics*, 22:2761–2767, 2006.

J. Dauparas, I. Anishchenko, N. Bennett, et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378:49–56, 2022.

J. Dunbar and C. Deane. Anarci: antigen receptor numbering and receptor classification. *Bioinformatics*, 32:298–300, 2016.

J. Gehin, G. Goll, M. Brun, et al. Assessing immunogenicity of biologic drugs in inflammatory joint diseases: Progress towards personalized medicine. *BioDrugs*, 36:731–748, 2022.

V. Giudicelli, D. Chaume, and M.-P. Lefranc. Imgt/gene-db: a comprehensive database for human and mouse immunoglobulin and t cell receptor genes. *Nucleic Acids Research*, 33:D256–D261, 2005.

F. Gonzalez-Galarza, A. McCabe, E. d. Santos, et al. Allele frequency net database (afnd) 2020 update: gold-standard data classification, open access genotype data and new query tools. *Nucleic Acids Research*, 2019.

E. Howard, M. Goens, L. Susta, et al. Anti-drug antibody response to therapeutic antibodies and potential mitigation strategies. *Biomedicines*, 13:299, 2025.

W. Jankowski, Y. Park, J. McGill, et al. Peptides identified on monocyte-derived dendritic cells: a marker for clinical immunogenicity to fviii products. *Blood Advances*, 3:1429–1440, 2019.

V. Jawa, F. Terry, J. Gokemeijer, et al. T-cell dependent immunogenicity of protein therapeutics preclinical assessment and mitigation-updated consensus and review 2020. *Frontiers in Immunology*, 11:1301, 2020.

M. Jurewicz and L. Stern. Class ii mhc antigen processing in immune tolerance and inflammation. *Immunogenetics*, 71:171–187, 2019.

A. Karle. Applying mapps assays to assess drug immunogenicity. *Frontiers in Immunology*, 11:698, 2020.

K. Lamberth, S. Reedtz-Runge, J. Simon, et al. Post hoc assessment of the immunogenicity of bioengineered factor viia demonstrates the use of preclinical tools. *Science Translational Medicine*, 9, 2017.

Z. Lin, H. Akin, R. Rao, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379:1123–1130, 2023.

A. Moslemi, A. Briskina, Z. Dang, et al. A survey on knowledge distillation: Recent advancements. *Machine Learning with Applications*, 18:100605, 2024.

J. Nilsson, S. Kaabinejadian, H. Yari, et al. Accurate prediction of hla class ii antigen presentation across all loci using tailored data acquisition and refined machine learning. *Science Advances*, 9: eadj6367, 2023.

F. Pedregosa, G. Varoquaux, A. Gramfort, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

B. Reynisson, B. Alvarez, S. Paul, et al. 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:W449–W454, 2020.

J. Watson, D. Juergens, N. Bennett, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620:1089–1100, 2023.

V. Yip, A. Alfirevic, and M. Pirmohamed. Genetics of immune-mediated adverse drug reactions: a comprehensive and clinical review. *Clinical Reviews in Allergy & Immunology*, 48:165–175, 2015.

A Supplementary figures

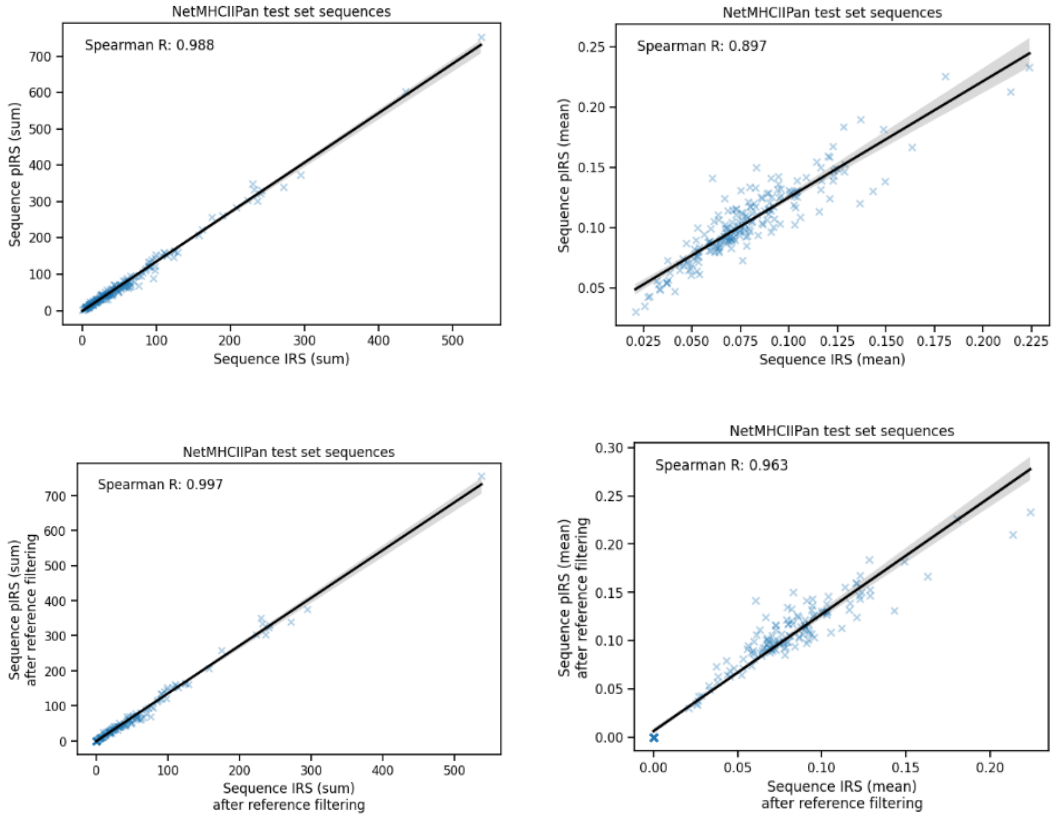

Figure 3: **ImmunoGeNN strongly recovers IRS ranking of sequences by immunogenicity risk.** A) Ranking of NetMHCIIpan test set sequences, calculating a single sequence score either by their summing or averaging across all its peptide scores B) Ranking of sequences following reference filtering (see Methods).

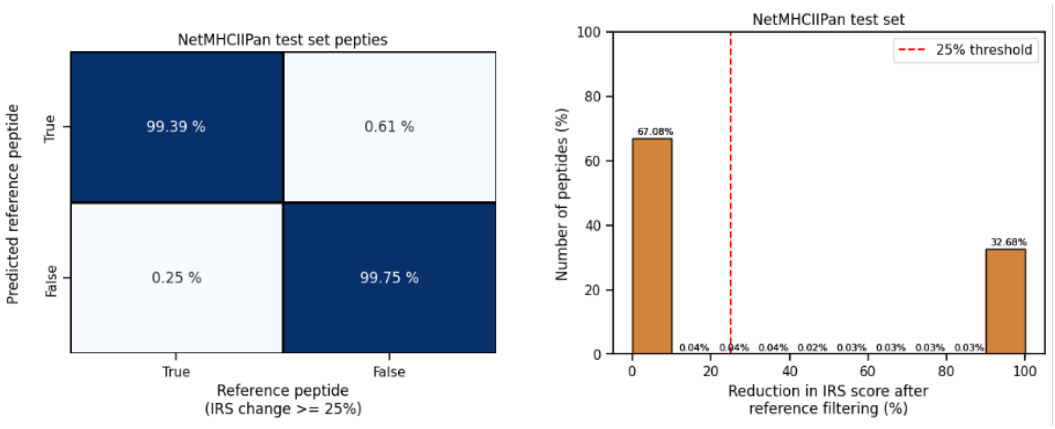

Figure 4: **ImmunoGeNN identifies reference peptides with >99% accuracy.** A) Confusion matrix of peptides in the human reference with at least 25% IRS score reduction after reference filtering (see Methods), and ImmunoGeNN predicted reference peptides (i.e. predicted binding core match with human reference). B) Distribution of reduction in IRS scores after reference filtering, with ~67% of peptides showing no change in IRS score and ~33% a 100% reduction in score.

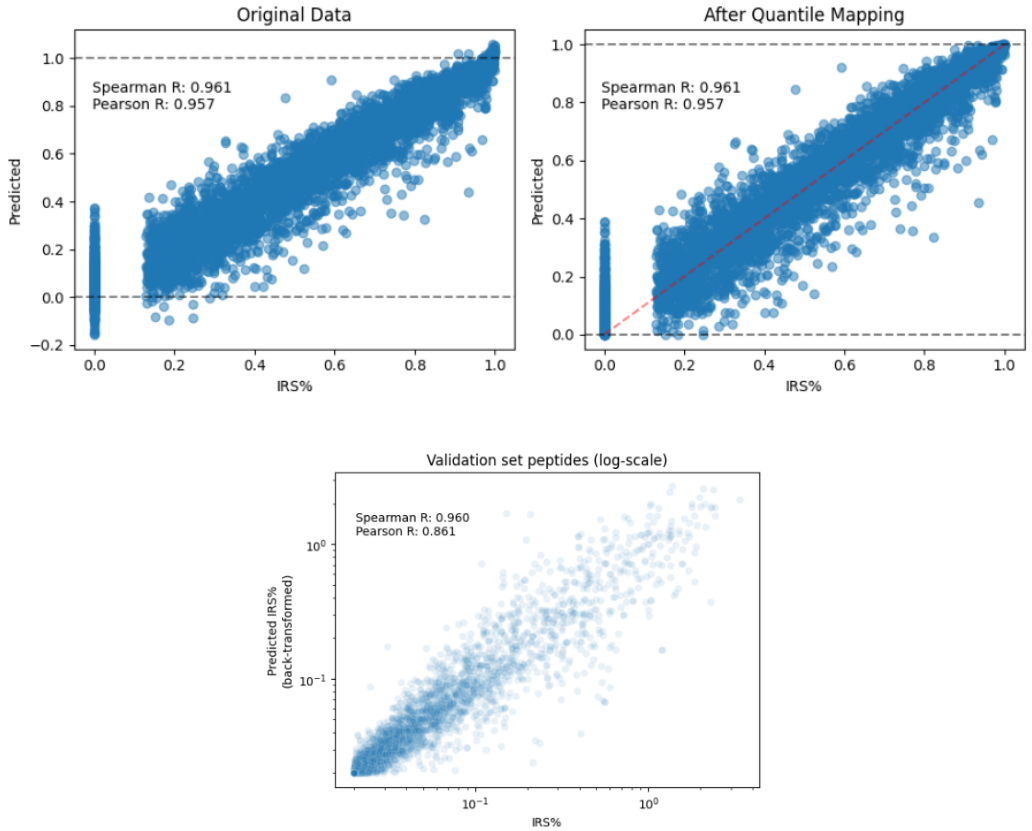

Figure 5: **Back-normalization of validation set peptides.** A) Predicted peptide scores are first quantile transformed back to a 0-100% range (with a QuantileTransformer fitted on the training set predictions), and next B) back-transformed to a raw IRS score distribution using an inverse transform (with a QuantileTransformer fitted on the training set IRS score distribution) C) Predicted IRS values after back-transformation versus IRS values.

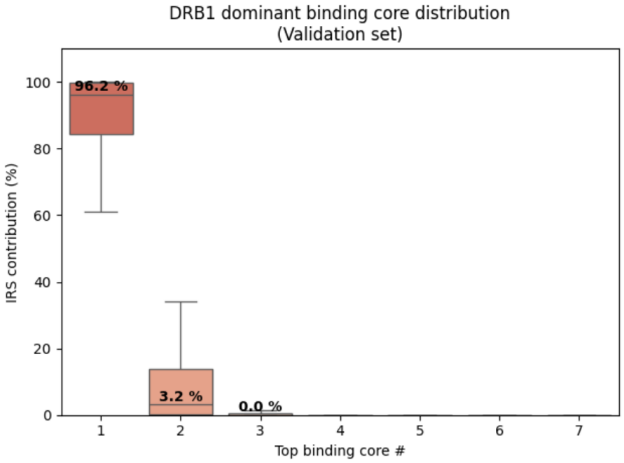

Figure 6: **IRS contribution across top 7 binding cores for peptides in the NetMHCIIpan validation set.** Filtered for peptides with an IRS > 85%. Scores are calculated from the peptide-allele IRS score and NetMHCIIpan identified binding core, and their relative contributions to the final peptide IRS.

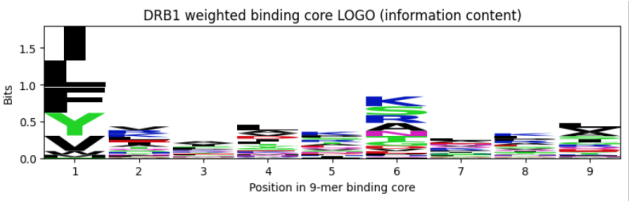

Figure 7: **LOGO plot of DRB1 binding core motifs as assessed by their IRS.** Calculated from the NetMHCIIpan validation set, only including peptides with an IRS > 85%. Calculated using LOGOmaker (Logomaker: beautiful sequence logos in Python | Bioinformatics | Oxford Academic)