Fragment ion intensity prediction improves the identification rate of non-tryptic peptides in TimsTOF

Charlotte Adams[1,2], Wassim Gabriel[3], Kris Laukens[1], Mathias Wilhelm[3], Wout Bittremieux[*,1], Kurt Boonen[*,2,4]

1 Department of Computer Science, University of Antwerp, Antwerp, Belgium

2 Centre for Proteomics (CFP), University of Antwerp, Antwerp, Belgium

3 Computational Mass Spectrometry, Technical University of Munich, 85354 Freising, Germany

4 Sustainable Health Department, Flemish Institute for Technological Research (VITO), Antwerp, Belgium

Abstract

Immunopeptidomics plays a crucial role in identifying targets for immunotherapy and vaccine development. Because the generation of immunopeptides from their parent proteins does not adhere to clear-cut rules, rather than being able to use known digestion patterns, every possible protein subsequence within HLA class-specific length restrictions needs to be considered. This leads to an inflation of the search space and results in lower spectrum annotation rates. Rescoring is a powerful enhancement of standard sequence database searching that boosts the spectrum annotation performance. In the field of immunopeptidomics low abundant peptides often occur, which is why the highly sensitive timsTOF instruments are increasingly gaining popularity. To improve rescoring for immunopeptides measured using timsTOF instruments, we trained a deep learning-based fragment ion intensity prediction model. Over 300,000 synthesized non-tryptic peptides from the ProteomeTools project were analyzed on a timsTOF-Pro to generate a dataset that was used to fine-tune an existing Prosit model. By applying our fragment ion intensity prediction model, we demonstrate up to 3-fold improvement in the identification of immunopeptides. Furthermore, our approach increased detection of immunopeptides even from low input samples.

Introduction

The adaptive immune system can eradicate pathogen-infected and cancerous cells by recognising peptides bound to major histocompatibility complex (MHC) molecules present on the cell surfaces. Even in the absence of infectious agents or cancerous transformation, the continuous yet dynamic process of peptide presentation informs the adaptive immune system about the health state of cells [1]. In immunopeptidomics, MHC-bound peptides—commonly termed immunopeptides—are isolated and characterized using mass spectrometry (MS). The identification of immunopeptides is critical for the development of immunotherapies and vaccines. In recent years MS-based immunopeptidomics has been used to discover T cell targets against tumors, autoimmune diseases, and pathogens [2–5]. As even a single immunopeptide could elicit an immune response [6], potential targets can be based on a single

peptide-spectrum match (PSM). This underscores the importance of the specificity of PSM annotations.

Unfortunately, however, it remains challenging to identify immunopeptides from MS data. Because the generation of immunopeptides from their parent proteins lacks clear-cut rules, rather than being able to use known digestion patterns, every possible protein subsequence within HLA class-specific length restrictions needs to be considered. As a result, there is a significant inflation of the search space, leading to an increased false positive rate and a low peptide identification sensitivity [7]. In immunopeptidomics the search space is often expanded further by incorporating somatic mutations, pathogen genomes, and unannotated open reading frames (nuORFs). A recent study highlights the significance of nuORFs as an underexplored source of MHC-I-presented, tumor-specific peptides that hold potential as targets for immunotherapy [8].

To minimize false positives and improve identification rates, PSM rescoring can be used. This involves post-processing results from an unfiltered database search using machine learning algorithms, such as Percolator [9], to use multiple PSM features to distinguish between correct and incorrect PSMs. Recently, driven by powerful prediction tools, there has been significant interest in using additional features for PSM rescoring. One example is using MS/MS spectrum prediction tools to generate spectral features based on the similarity between experimental and predicted fragment ion intensities. This approach is especially relevant for immunopeptidomics, where the use of specialized fragment ion intensity prediction tools has yielded promising results [10–12]. Specifically, the use of Prosit led to a more than two-fold increase in the identification of HLA ligands [11].

A timsTOF mass spectrometer (Bruker) combines two stages of trapped ion mobility spectrometry (TIMS) with a quadrupole and a high-resolution time-of-flight (TOF) mass analyzer. This configuration introduces an additional dimension, the collisional cross section, that can separate isobaric peptides. During a single TIMS scan, multiple precursors can be selected as a function of ion mobility, while the first TIMS accumulates ions for the next TIMS scan. This scan mode, termed parallel accumulation-serial fragmentation (PASEF), increases MS/MS rates more than ten-fold without any loss in sensitivity [13].

In the context of immunopeptidomics, it is critical to use highly sensitive instrumentation due to the relatively low abundance of immunopeptides. A timsTOF-based approach has been shown to significantly increase HLA peptide identifications compared to immunopeptidomics using an Orbitrap mass spectrometer [14]. Furthermore, optimization of the timsTOF acquisition method has been recently demonstrated to further improve HLA peptide identification rates [15]. Moreover, a recent study has revealed that MS/MS spectra from timsTOF instruments exhibit more reproducibility at low abundances compared to MS/MS spectra from Orbitrap instruments [16]. Notably, when analyzing a hybrid proteome mixture using different instruments, substantial differences in fragment ion intensities were observed between timsTOF Pro and Orbitrap QE HF-X mass spectrometers [17]. While PSM rescoring has been proven to be highly effective for immunopeptides measured on an Orbitrap [11], the considerable dissimilarity in MS/MS spectra produced by timsTOF and Orbitrap instruments necessitates the development of fragment ion intensity prediction models that are optimized for predicting timsTOF data.

In this study we measured over 300,000 synthesized non-tryptic peptides from the ProteomeTools project [18] on a timsTOF-Pro to fine-tune the existing Prosit model [11]. The integration of fragment ion intensity predictions into the database searching process significantly improved the identification rate of HLA peptides measured on a timsTOF

compared to an Orbitrap. In addition, we rescored timsTOF data from low-input samples and successfully identified immunopeptides derived from novel unannotated ORFs (nuORFs).

Results

Measuring non-tryptic peptides on a timsTOF

The ProteomeTools project is a large-scale effort in which peptides were synthesized and analyzed. Initially it contained measurements of 330,000 synthetic tryptic peptides covering essentially all canonical human proteins [18]. Subsequently, the project expanded to include post-translational modifications [19] and non-tryptic peptides [11]. This valuable dataset was used to train the deep neural network, Prosit, for the prediction of retention time and fragment ion intensity [20]. However, all measurements conducted in previous studies were performed on Orbitrap and ion trap instruments.

The considerable dissimilarity in MS/MS spectra generated by timsTOF and Orbitrap instruments for the same peptide (Fig. 1, Supplementary Fig. S1) underscores the need to develop fragment ion intensity prediction models optimized for timsTOF data. To address this, we measured over 300,000 non-tryptic peptides from the ProteomeTools project [11]. Our measurements encompassed a range of collision energies from 20.81 eV to 69.77 eV, enabling us to investigate the impact of collision energy on fragment ion intensities. Consequently, we compiled a dataset consisting of 93,227 non-tryptic MS/MS spectra, complemented by 184,554 previously published tryptic MS/MS spectra [21]. This extensive dataset, comprising a total of 277,781 MS/MS spectra, serves as a unique training dataset for the development of machine learning tools tailored to timsTOF instruments.

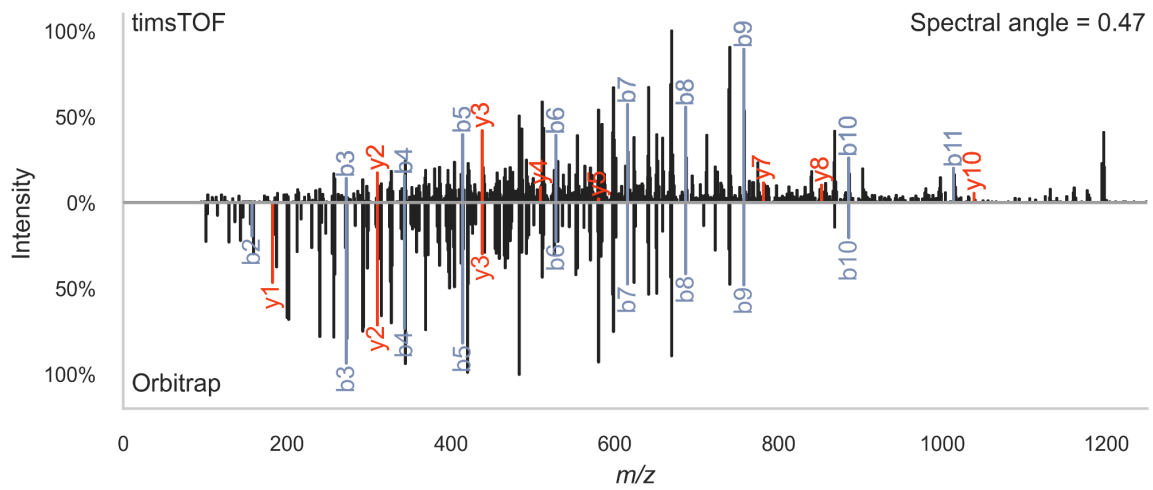

Fig. 1 Mirror plot of the singly charged non-tryptic synthetic ProteomeTools [18] peptide GVDAANSAAQQY measured on a timsTOF (top; mzspec:PXD043844:HLAI_p2_97_178_p2-D1_S1-D1_1_6866.mgf:index:3378:GVDAANSAAQQY/1) and on an Orbitrap (bottom; mzspec:PXD021013:02446d_GD1-TUM_HLA_133_01_01-3xHCD-1h-R4:scan:14565:GVDAANSAAQQY/1) instrument. The spectral similarity measured by the normalized spectral contrast angle based on annotated fragments between the two spectra is 0.47. This illustrates how different timsTOF MS/MS spectra can look compared to Orbitrap data. This randomly chosen peptide was measured several times on the Orbitrap, after which the spectrum with the highest similarity to all other Orbitrap spectra for this peptide was

selected (the medoid spectrum). In the timsTOF data the displayed MS/MS spectrum was the only measurement of this peptide.

Optimized Prosit model allows accurate prediction of tryptic and non-tryptic peptide timsTOF MS/MS spectra

To optimize the Prosit fragment ion intensity prediction model towards timsTOF instruments, we fine-tuned the HCD Prosit 2020 model using the 277,781 MS/MS spectra obtained in this study, split into training, validation, and test sets (Fig. 2a). The HCD Prosit 2020 model was originally trained on approximately 30 million MS/MS spectra, consisting of 9 million MS/MS spectra of non-tryptic peptides [11] and 21 million previously published tryptic MS/MS spectra [18,20]). The comparison between the HCD Prosit 2020 model and the newly developed TOF Prosit 2023 model (Fig. 2b-d) reveals a substantial improvement in normalized spectral contrast angle (SA) between predicted and experimental timsTOF MS/MS spectra for non-tryptic peptides (SA ≥ 0.9 for 26.3% of spectra, compared to 2.4% with HCD Prosit 2020) and for tryptic peptides (SA ≥ 0.9 for 42.1% of spectra, compared to 0.2% with HCD Prosit 2020). The TOF Prosit 2023 model demonstrates consistent performance across different precursor charges, peptide lengths, and collision energies (Supplementary Fig. S2a-c), with minimal bias towards C- and N-terminal amino acids (Supplementary Fig. S3a-b).

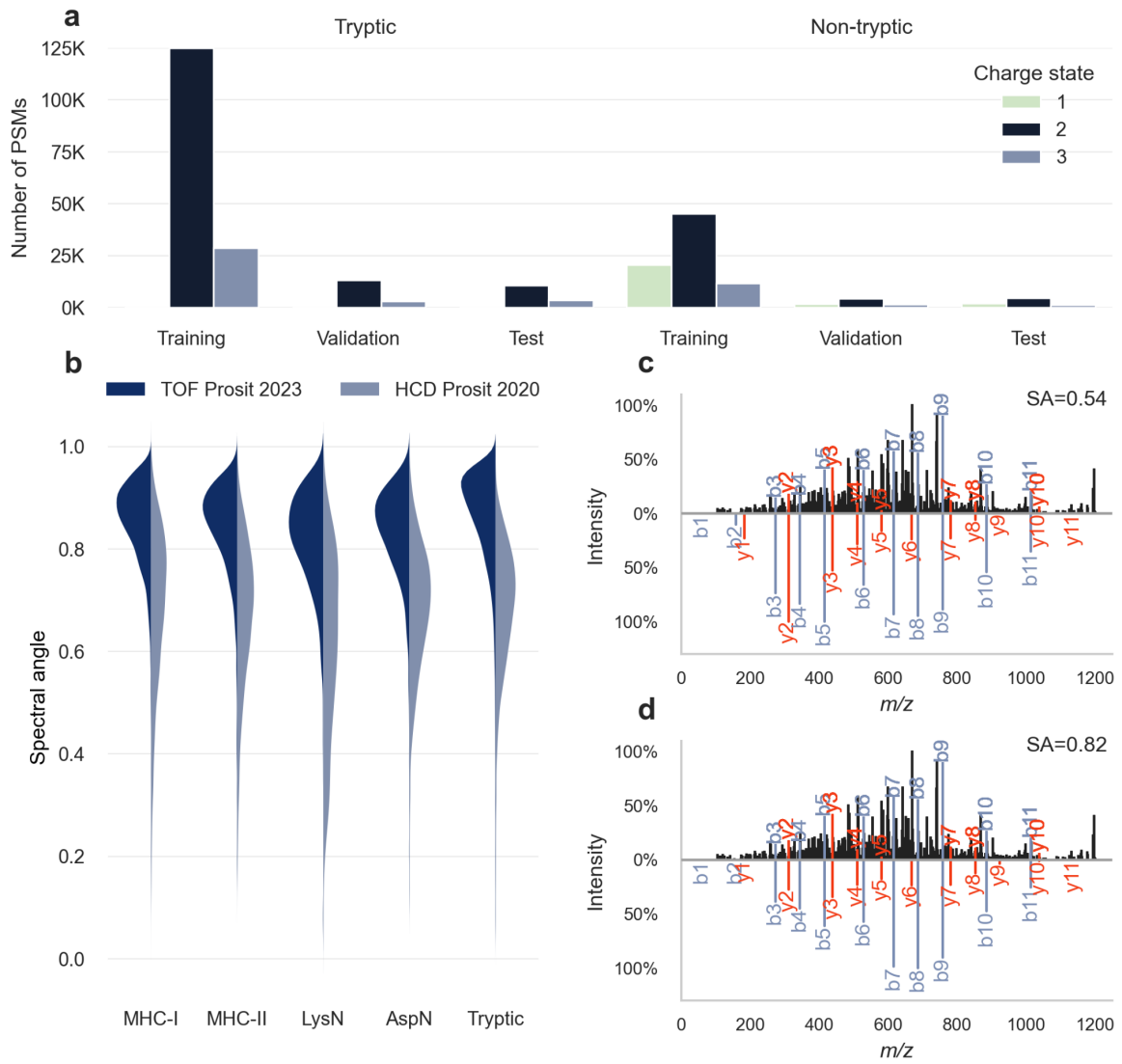

Fig. 2 Deep learning framework Prosit for tryptic and non-tryptic peptide fragment ion intensity prediction. a The 277,781 MS/MS spectra from tryptic and non-tryptic peptides measured on timsTOF instruments were split into training, validation, and test sets and used to fine-tune Prosit. **b** Violin plots comparing the prediction accuracy of the TOF Prosit 2023 model against the the previously published HCD Prosit 2020 model [11] for non-tryptic (MHC-I, MHC-II, LysN, and AspN) and tryptic peptides. **c** Mirror plot of the randomly chosen singly charged non-tryptic synthetic peptide GVDAANSAAQQY measured on a timsTOF (top; mzspec:PXD043844:HLAI_p2_97_178_p2-D1_S1-D1_1_6866.mgf:index:3378:GVDAANSAAQQY/1) and the predicted spectrum for this peptide at the aligned collision energy with the HCD Prosit 2020 model (bottom). **d** Mirror plot of the same measurement (top; mzspec:PXD043844:HLAI_p2_97_178_p2-D1_S1-D1_1_6866.mgf:index:3378:GVDAANSAAQQY/1) and the predicted spectrum for this peptide at the aligned collision energy with the TOF Prosit 2023 model (bottom). Fragment ions are labeled in red and blue for y and b ions, respectively. The mirror plot was generated using spectrum_utils version 0.4.1 [22]. SA = normalized spectral contrast angle.

PSM rescoring boosts immunopeptide identification on timsTOF compared to Orbitrap

We hypothesized that integrating fragment ion intensity predictions into the database searching process would improve the identification rate of HLA peptides measured on a timsTOF, similar to what was previously observed for tryptic and non-tryptic peptides measured on other instruments [11]. To investigate this, we reanalyzed data from a recently published benchmarking study on timsTOF-based immunopeptidomics for tumor antigen discovery [14]. The study compared timsTOF-based immunopeptidomics to immunopeptidomics using Orbitrap technology and demonstrated a significant increase in the identification of immunopeptides from various benign and malignant primary samples of solid tissue and hematological origin.

In this analysis, the dataset was reprocessed with MaxQuant and all proposed PSMs were rescored by integrating Prosit's fragment ion intensity predictions, using Oktoberfest (https://github.com/wilhelm-lab/oktoberfest). This allowed us to compare rescoring of timsTOF data using the TOF 2023 model to rescoring of Orbitrap data using the CID 2020 model for HLA-I and HCD 2020 model for HLA-II (Fig. 3a-d). Rescoring the Orbitrap data resulted in on average 2.5-fold more unique HLA-I peptides and 1.4-fold more unique HLA-II peptides. In contrast, rescoring timsTOF data resulted in a higher increase, with on average 2.8-fold more unique HLA-I peptides and 1.7-fold more unique HLA-II peptides.

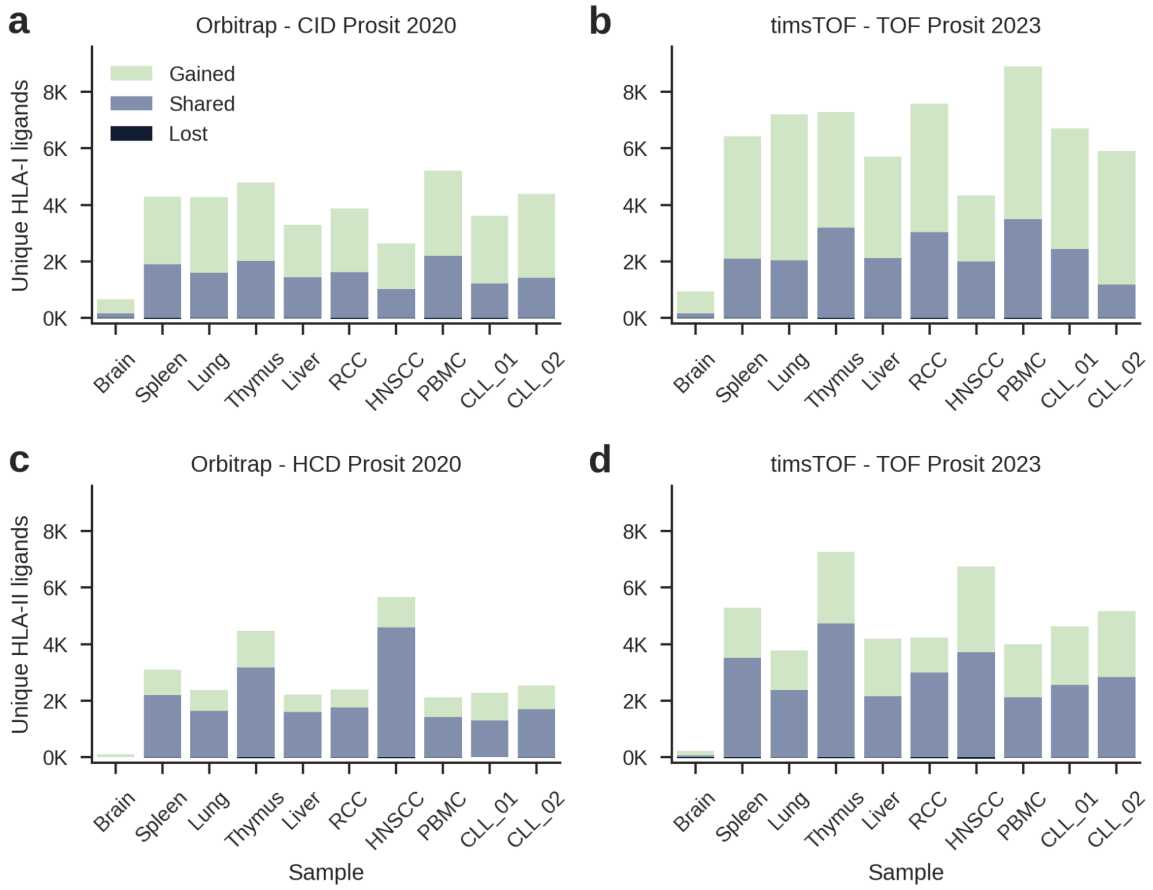

Figure 3. Gained, shared, and lost peptide identifications for different sample types to compare rescoring on Orbitrap data with rescoring on timsTOF data. In general rescoring was able to boost the confidence in peptide identifications, retaining true PSMs, gaining new PSMs, and losing only a small number of previously incorrect PSMs. **a** Rescoring of HLA-I

Orbitrap data with the CID Prosit 2020 model resulted in a 2.5-fold increase. **b** Rescoring of HLA-I timsTOF data with the TOF Prosit 2023 model resulted in a 2.8-fold increase. **c** Rescoring of HLA-II Orbitrap data with the HCD Prosit 2020 model resulted in a 1.4-fold increase. **d** Rescoring of HLA-II timsTOF data with the TOF Prosit 2023 model resulted in a 1.7-fold increase. RCC = renal cell carcinoma; HNSCC = head and neck squamous-cell carcinoma; PBMC = peripheral blood mononuclear cell; CLL = chronic lymphocytic leukemia.

Rescoring of melanoma immunopeptides reveals novel neo-epitopes

To enable the detection of rare and clinically relevant antigens from a limited cell input, Phulphagar et al. [23] developed a high-throughput single-shot MS-based immunopeptidomics workflow using the timsTOF single-cell proteomics system (SCP). This workflow was applied to sample inputs ranging from 1 million to 40 million A-375 cells, a melanoma cell line which expresses the following HLA genes: A*01:01, A*02:02, B*57:01, B*44:03, C*16:02, and C*06:02.

Low sample inputs frequently suffer from missing peaks, as fragments with low intensities fail to surpass the noise level. This leads to low database search engine scores and low identification rates. Consequently, the benefit of rescoring is expected to be even higher for low input samples. To validate this assumption, we performed a reanalysis on this timsTOF SCP dataset using the TOF Prosit 2023 model.

Individual spectrum peak files were searched against a compiled database consisting of the human reference proteome, common laboratory contaminants, curated small open reading frames (ORFs), and novel unannotated ORFs (nuORFs) supported by ribosomal profiling [8]. All proposed PSMs by MaxQuant were subsequently rescored using Oktoberfest. The results showed an average increase in identified HLA-I ligands across different cell input sizes, ranging from 1.3-fold at 1 million cells to 1.9-fold at 40 million cells (Fig. 4a).

To validate the peptide identifications obtained through PSM rescoring, we employed Gibbs clustering [24] on the gained, shared, and lost peptides separately. We then compared the cluster motifs with the known binding motifs of the HLA alleles expressed by the cells. The selection of motifs shown in Fig. 4b was based on the cluster with the highest Kullback-Leibler distance. The Kullback-Leibler distance provides a measure of similarity between clusters, thus identifying the cluster that differs the most from the other clusters found. Notably, we observed that the clusters with the highest Kullback-Leibler distances to the other clusters among the shared and gained peptides exhibited a striking resemblance to the motif of A*01:01. Conversely, the motifs of the clusters of the lost peptides did not correspond to any of the motifs of the HLA types present in the cell line (Fig. 4b, Supplementary Fig. S4). The motifs of the other clusters based on the gained and shared peptides were consistent with other HLA alleles present in the cell, namely A*02:02, B*44:03, and B*57:01 (Supplementary Fig. S4).

To further validate the peptide identifications obtained through PSM rescoring, we assessed the predicted binding affinity of the gained, shared, and lost peptides. Using thresholds provided by NetMHCpan for weak binders and strong binders, we found that 88% of peptides gained after rescoring were weak binders of at least one of the HLA types present in the cell, with 80% being a strong binder (Fig. 4c). For the shared peptides this was 89% and 85%, and for the lost peptides this was 44% and 24%, respectively. This implies that 56% of the peptides lost after rescoring were predicted to not bind any of the HLA molecules present in the cell.

Among the identified immunopeptides, a subset of 2251 peptides (2%) originated from nuORF source proteins (Fig. 4d). Recent studies have provided evidence that peptides derived from noncanonical proteins can be displayed on HLA-I molecules [25,26]. These nuORFs may arise

from transcripts that are currently annotated as non-protein coding, including the 5' and 3' untranslated regions, overlapping yet out-of-frame alternative ORFs in annotated protein-coding genes, long noncoding RNAs, or pseudogenes [8]. HLA peptides derived from noncanonical proteins can expand the repertoire of potential immunotherapy targets in cancer. Notably, although we did not observe significant changes in the ratio of nuORFs after rescoring, we did identify twice as many nuORF source proteins, which is of great interest. Furthermore, we examined the binding affinity of peptides originating from nuORFs and found that 90% of peptides can be considered a weak binder to at least one of the HLA types present in the cell, with 81% being a strong binder. This suggests that these peptides are actually presented by the cell.

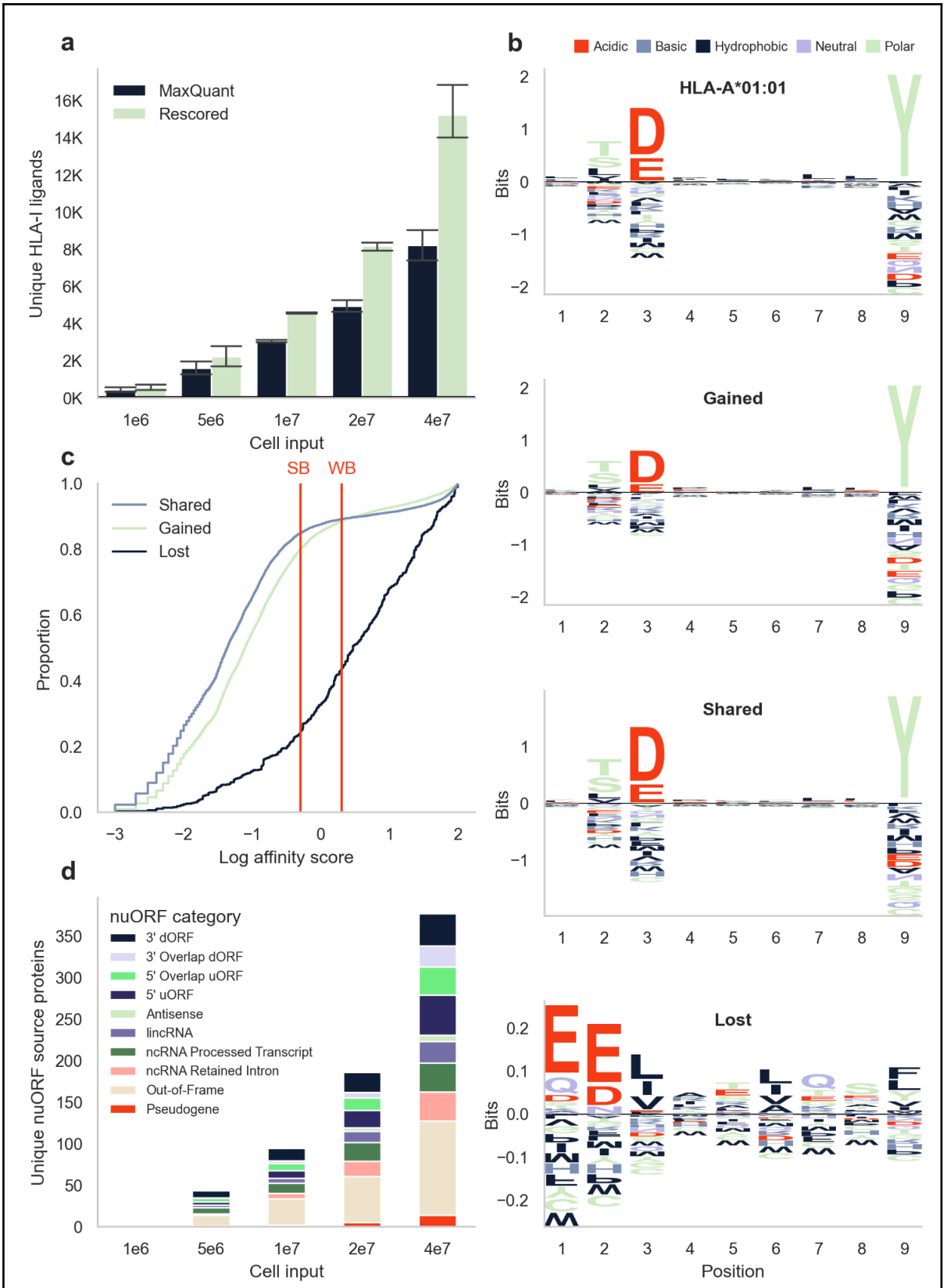

Figure 4. Rescoring timsTOF SCP immunopeptidomics data. a Mean and standard deviation showing the number of unique HLA-I peptides identified from 1 million to 40 million A-375 cells. **b** Peptide motif plots of 1,406 unique peptides confidently identified to be present in the cell line expressing allele A*01:01 [27], 16,641 unique gained peptides, 13,208 unique shared peptides, and 447 unique lost peptides. Amino acids are colored according to

their physicochemical properties (red acidic, blue basic, black hydrophobic, purple neutral, and green polar amino acids). **c** Cumulative distribution function (CDF) plot illustrating the distribution of the shared, gained, and lost peptides across the log10 affinity scores calculated with NetMHCpan version 4.1 [28]. The cut-off values for strong binders (SB) and weak binders (WB) are indicated in red. **d** Mean unique nuORF source proteins contributing to the HLA-I immunopeptidome of 1 million to 40 million A-375 cells [8].

Discussion

The identification of immunopeptides is critical for the advancement of vaccine and immunotherapy development. Previous studies have shown that using fragment ion intensity predictions in rescoring can greatly increase the identification rate of HLA ligands [10–12]. In this study, we established a comprehensive dataset consisting of MS/MS spectra from synthetic non-tryptic and tryptic peptides measured on a timsTOF instrument. This dataset served as the foundation for training the novel TOF Prosit 2023 model. By employing this model for rescoring immunopeptides measured on a timsTOF, we achieved a nearly 3-fold increase in the identification of HLA-I ligands. In addition, we demonstrated the effectiveness of our model for PSM rescoring of low sample inputs measured using a timsTOF SCP instrument, resulting in improved identification rates. Importantly, the immunopeptides identified after rescoring are likely to be HLA binders, as supported by the motif analysis and binding affinity assessment, providing an orthogonal validation of our method. Moreover, PSM rescoring led to an almost 2-fold increase in the identification of unique nuORF source proteins, which hold the potential to serve as valuable targets for immunotherapy [25,26].

To ensure the TOF Prosit 2023 model's strong predictive capabilities towards immunopeptides, we generated MS/MS spectra from synthesized non-tryptic peptides to compile the training data. This enables the model to generalize over different peptide types, whereas machine learning models that are solely trained on tryptic data often fail to do so, for example, by exhibiting a bias towards C-terminal arginine or lysine residues. A potential limitation could be that while our current model only predicts fragment ion intensities for canonical y and b ions, non-tryptic peptides exhibit distinct MS/MS characteristics compared to tryptic peptides, often displaying strong internal ion series and neutral losses. However, as PSM rescoring using Prosit has demonstrated robustness against the presence of a large number of neutral loss or internal ion series [11], we do not expect this to be overly detrimental. In addition to the analysis of immunopeptidomics data, our model holds promise for numerous other biological and biomedical applications. One such area is deep proteome sequencing, where multiple proteases are used to enhance proteomic coverage [29], particularly in regions with suboptimal trypsin cleavage sites, such as membrane-spanning domains and splice junctions. Our model can effectively enhance the confidence of peptide identifications in such studies, enabling valuable insights into alternative splicing and facilitating a comprehensive exploration of its impact on the proteome.

It is important to note that the applied collision energy has a profound impact on the information content of the obtained MS/MS spectra [30]. Thus, collision energy calibration is needed for accurate fragment ion intensity predictions. The impact of collision energy in the timsTOF instrument is a bit more complicated compared to its impact in the Orbitrap. During IMS, ions are subjected to a series of collisions. This kinetic energy can be transferred to internal energy, similarly to what takes place during the activation of ions in collision-induced dissociation. Because IMS energizes the peptides significantly, the use of lower collision energies is advised [30]. Similarly to what has been observed for retention time alignment [31], we expect a benefit from collision energy alignment to account for the run-to-run fluctuations.

Although currently the TOF Prosit 2023 model is dependent on MaxQuant, as it relies on the search engine to sum the MS/M scans, in the future, it will be further extended to support other search engines as well and become search engine agnostic, similar to how Oktoberfest has recently extended applicability of Prosit Orbitrap predictions beyond MaxQuant. The TOF Prosit 2023 model is available on koina (https://koina.proteomicsdb.org) and can be used via Oktoberfest (Prosit_2023_intensity_TOF).

Methods

Data acquisition

Synthetic non-tryptic peptides data acquisition

Within the ProteomeTools project, 305,331 non-tryptic peptides were synthesized, comprising 168,688 HLA class I, 73,464 HLA class II, 31,744 AspN, and 31,435 LysN sequences. For detailed information on the peptide origins, please refer to the original publication by Wilhelm et al. [11]. Peptide pools for synthesis and measurement contained roughly 1000 peptides each. Near-isobaric peptides (±10 p.p.m.) were distributed across different pools of similar length to avoid ambiguous masses in pools wherever possible. Ten microliters of the stock solution were transferred to a 96-well plate and spiked with two retention time standards (Pierce Retention Time Standard and PROCAL [32]) at 100 fmol per injection. An equimolar amount of approximately 50 fmol of each peptide was injected into an Evosep One HPLC system (Evosep) coupled to a hybrid TIMS-quadrupole TOF mass spectrometer (Bruker Daltonik timsTOF Pro) via a nano-electrospray ion source (Bruker Daltonik Captive Spray). The 100 SPD method was used. The Endurance Column 15 cm x 150 µm ID, 1.9 µm beads (EV1106, Evosep) was connected to a Captive Spray emitter (ZDV) with a diameter 20 µm (1865710, Bruker) (both from Bruker Daltonics).

The timsTOF Pro was calibrated according to the manufacturer's guidelines. The source parameters were: capillary voltage 1500 V, dry gas 3.0 l/min, and dry temp 180°C. The temperature of the ion transfer capillary was set to 180°C. The column was kept at 40°C. The data-dependent Parallel Accumulation–Serial Fragmentation (PASEF) method was used to select precursor ions for fragmentation with 1 TIMS-MS scan and 10 PASEF MS/MS scans, as described by Meier et al. [13]. The TIMS-MS survey scan was acquired between 0.70 and 1.45 Vs/cm^2 and 100–1,700 *m/z* with a ramp time of 100 ms. The *m/z* and ion mobility information was used to select precursors with charges ranging from 1–3. Dynamic exclusion was used to avoid re-sequencing of precursors that reached a target value of 20,000 a.u. The timsTOF Pro was controlled by the OtofControl 6.0 software (Bruker Daltonik GmbH). The collision energy was increased as a function of increasing ion mobility (ranging from 0.76–1.68 Vs/cm^2), starting from 20 eV to 70 eV.

Synthetic tryptic peptides data acquisition

The "proteotypic" synthetic peptide set from ProteomeTools [18], covering confidently and frequently identified proteins (124,875 peptides covering 15,855 human annotated genes), was obtained by Meier et al. [21]. The data was downloaded from the PRIDE repository with identifier PXD019086.

As per Meier et al., LC–MS was performed on an EASY-nLC 1200 (Thermo Fisher Scientific) system coupled to timsTOF Pro mass spectrometer (Bruker Daltonik, Germany) via a nano-electrospray ion source (Bruker Daltonik Captive Spray). Approximately 200 ng of peptides were separated on an in-house 45 cm × 75 µm reversed-phase column at a flow rate of 300 nL/min in an oven compartment heated to 60°C. The column was packed in-house with 1.9 µm C18 beads (Dr. Maisch Reprosil-Pur AQ, Germany) up to the laser-pulled electrospray

emitter tip. Mobile phases A and B were water and 80%/20% ACN/water (v/v), respectively, and both buffered with 0.1% formic acid (v/v). The pooled synthetic peptides were analyzed with a gradient starting from 5% B to 30% B in 35 min, followed by linear increases to 60% B and 95% in 2.5 min each before washing and re-equilibration for a total of 5 min.

The timsTOF Pro was operated in data-dependent PASEF [33] mode with 1 survey TIMS-MS and 10 PASEF MS/MS scans per acquisition cycle. They analyzed an ion mobility range from 1/K0 = 1.51 to 0.6 Vs/cm^2 using equal ion accumulation and ramp time in the dual TIMS analyzer of 100 ms each. Suitable precursor ions for MS/MS analysis were isolated in a window of 2 Th for $m/z < 700$ and 3 Th for $m/z > 700$ by rapidly switching the quadrupole position in sync with the elution of precursors from the TIMS device. The collision energy was lowered stepwise as a function of increasing ion mobility, starting from 52 eV for 0–19% of the TIMS ramp time, 47 eV for 19–38%, 42 eV for 38–57%, 37 eV for 57–76%, and 32 eV until the end. The m/z and ion mobility information was used to exclude singly charged precursor ions with a polygon filter mask. Dynamic exclusion was used to avoid re-sequencing of precursors that reached a target value of 20,000 a.u.

Preparation of the training data

The raw Bruker data from synthetic peptides from ProteomeTools [18] were analyzed with MaxQuant version 2.1.2.0 [34]. Individual spectrum peak files were searched against pool-specific databases [35]. Default parameters were used, unless mentioned otherwise: carbamidomethylated cysteine was specified as a fixed modification and methionine oxidation as a variable modification. The minimal sequence length was set to 7 and the maximum sequence length was set to the maximum length of peptides in the pool. PSMs were filtered at a 1% false discovery rate (FDR). Only peptides expected in the pool, including full-length and N-terminally truncated peptides, were selected. All PSMs, even for the same peptide, with an Andromeda score ≥ 70 were included.

Unprocessed spectra were extracted from the raw Bruker files with OpenTIMS [36], using the precursorID from the accumulatedMsmsScans.txt and the frameID from the pasefMsmsScans.txt MaxQuant output files. Frame-level scans were summed based on the scan number from msms.txt with MasterSpectrum version 1.1 [37]. The y and b ions were annotated for fragment charges ranging from 1 to 3.

The data was split into three distinct sets with each peptide and subsequence of a peptide only included in one of the three: training (80%, 153,809 tryptic PSMs and 77,577 non-tryptic PSMs), validation (10%, 16,483 tryptic PSMs and 7,778 non-tryptic PSMs), and test (10%, 14,262 tryptic PSMs and 7,872 non-tryptic PSMs). For each PSM in the training set, MS/MS spectra were predicted with the HCD Prosit 2020 model across collision energies ranging from 5 to 45 eV. The SA was calculated between the observed spectra and the predicted spectra, and the collision energy corresponding to the top-scoring predicted spectra was selected as the optimal collision energy. This process was performed separately for each peptide type (tryptic, non-tryptic) and precursor charge state (1-3). A robust linear model was trained using RANSAC regression in scikit-learn version 1.2.2 [38] to predict the difference between the reported collision energy and the optimal collision energy, based on the peptide mass.

To calibrate the validation and test set, the collision energy difference was predicted for each peptide mass, and this difference was applied to obtain the aligned collision energy. The models used for the collision energy calibration are available on the PRIDE repository (PXD043844).

Prosit 2023 model training

The HCD Prosit 2020 model [11] was fine-tuned using the training set. To control for overfitting, the validation set was used with early stopping, employing a patience of 5 epochs. The test set was used after the model was fully trained to evaluate its generalization and potential biases.

The model architecture remained unchanged, and the normalized spectral contrast loss [20] was used as a loss function. We used the Adam optimizer [39] with a cyclic learning rate algorithm [40]. During training, the learning rate cycled between a constant lower limit of 0.00001 and an upper limit of 0.0002 which is continuously scaled by a factor of 0.95 with the "triangular" mode. The model was trained with a batch size of 2000 on an Nvidia V100 GPU. The model improved significantly in predicting fragment ion intensity during the initial epochs, as depicted in Supplemental Fig. S5, and converged at epoch 28 with a median SA of 0.86.

General rescoring pipeline

Before rescoring, all spectrum peak files were searched using MaxQuant version 2.0.3.1 with default parameters unless specified otherwise: carbamidomethylated cysteine was specified as a fixed modification and methionine oxidation as a variable modification. The minimum peptide length was set to 8 amino acids and the maximum peptide length depended on the HLA class, with a length set to 16 amino acids for HLA-I and 30 amino acids for HLA-II. Specific settings for the individual datasets are detailed below.

The unfiltered search results, including decoy PSMs, were used as an input for the spectral intensity-based rescoring. The rescoring was performed using Oktoberfest v1.1 (https://github.com/wilhelm-lab/oktoberfest). In brief, unprocessed MS/MS spectra corresponding to the identifications were extracted from the raw Bruker files and the y and b ions were annotated at fragment charges 1 up to 3. Both retention time and fragment ion intensities were predicted and features were generated to add to Percolator [41], which was used for the PSM and peptide FDR estimation.

Application of Prosit and rescoring to external datasets

Re-analysis of the comparison dataset

To compare the rescoring performance on Orbitrap versus timsTOF data, we utilized a comparison dataset comprising both HLA-I and HLA-II peptides measured on an Orbitrap and on a timsTOF. For detailed information on data acquisition, please refer to the original publication by Gravel et al. [14]. In brief, 10 samples were measured in technical triplicate (two technical replicates for the HNSCC sample) on the Orbitrap Fusion Lumos mass spectrometer (Thermo Fisher Scientic, Waltham, USA) and on the timsTOF Pro (Bruker Daltonik, Germany). The fragmentation methods used for the Orbitrap instrument were collisional-induced dissociation (CID) at a normalized collision energy of 35% for HLA-I peptides and higher collisional dissociation (HCD) at a normalized energy of 30% for HLA-II peptides. The data was downloaded from the PRIDE repository with identifier PXD038782.

Individual spectrum peak files were searched against a database containing 20,598 human UniProt entries downloaded from https://www.ebi.ac.uk/reference_proteomes/ [35]. To perform rescoring on the Orbitrap data we employed the 2020 CID Prosit model with a collision energy set to 35 for HLA-I peptides, and the 2020 HCD Prosit model with collision energy set to 30 for the HLA-II peptides. For timsTOF data, rescoring was performed using the TOF Prosit 2023 model with the reported collision energies for each PSM.

Re-analysis of an immunopeptidomics dataset measured on timsTOF SCP

To investigate whether low input samples benefit from rescoring, we rescored a timsTOF SCP dataset. For detailed information on data acquisition, please refer to the original publication by Phulphagar et al. [23]. In brief, HLA-I peptides were directly enriched from 1 million to 40 million A-375 cells by single shot injections on timsTOF SCP. Each sample was measured in technical triplicate (four technical replicates for the 40 million sample). Individual spectrum peak files were searched against a compiled database comprised of the human reference proteome Gencode 34 (ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_34) with 47,429 non redundant protein-coding transcript biotypes mapped to the human reference genome GRCh38, 602 common laboratory contaminants, 2043 curated small ORFs (lncRNA and upstream ORFs), 237,427 novel unannotated ORFs (nuORFs) supported by ribosomal profiling nuORF DB v1.037, for a total of 287,501 entries [8]. The data was downloaded from the PRIDE repository with identifier PXD040740.

To validate the peptide identifications acquired through rescoring, gained, shared, and lost peptides were clustered with GibbsCluster version 2.0 [24] with parameters for MHC class I ligands of length 8-13. Based on the Kullback-Leibler distance in function of the number of clusters, the optimal number of motifs in the data was selected. For each motif the position-specific scoring matrix was extracted and put into Seq2logo version 2.0 [42] to get the position-specific frequency matrix of the Kullback-Leibler logos. All logos were visualized using the Python package Logomaker version 0.8 [43]. The logos from gained, shared, and lost peptides were plotted next to the logos of the HLA-types present in the cell line to which they had the lowest Kullback-Leibler distance. For the HLA motif, peptide lists of the large monoallelic HLA class I cell line study by Sarkizova and Klaeger et al. [27] were used.

For each peptide we calculated the binding affinity to every HLA allele present in the cell line, using NetMHCpan version 4.1 [28]. For each peptide the best, i.e. lowest, affinity score was retained. A percentile rank cutoff of 2 was used for weak binders and 0.5 for strong binders [28].

Data Availability

The MS datasets are available via the PRIDE partner repository with the identifier PXD043844 (non-tryptic timsTOF dataset), PXD019086 (tryptic timsTOF dataset; reanalysis available on MSV000092462), PXD038782 (comparison dataset; reanalysis available on MSV000092461), and PXD040740 (SCP dataset; reanalysis available on RMSV000000693.1). All protein databases used in this study are deposited alongside the result files.

Code Availability

Source code and scripts are available on GitHub at https://github.com/wilhelm-lab/koina, https://github.com/wilhelm-lab/oktoberfest, and https://github.com/adamscharlotte/timsTOF-immunopeptide-prediction.

Competing interests

K.L. holds shares in ImmuneWatch BV: an immunoinformatics company. M.W. is a co-founder and shareholder of MSAID GmbH. M.W. is also a co-founder and shareholder of OmicScouts GmbH, he has no operational role in both companies. K.B. is a co-founder and shareholder of ImmuneSpec BV.

References

1. Vaughan, K., Xu, X., Caron, E., Peters, B. & Sette, A. Deciphering the MHC-associated peptidome: a review of naturally processed ligand data. *Expert Rev. Proteomics* **14**, 729–736 (2017).
2. Leddy, O. K., White, F. M. & Bryson, B. D. Leveraging Immunopeptidomics To Study and Combat Infectious Disease. *mSystems* **6**, e00310-21 (2021).
3. Peltonen, K. *et al.* Therapeutic Cancer Vaccination with Immunopeptidomics-Discovered Antigens Confers Protective Antitumor Efficacy. *Cancers* **13**, 3408 (2021).
4. Prinz, J. C. Immunogenic self-peptides - the great unknowns in autoimmunity: Identifying T-cell epitopes driving the autoimmune response in autoimmune diseases. *Front. Immunol.* **13**, (2023).
5. Mayer, R. L. & Impens, F. Immunopeptidomics for next-generation bacterial vaccine development. *Trends Microbiol.* **29**, 1034–1045 (2021).
6. George, A. J. T., Stark, J. & Chan, C. Understanding specificity and sensitivity of T-cell recognition. *Trends Immunol.* **26**, 653–659 (2005).
7. Nesvizhskii, A. I. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J. Proteomics* **73**, 2092–2123 (2010).
8. Ouspenskaia, T. *et al.* Unannotated proteins expand the MHC-I-restricted immunopeptidome in cancer. *Nat. Biotechnol.* **40**, 209–217 (2022).
9. Käll, L., Canterbury, J. D., Weston, J., Noble, W. S. & MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **4**, 923–925 (2007).
10. Declercq, A. *et al.* MS2Rescore: Data-Driven Rescoring Dramatically Boosts Immunopeptide Identification Rates. *Mol. Cell. Proteomics* **21**, 100266 (2022).
11. Wilhelm, M. *et al.* Deep learning boosts sensitivity of mass spectrometry-based

immunopeptidomics. *Nat. Commun.* **12**, 3346 (2021).

12. Yang, K. L. *et al.* MSBooster: Improving Peptide Identification Rates using Deep Learning-Based Features. 2022.10.19.512904 Preprint at https://doi.org/10.1101/2022.10.19.512904 (2022).
13. Meier, F. *et al.* Online Parallel Accumulation–Serial Fragmentation (PASEF) with a Novel Trapped Ion Mobility Mass Spectrometer. *Mol. Cell. Proteomics MCP* **17**, 2534–2545 (2018).
14. Gravel, N. H. *et al. timsTOF mass spectrometry-based immunopeptidomics refines tumor antigen identification*. https://www.researchsquare.com/article/rs-2402111/v1 (2023) doi:10.21203/rs.3.rs-2402111/v1.
15. Gomez-Zepeda, D. *et al. Thunder-DDA-PASEF enables high-coverage immunopeptidomics and identifies HLA class-I presented SarsCov-2 spike protein epitopes*. (2023). doi:10.21203/rs.3.rs-2625909/v1.
16. Orsburn, B. C. Time-of-Flight Fragmentation Spectra Generated by the Proteomic Analysis of Single Human Cells Do Not Exhibit Atypical Fragmentation Patterns. *J. Proteome Res.* (2023) doi:10.1021/acs.jproteome.2c00715.
17. Van Puyvelde, B. *et al.* A comprehensive LFQ benchmark dataset on modern day acquisition strategies in proteomics. *Sci. Data* **9**, 126 (2022).
18. Zolg, D. P. *et al.* Building ProteomeTools based on a complete synthetic human proteome. *Nat. Methods* **14**, 259–262 (2017).
19. Zolg, D. P. *et al.* ProteomeTools: Systematic Characterization of 21 Post-translational Protein Modifications by Liquid Chromatography Tandem Mass Spectrometry (LC-MS/MS) Using Synthetic Peptides*. *Mol. Cell. Proteomics* **17**, 1850–1863 (2018).
20. Gessulat, S. *et al.* Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat. Methods* **16**, 509–518 (2019).
21. Meier, F. *et al.* Deep learning the collisional cross sections of the peptide universe from a

million experimental values. *Nat. Commun.* **12**, 1185 (2021).

22. Bittremieux, W. spectrum_utils: A Python Package for Mass Spectrometry Data Processing and Visualization. *Anal. Chem.* **92**, 659–661 (2020).
23. Phulphagar, K. M. *et al.* Sensitive, High-Throughput HLA-I and HLA-II Immunopeptidomics Using Parallel Accumulation-Serial Fragmentation Mass Spectrometry. *Mol. Cell. Proteomics* **22**, (2023).
24. Andreatta, M., Alvarez, B. & Nielsen, M. GibbsCluster: unsupervised clustering and alignment of peptide sequences. *Nucleic Acids Res.* **45**, W458–W463 (2017).
25. Chong, C. *et al.* Integrated proteogenomic deep sequencing and analytics accurately identify non-canonical peptides in tumor immunopeptidomes. *Nat. Commun.* **11**, 1293 (2020).
26. Erhard, F., Dölken, L., Schilling, B. & Schlosser, A. Identification of the Cryptic HLA-I Immunopeptidome. *Cancer Immunol. Res.* **8**, 1018–1026 (2020).
27. Sarkizova, S. *et al.* A large peptidome dataset improves HLA class I epitope prediction across most of the human population. *Nat. Biotechnol.* **38**, 199–209 (2020).
28. Reynisson, B., Alvarez, B., Paul, S., Peters, B. & Nielsen, M. 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 Res.* **48**, W449–W454 (2020).
29. Sinitcyn, P. *et al.* Global detection of human variants and isoforms by deep proteome sequencing. *Nat. Biotechnol.* 1–11 (2023) doi:10.1038/s41587-023-01714-x.
30. Nagy, K., Gellén, G., Papp, D., Schlosser, G. & Révész, Á. Optimum collision energies for proteomics: The impact of ion mobility separation. *J. Mass Spectrom.* **n/a**, e4957 (2023).
31. Searle, B. C. *et al.* Generating high quality libraries for DIA MS with empirically corrected peptide predictions. *Nat. Commun.* **11**, 1548 (2020).
32. Zolg, D. P. *et al.* PROCAL: A Set of 40 Peptide Standards for Retention Time Indexing,

Column Performance Monitoring, and Collision Energy Calibration. *Proteomics* **17**, (2017).

33. Meier, F. *et al.* Parallel Accumulation–Serial Fragmentation (PASEF): Multiplying Sequencing Speed and Sensitivity by Synchronized Scans in a Trapped Ion Mobility Device. *J. Proteome Res.* **14**, 5378–5387 (2015).
34. Tyanova, S., Temu, T. & Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc.* **11**, 2301–2319 (2016).
35. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
36. Łącki, M. K., Startek, M. P., Brehmer, S., Distler, U. & Tenzer, S. OpenTIMS, TimsPy, and TimsR: Open and Easy Access to timsTOF Raw Data. *J. Proteome Res.* **20**, 2122–2129 (2021).
37. Tobias. kusterlab/MasterSpectrum: v1.1. (2020) doi:10.5281/zenodo.3706673.
38. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
39. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. Preprint at https://doi.org/10.48550/arXiv.1412.6980 (2017).
40. Smith, L. N. Cyclical Learning Rates for Training Neural Networks. *2017 IEEE Winter Conf. Appl. Comput. Vis. WACV* 464–472 (2017) doi:10.1109/WACV.2017.58.
41. The, M., MacCoss, M. J., Noble, W. S. & Käll, L. Fast and Accurate Protein False Discovery Rates on Large-Scale Proteomics Data Sets with Percolator 3.0. *J. Am. Soc. Mass Spectrom.* **27**, 1719–1727 (2016).
42. Thomsen, M. C. F. & Nielsen, M. Seq2Logo: a method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion. *Nucleic Acids Res.* **40**, W281–W287 (2012).

43. Tareen, A. & Kinney, J. B. Logomaker: beautiful sequence logos in Python. *Bioinformatics* **36**, 2272–2274 (2020).