Journal Pre-proof

inSPIRE: An open-source tool for increased mass spectrometry identification rates using Prosit spectral prediction

John A. Cormican, Yehor Horokhovskyi, Wai Tuck Soh, Michele Mishto, Juliane Liepe

PII: S1535-9476(22)00240-7

DOI: https://doi.org/10.1016/j.mcpro.2022.100432

Reference: MCPRO 100432

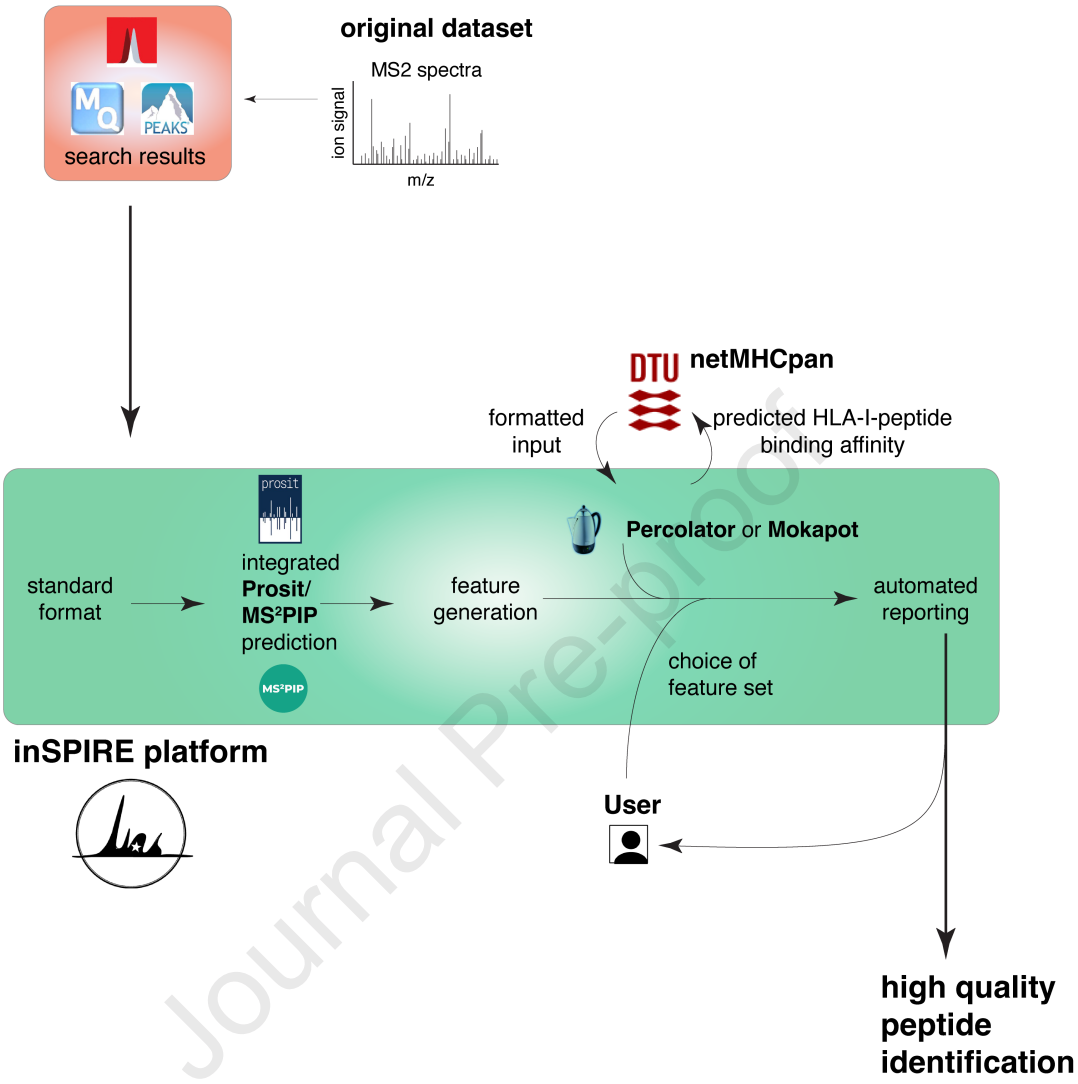
To appear in: *Molecular & Cellular Proteomics*

Received Date: 26 May 2022

Revised Date: 17 October 2022

Accepted Date: 19 October 2022

Please cite this article as: Cormican JA, Horokhovskyi Y, Soh WT, Mishto M, Liepe J, inSPIRE: An open-source tool for increased mass spectrometry identification rates using Prosit spectral prediction, *Molecular & Cellular Proteomics* (2022), doi: https://doi.org/10.1016/j.mcpro.2022.100432.



inSPIRE: An open-source tool for increased mass spectrometry identification rates using Prosit spectral prediction

John A. Cormican[1], Yehor Horokhovskyi[1], Wai Tuck Soh[1], Michele Mishto[2,3,$], Juliane Liepe[1,$]

[1] Max-Planck-Institute for Multidisciplinary Sciences (MPI-NAT), 37077 Göttingen, Germany
[2] Centre for Inflammation Biology and Cancer Immunology (CIBCI) & Peter Gorer Department of Immunobiology, King's College London, SE1 1UL London, United Kingdom
[3] The Francis Crick Institute, WC2A 3LY London, United Kingdom

[$] Correspondence to: michele.mishto@kcl.ac.uk, jliepe@mpinat.mpg.de.

Keywords
mass spectrometry, rescoring, Prosit, Percolator,

Abstract
Rescoring of mass spectrometry (MS) search results using spectral predictors can strongly increase Peptide Spectrum Match (PSM) identification rates. This approach is particularly effective when aiming to search MS data against large databases, for example when dealing with non-specific cleavage in immunopeptidomics or inflation of the reference database for noncanonical peptide identification. Here, we present inSPIRE (*in silico* Spectral Predictor Informed REscoring), a flexible and performant open-source rescoring pipeline built on Prosit MS spectral prediction, which is compatible with common database search engines. inSPIRE allows large scale rescoring with data from multiple MS search files, increases sensitivity to minor differences in amino acid residue position, and can be applied to various MS sample types, including tryptic proteome digestions and immunopeptidomes. inSPIRE boosts PSM identification rates in immunopeptidomics, leading to better performance than the original Prosit rescoring pipeline, as confirmed by benchmarking of inSPIRE performance on ground truth datasets. The integration of various features in the inSPIRE backbone further boosts the PSM identification in immunopeptidomics, with a potential benefit for the identification of noncanonical peptides.

Introduction

Tandem mass spectrometry (MS) has been a successful tool for large scale identification in proteomics and peptidomics (1-3). In tandem MS, peptides in a sample are first ionized and separated by their mass-to-charge ratio (m/z), resulting in MS1 spectra. Selected ions are then fragmented, often by collision-induced dissociation (CID) or high collision-induced dissociation (HCD), resulting in MS2 spectra. Commonly, peptide identifications are performed via database search engines comparing the experimentally observed MS2 spectra to the theoretical fragments produced by all possible peptides in a reference proteome (*i.e*., the search space) (4). The highest scoring match between theoretical and experimental spectra is then assigned to produce a peptide spectrum match (PSM). In order to quantify the probability that a PSM is correct, it is of importance to compute the false discovery rate (FDR). This is commonly estimated by searching a decoy database of reversed or randomized sequences, with a similar composition to the reference (or target) database (5). The decoy database should contain no true peptide sequences present in the analyzed sample, and so the number of false discoveries above a given scoring threshold can be estimated by the number of PSMs from the decoy database with scores above that threshold.

This approach is adopted by many traditional search engines such as Mascot, MaxQuant or PEAKS DB, and it has been successful, particularly when dealing with a small well informed reference proteome (6). For example, in proteomics experiments, the digestion of a protein containing sample with a specific protease such as trypsin results in a reduced search space compared to the digestion of the same sample with an unspecific protease. However, it is not always possible to achieve this reduced search space. One such case is the field of immunopeptidomics, where MS-mediated identifications of peptides presented by Human Leucocyte Antigen class I and II (HLA-I and -II) complexes can provide valuable insights into specific immune responses and potentially aid the development of targeted immunotherapies (7). These immunopeptides are generated within the cell via various processing steps, including proteasomal processing (8). Proteasomes are proteases that, in contrast to trypsin, can cleave after any amino acid, following complex dynamics (9, 10), thereby impacting on the database search space. The size of the search space can be expanded even further with increasing interest in noncanonical peptides outside of commonly used reference proteomes (11-14). These enlarged search spaces can create a number of issues to traditional target-decoy search approaches, with a significant negative impact on both peptide yield and FDR estimation (15, 16).

Post-processing or rescoring approaches, which perform additional validation on the target and decoy PSMs outputted from the original database search, have been developed to achieve high peptide yield at low FDR estimates, even when confronted with enlarged search spaces. These algorithms have been used for many years as a method of validating peptide identifications and increasing identification rates in MS search results (17-19). Percolator is one such algorithm that uses a semi-supervised machine learning approach, which considers features beyond the original search engine score using Support Vector Machine (SVM) models, and it has become the dominant approach for MS2 post-processing (20). Percolator makes use of a subset of high confidence target PSMs as positive samples and all decoy PSMs as negative samples for training its model. The trained model then provides a better separation between target and decoy peptides, allowing a larger number of peptides to be identified at a similar FDR. Throughout this process, Percolator employs a cross-validation mechanism to avoid overfitting (21).

Percolator is highly flexible and allows its user to consider any set of features to describe a dataset of PSMs. Multiple researchers have taken advantage of this by integrating features from newly developed predictors to increase the number of high confidence peptides identified. Such applications include the use of retention time predictors (22), and the use of predictors of HLA-peptide binding affinities in immunopeptidomics (23). One particularly fruitful approach has been the use of spectral predictors, as proposed by C. Silva *et al*. (24).

Accurate prediction of the MS2 spectrum of a peptide has been an active area of research for a number of years (25). While classic database search engines considered only the presence or absence of possible MS2 fragment ions, *i.e*., y- or b- ions for HCD, in an experimental MS2 spectrum, modern spectral predictors can accurately predict the relative intensities of the fragment ions (26-29). One of the most significant accomplishments in this field is the Prosit spectral predictor (28, 29). Prosit is a deep learning tool that was trained on more than 20 million high quality experimental MS2 spectra, and that has demonstrated state of the art performance on both tryptic and immunopeptidomics datasets.

Wilhelm *et al*. (29) combined metrics describing the match between the Prosit predicted spectrum and the experimentally observed spectrum as Percolator input features to increase discovery rates on immunopeptidome search spaces. Significant increases in performance have also been demonstrated for the MS²Rescore software, which uses predictions from the MS²PIP spectral predictor rather than from Prosit (30). Both tools have also been compared when applied to a tryptic digest with an enlarged search space, showing a similar performance (31).

Despite the clear potential of using Prosit predicted spectra in rescoring pipelines, its use is limited by the fact that no fully open-source pipeline is available. The Prosit rescoring pipeline presented by Wilhelm *et al*. (29) and recently updated (32), is only available via a web server, which only allows search results from a single MS RAW file and only processes search results from the MaxQuant search engine. Reproducibility is also limited with this system as no versioning information is available. The alternative INFERYS pipeline removes some of these technical limitations but is only available as part of the commercial Thermo Fisher Proteome Discoverer software (33).

To address these gaps we developed inSPIRE, which stands for *in silico* Spectral Predictor Informed Rescoring. inSPIRE is a flexible and performant open-source rescoring pipeline, primarily built on Prosit retention time and MS spectral prediction. In contrast to the Prosit rescoring pipeline, inSPIRE is compatible with multiple major database search engines and allows large scale rescoring with data from multiple search files. Furthermore, the inSPIRE pipeline can perform spectral prediction with Prosit on standard CPU hardware, whereas the original Prosit release required a specialized GPU. For added flexibility, inSPIRE can also use MS^{2}PIP as the spectral predictor instead of Prosit and Pyteomics for retention time prediction (34, 35), though this was not our primary focus given that the MS^{2}Rescore pipeline already provides fully open-source rescoring using MS^{2}PIP predictions. inSPIRE can be applied to various sample types including tryptic proteome digestions and immunopeptidomes. It is specifically optimized for enlarged immunopeptidome search spaces with increased sensitivity to minor differences in amino acid residue position.

In addition, we developed an inSPIRE variant, specifically for HLA-I immunopeptidomics – *i.e*., inSPIRE-affinity – that allows the integration of NetMHCpan predictions to the inSPIRE backbone. NetMHCpan also uses a deep learning framework, in this case, to predict the binding affinity of a given peptide to given HLA molecules (36-38).

Experimental Procedures

Cell lines

K562-B*07:02 and -A*02:01 cell line clones express either the single HLA-B*07:02 or -A*02:01 alleles. They derive from the leukemia K562 cell line (ATCC®CCL-243™), which does not express endogenous HLA-I and -II molecules, and its generation and growing conditions are described elsewhere (13).

HLA-I immunopeptidome elution and tryptic proteome digestion

HLA-I-bound peptides were isolated from 10^9 cells of K562-B*07:02 and -A*02:01 cell line clones, through HLA-I-peptide elution using W6/32 antibody, as described elsewhere (13). The MS files were already published in (13).

The HLA-I immunopeptidomes used as training datasets were previously published by Paes *et al*. (39) and Bassani-Sternberg *et al.* (40).

Tryptic digestions of cell proteome obtained from the K562 cell line were carried out as follows: cell pellet was lysed in cell lysis buffer (50 mM HEPES, pH 7.5, 150 mM NaCl, 4% SDS, 2 mM DTT, 0.5% NP40) and heated at 95°C for 10 min. The cell lysate was then diluted to a final concentration of 1% SDS with 50 mM HEPES, pH 7.5. Pierce™ Universal nuclease (ThermoFisherScientific) was added according to the manufacturer's recommendations and incubated at 37°C for 30 min under shaking condition (300 rpm). Protein concentration was determined using Pierce™ BCA protein assay kit (ThermoFisherScientific) and 50 μg of protein was used for proteome digestion. Proteins were reduced with 5 mM DTT for 30 min at 37°C and alkylated by the addition of 20 mM iodoacetamide and incubation for 30 min at room temperature in the dark. The reaction was quenched by incubation with 20 mM DTT

for 15 min at room temperature before purification with SP3 beads (41), and elution for proteome digestion with trypsin (Promega) at protease to proteome weight ratio of 1:25 at 37°C for 16 hours.

Synthetic peptide library

The synthetic peptide library contained 9, 10, or 15 amino acid long peptides (n = 6,876 unique peptide sequences and 13,868 PSMs) related to $CD4^+$ and $CD8^+$ T cell response to Dengue and VZV viruses, as described elsewhere (42). The Dengue and VZV synthetic peptides utilized in this study were selected for analysis because they were already available in-house and synthesized for separate epitope identification studies. The selection and characterization of these peptides has been described previously (43-50). Each of the peptides in synthetic peptide libraries was derived from respective Dengue and VZV proteomes. Peptides were originally selected for other studies based on bioinformatic analyses of predicted capacity to bind various common HLA-I and -II alleles in the general worldwide population. The set of Dengue protein sequences of provenance represents all four Dengue serotypes and several different variant isolates. The VZV peptides were primarily derived from the attenuated varicella vaccine strain vOka and a few variant isolates. Peptides were grouped in 8 library batches, with each peptide measured at the concentration of 0.0625 pmol/μl. For each pool, 8 μl was injected in the instrument, thereby measuring 500 fmol of each peptide. The synthetic peptide libraries are reported in **File S5**.

Mass spectrometry

MS data of HLA-I immunopeptidomes were collected using Orbitrap Fusion Lumos mass spectrometer coupled to an Ultimate 3000 RSLC nano pump (both from ThermoFisherScientific), as described elsewhere (13). The same method and instrument were used for the synthetic peptide library measurement. MS data of tryptic digestions of cell proteome were measured through Thermo Scientific Orbitrap Exploris™ 480 mass spectrometer. Digested proteome samples were injected using an Ultimate 3,000 RSLC nano pump (both from ThermoFisherScientific). Briefly, 0.5 μg of each sample was loaded and separated by a nanoflow HPLC (RSLC Ultimate 3000) on an Easy-spray C18 nano column (30 cm length, 75 μm internal diameter). Peptides were eluted with a linear gradient of 5% – 45% buffer B (80% ACN, 0.1% formic acid) at a flow rate of 300 nl/min over 58 min at 50°C. The instrument was programmed within Xcalibur 3.1.66.10 to acquire MS data in a Data Dependent Acquisition mode using Top 30 precursor ions. We acquired one full-scan MS spectrum at a resolution of 60,000 with a normalized automatic gain control (AGC) target value of 300% and a scan range of 350-1,600 m/z. The MS/MS fragmentation was conducted using HCD collision energy (28%) with an orbitrap resolution of 15,000. The normalized AGC target value was set up at 100% with a max injection time of 40 ms. A dynamic exclusion of 22s and 2 - 6 included charged states were defined within this method.
The MS files used in each figure are reported in **Table S4**.

MS software settings

For all sections where MaxQuant was used, RAW MS files were searched using MaxQuant GUI version 1.6.17. First search peptide tolerance was set to 20 ppm and the main search peptide tolerance to 4.5 ppm. Minimum peptide length was set to 7 and maximum peptide mass to 4,600 Da. The mass tolerance for the fragment ions was set to 20ppm. For identification both PSM FDR and Protein FDR were set to 1.0, allowing the maximum possible number of PSMs to be exported.
For the tryptic searches, we performed a specific search against the reference proteome with enzyme set to trypsin, allowing cleavage after proline. Up to 2 missed cleavages were allowed. Oxidation of methionine was set as the only variable post-translational modification (PTM) and carbamidomethylation of cysteine was set as a fixed modification. For the immunopeptidome searches we performed an unspecific search against the reference proteome. Oxidation of methionine was set as the only variable PTM and no fixed PTMs were set. Before rescoring with inSPIRE or Prosit-rescoring, all hits containing cysteine were removed as Prosit assumes carbamidomethylation of cysteine. For the ground truth dataset, a non-specific search was used. In this case, no modifications were selected and again the hits containing unmodified cysteine residues were removed before rescoring with inSPIRE or Prosit-rescoring.

For the identification of the synthetic peptides used in the synthetic peptide library, we searched RAW files using PEAKS version 10.6 with precursor mass tolerance of 5 ppm and fragment ion mass tolerance of 0.02 Da. No PTMs were allowed and results were exported at FDR of 1%.
For the comparison between rescoring on different search engines, we searched RAW files using PEAKS version 10.6 with precursor mass tolerance of 5 ppm and fragment ion mass tolerance of 0.02 Da. As with the MaxQuant searches, the tryptic searches were performed with 2 missed cleavages allowed and cleavage was allowed after proline. The same PTM settings were used and again all hits containing cysteine were filtered out before rescoring. Results were exported for all PSMs with PEAKS -10lgP score greater than 0.
In the case of Mascot, we used Mascot Distiller version 2.8.0.1 to process the MS RAW files. To allow the detection of chimeric spectra with Mascot we set the Maximum number of precursor m/z values to 2 and set *Allow multiple precursors per scan* to true. Precursor mass tolerance was set to 5 ppm and fragment ion mass tolerance was set to 0.02 Da for both tryptic and immunopeptidome searches. The tryptic searches were performed with 2 missed cleavages allowed and cleavage was allowed after proline. The same PTMs were allowed as with PEAKS and MaxQuant and hits containing unmodified cysteine residues were filtered out before rescoring. We used Mascot's automatic decoy search and exported both target and decoy results with a homology significance threshold set to 0.999999 (*i.e.*, exporting essentially all hits).
In order to provide the experimental spectra to the inSPIRE pipeline, RAW files were converted to mgf format using the ms-convert GUI. The ThermoRawFileParser version 1.4.0 was used to generate data for mgf input in Prosit-*delta* training pipeline (51).
Percolator version 3.0.5 was for all rescoring jobs. All Prosit-rescoring jobs were submitted to the web server between March 10th and August 16th 2022. Rescoring with MS^{2}Rescore was performed with version 2.1.2.
The search result files used for each figure are reported in **Table S5**. The final identifications for all pipelines for all datasets are provided in **File S6-S9**.

Application of MS^{2}Rescore
For both tryptic and immunopeptidome datasets the general settings for MS^{2}Rescore were set with "pipeline" to "infer", "feature_sets" to a list of "searchengine", "ms2pip", and "rt", "run_percolator" to false, "id_decoy_patter" to null, "num_cpu" to -1, "config_file" to null, "tmp_path" to null, "mgf_path" to null, "output_filename" to null, "log_level" to info and "plotting" to false. The "ms2pip" settings were set with "model" as "Immuno-HCD" for the immunopeptidome datasets and "HCD2021" for the tryptic datasets. The "frag_error" was set to 0.02. Variable modification of oxidation of methionine was set in either case with the "modification_mapping" set with "Oxidation (M)" mapping to "Oxidation" for both datasets. In the case of the tryptic proteome digestion, "fixed_modifications" was also set with "C" mapping to "Carbamidomethyl".

Application of Percolator
We reran Percolator for all pipelines due to the use of --subset-max-train command line argument in the Prosit rescoring pipeline. This command line argument can lead to a breakdown of the Percolator cross-validation algorithm and should not be applied on small datasets according to The *et al.* (20), as confirmed by the Percolator team via GitHub (personal communication). We have communicated this issue to the Prosit team via GitHub. In reapplying Percolator with the same command line arguments to all pipelines, we ensured that the only variation in the PSMs identified by a rescoring pipeline was not due to different applications of Percolator.
MS^{2}Rescore allows the user to select the command line arguments passed to Percolator but for convenience we simply reran Percolator on the .pin file produced by MS2Rescore via terminal with the same command line arguments used for inSPIRE and Prosit rescoring.

RNA sequencing and reference databases
The K562 RNA was extracted from K562 cell line pellets, processed for polyA enrichment and sequenced by using NEBNext Ultra RNA Library Preparation Kit with random priming. Sequencing was performed using HiSeq 2x150 PE HO with a depth of 20-25 million reads per sample. Details about reads trimming, quantification and data processing are described elsewhere (13).

The RNA sequencing dataset generated Paes *et al.* (39) was generated as described in the original paper and is available upon request to the authors.
RNA-informed reference databases were generated by imposing an expression cutoff of 10 estimated counts per transcript in Gencode transcriptome main annotation Release 33 (GRCh38.p13) (52). Protein-coding transcript translation sequences from these transcripts were kept in an RNA-informed reference database.
The Gencode transcriptome main annotation Release 33 (GRCh38.p13) (52) was searched alongside the RNA-informed reference database so that performance across different database size could be compared.
The Uniprot *Homo Sapiens* proteome reference database used for PEAKS DB searches to generate PSMs for the Prosit-*delta* training data was downloaded on the 14th of July 2022.

HLA-I-peptide binding affinity prediction
HLA-I-peptide binding affinity was predicted by applying NetMHCpan 4.1. Specifically, we used a custom docker image. The NetMHCpan input file is provided as part of the inSPIRE "prepare" pipeline, provided that "useBindingAffinity" setting in the configuration file is specified as "asValidation" or "asFeature". When using binding affinity predictions as a validation (*i.e*., comparing number of predicted HLA-I-peptide binders for inSPIRE compared to Prosit-rescoring) we only considered NetMHCpan predictions for peptides with length between 8 to 14 residues due to software limitations. For inSPIRE-affinity, we generated predictions for all peptides as null values were not allowed in the Percolator input file.
For our validation and reporting pipelines, we defined a peptide predicted by NetMHCpan to bind a given HLA-I complex, by evaluating against the %Rank value, according to Reynisson *et al*. (37). The %Rank is a transformation on the original prediction, allowing comparison across HLA-I-peptide binding specificities. This system defined a 'strong HLA-I binder' as a peptide with a %Rank < 0.5% for a given HLA-I allele, and a 'HLA-I binder' as a peptide with a %Rank < 2% for a given HLA-I allele.
We also used the values for the Positive Predictive Value (PPV) reported by Reynisson and colleagues (37) as a metric to understand the variation between NetMHCpan performance on different alleles. Reynisson and colleagues (37) defined PPV as the number of positive binding peptides correctly predicted divided by 0.95 times the number of ligands predicted. By considering this metric for the different alleles analyzed, we could study how the strength of the NetMHCpan predictor for an HLA-I allele impacted the use of predicted HLA-I-peptide binding affinity both as an evaluation metric and as a feature for rescoring.

Experimental Design and Statistical Rationale
This study aimed to benchmark inSPIRE performance against other state of the art tools, in particular the Prosit rescoring pipeline, to demonstrate its value on datasets that the original Prosit rescoring pipeline could not allow rescoring, and to demonstrate the value of our novel Prosit-*delta* predictor.
In benchmarking we focused our analysis on HLA-I immunopeptidome datasets of the K562-A*02:01 and K562-B*07:02 cell lines, for which we had an RNA-informed dataset and for which the NetMHCpan predictor performs strongly (37).
Since the Prosit web server only allowed the analysis of a single raw file, and given that all of the immunopeptidomics dataset came from previously published studies, we generally did not favor running multiple replicates of the same allele. This allowed us to explore a wider variety of HLA alleles with differing motifs rather than focusing on many replicates of a limited variety.
For **Fig. 2**, the MaxQuant search results of the K562-A*02:01 derived immunopeptidome datasets searched with the RNA-informed and Gencode reference databases contained 14,689 and 15,065 PSMs respectively. The equivalent datasets for the K562-B*07:02 derived immunopeptidome contained 14,738 and 14,741 PSMs, respectively, and for the tryptic proteome digestion they contained 55,396 and 56,560 PSMs, respectively.
For **Fig. 3**, in all cases the total number of PSMs used to generate the figures was 12,924 PSMs.
For **Fig. 4**, the immunopeptidome rescoring for RNA-informed and Gencode reference database searches was based, respectively, on 41,188 and 40,929 PSMs for MaxQuant, 29,802 and 30,833 PSMs for Mascot, and 22,958 and 22,504 PSMs for PEAKS DB. The tryptic proteome digestion rescoring using the same reference databases was based, respectively, on 339,609 and 339,470 PSMs

for MaxQuant, 244,697 and 244,698 PSMs for Mascot, and 316,095 and 310,219 PSMs for PEAKS DB.
The p-values relevant to **Fig. S10** were calculated using Student's t-test.
The R^2 values in **Fig. 5E,F** were based on 128,087 and 253,478 Prosit-*delta* values, respectively.
All analysis has been implemented in Python, if not stated otherwise. All statistics for performance measurement are described in the benchmarking framework.

Metrics validating rescoring performance
Our simplest analysis of the performance of an identification method was to compare the number of PSMs identified at 1% FDR as estimated via Percolator, which was used as the final identification method for all rescoring pipelines presented.
In an attempt to ensure that all pipelines were applying FDR estimation fairly, we used two independent validations for our K562, K562-A*02:01, K562-B*07:02 cell line datasets. Firstly, for HLA-I immunopeptidome data, if NetMHCpan was not used in rescoring, we investigated the percentage of HLA-I binders and strong binders predicted by NetMHCpan among the peptides identified. Secondly, when rescoring search results obtained using the Gencode reference database, we investigated the percentage of peptides identified, which were also found by the search engine at any confidence level when searching the RNA-informed reference database.
We acknowledge that neither of these validation techniques was perfect; it is possible that a correct peptide sequence was not predicted to be an HLA-I binder by NetMHCpan or that the RNA sequencing evidence was not sufficient for its substrate protein to be included in our RNA-informed database. Equally, it is possible that an incorrectly identified peptide was predicted as a strong HLA-I binder and was also found in the RNA-informed database.
However, we estimated that reasonable consistency between these metrics across identifications from different pipelines, combined with Percolator's well established FDR estimation method (21) was a sufficient validation that an increased number of PSMs identified at a given threshold did represent better identification performance, indeed.

Benchmarking with the synthetic peptide library as ground truth dataset
As a validation of the increased number of PSMs for the inSPIRE pipeline we benchmarked all rescoring methods using 'ground truth' datasets, along the line of the benchmarking tool iBench (53). In the approach applied in this study, we measured synthetic peptides via MS, and selected MS2 scans that were identified with 1% FDR using PEAKS search engine. These peptides and their PSMs formed our 'ground truth' dataset, although we note that our 'ground truth' datasets represented an approximation to an absolute ground truth. Indeed, this strategy still had a minor degree of imperfection since it was based on MS measurement with 1% FDR rather than 0% FDR. Furthermore, although this strategy could contain a certain level of bias toward the PEAKS search engine, we estimated that this did not introduce any advantage for the rescoring methods used and the use of synthetic peptide libraries greatly reduced the risk of identification errors.
We then embedded two thirds of those synthetic peptide sequences into the Gencode reference database, thereby generating a constructed reference database, similarly as in (13). In this constructed reference database, these peptide sequences were labelled as the "discoverable". We ensured that the remaining one third of the synthetic peptide sequences were not in the constructed reference database, and, hence, were 'undiscoverable'. We also added fragments of these peptide sequences to the constructed reference database so that the composition of the database was not biased against these peptides after their removal. We then searched the RAW files with MaxQuant using the constructed database, and extracted all PSMs (FDR = 1.0%). Any identification found in the MaxQuant search result for an MS2 scan which was not identified by PEAKS in the original search of the synthetic peptides was filtered out before any rescoring was applied. This removed the possible confounding influence of contamination peptides of unknown origin.
In our study, we were initially limited by the fact that there were approximately 1,000 peptides identified per RAW file and the Prosit rescoring pipeline only allowed a single RAW file to be scored against. Hence, to overcome these potential limitations, we ran Prosit rescoring for each of the 8 RAW files of the synthetic peptide libraries (*i.e*. SPL1-2 to SPL8-2) separately. We then concatenated the *prosit.tab* files from each run, which contained all of the input for the final Percolator rescoring, and reran

Percolator with the concatenated files. We used the –override flag to ensure that Percolator used the full feature set for all executions.

In order to quantify the impact of the small dataset size on each pipeline, we performed rescoring on search results from 2, 4, and 8 RAW files, and calculated precision-recall (PR) curves for each method. To remove the effects of differences between RAW files, we ran all rescoring pipelines on 4 combinations of 2 RAW files and 2 combinations of 4 RAW files, so that in each case the final performance was measured on the same data.

The PR curves were generated by varying the cut off between the minimum and maximum Percolator score for each rescoring method. This involved combining Percolator scores across different runs. To note, Percolator normalized scores based on q-value and combined scores internally from the different cross-validated models. Hence, combining scores across model did not create any clear bias between the methods being benchmarked.

The precision at each cut off was calculated as the number of correctly identified PSMs divided by the total number of PSMs above the threshold, while the recall was calculated as the number of correctly identified PSMs divided by the number of discoverable peptides in the modified database. In each case the maximum possible recall was limited by the number of correct PSMs found in the original search engine results.

Development of inSPIRE Prosit-delta predictor

The motivation for the Prosit-*delta* is explained in detail in the results section. Briefly, we aimed to use a lightweight predictor to estimate the sensitivity of Prosit to adjacent residue permutation at each fragmentation site of the peptide. Although inSPIRE does allow for "brute force" computation of all Prosit predicted MS2 spectra and resulting Prosit-*delta* values this would result in doubling the run time and vastly increasing the memory consumption. As an alternative, we found that an xgboost Gradient Tree Boosting Regressor provided an appropriate and performant solution (54), without incurring the same computational burden.

We developed the *delta* predictor for Prosit only and not MS^{2}PIP for a number of reasons. For instance, inSPIRE was primarily developed to increase the availability of Prosit predictions. Also, one of the main reasons why an inSPIRE user would choose the inSPIRE-MS^{2}PIP pipeline could be to predict MS2 spectra for peptides containing PTMs not available with Prosit. Including a broad range of PTMs - particularly those linked to the termini of a peptide - would significantly complicate the current version of the *delta* predictor. In addition, spectral angle was not the primary metric on which MS^{2}PIP was trained, although it was an important feature in the inSPIRE-MS^{2}PIP pipeline. Hence, the development of *delta* scoring within MS^{2}PIP might need a specific investigation of the best metric to be considered.

In order to generate training data for the Prosit-*delta* predictor, we searched the HLA-I immunopeptidomes of Paes *et al*. (39) and Bassani-Sternberg et al. (40) using PEAKS 10.6, using an RNA-informed database and the Uniprot Homo Sapiens database, respectively, and exported all hits with PEAKS -10lgP greater than 0. This data was combined with synthetic immunopeptides used by Wilhelm *et al*. (29) for which we used the MaxQuant identifications provided in their Pride repository. A full description of the RAW files used and the number of PSMs is provided in **Table S2**. The PEAKS DB searches were run with oxidation of methionine and carbamidomethylation of cysteine set as variable modifications. All hits containing unmodified cysteines were discarded before training. The PSMs were then divided between train (80% of the data) and test (20% of the data) ensuring that there was no overlap in the peptides used between train and test.

While we collected data for peptides with lengths 7-30 and precursor charge 1-6, the vast majority of our training comes from peptides of length less than 13 and precursor charge 1-3 (**Fig. S11A,B**). This feature was in agreement with one of the main objectives of the Prosit-*delta* predictor, which was its application to immunopeptidome datasets. We also show sequence logo plots for the peptides of length 8-11 residues in the combined dataset in **Fig S11**(**C-F**), thereby illustrating that the dataset was not biased towards any specific motif.

For each PSM, we selected 5 positions at random in the peptide sequence and generated Prosit predictions of MS2 spectra for the peptide created by flipping the adjacent amino acids at those positions. The target variable was the difference between the spectral angle of the modified sequence and the spectral angle of the original sequence. Hence, each PSM in the training dataset generated 5 training data points. The features used as input for the Prosit-*delta* predictor are detailed in **Table S3**.

We performed hyperparameter tuning on the parameters minimum child weight, maximum tree depth, learning rate, gamma and columns sampled by tree. We then used randomized search with 5-fold cross-validation on the training set and compared performance for different sets of hyperparameters based both on predictive performance (r^2 score) and speed of execution. The results of this first round of hyperparameter tuning are shown in **File S10**. We then selected 5 sets of hyperparameters, which performed well in cross fold validation and which were then trained on the full training data and evaluated on the test set (**File S11**). From this second round of evaluation, it was clear that the model with maximum tree depth 16, minimum child weight 2, learning rate 0.15, gamma 0.1, and columns sampled by tree 0.9 was the most performant model. This model showed the best performance on the test data (r^2 score 0.74) despite showing slightly lower performance on the train data.
The trained model was packaged within inSPIRE and the minimum, maximum, median, first quartile, third quartile, fraction of predicted Prosit-*deltas* above -0.1 and fraction of predicted Prosit-*deltas* above 0.0 were passed as features for Percolator.
As with the inSPIRE source code, all of training code for the Prosit-*delta* predictor is fully open-source. Hence, a user of inSPIRE could retrain this predictor on their own data and use their Prosit-*delta* model in the inSPIRE pipeline.

inSPIRE Implementation and Application
All inSPIRE jobs presented in this study may be recreated by providing the required inSPIRE configuration file. Full details on the creation of the inSPIRE config file may be found in the README available on GitHub. For each experiment “rescoreMethod” was set to “percolator” and “mzAccuracy” was set to 0.02. The search engine and location of search results as well as the location of scan data converted to either mgf or mzML was provided via the config file. For inSPIRE-affinity “useBindingAffinity” was set to “asFeature”. The calibrated collision energy was also set, which agreed between inSPIRE and the Prosit web server in all cases.
To generate Prosit predictions without specialized GPU hardware, we downloaded the Prosit model details and changed the definition of the CuDNNGRU layers in the model.yml file to GRU layers with the following settings, activation equal to tanh, recurrent_activation equal to sigmoid, unroll equal to false, use_bias equal to true, and reset_after equal to true. We also had avoid using the tensorflow graph as in the original Prosit code. We were then able to reload the model definition and weights using Tensorflow version 2.5 and execute predictions by modifying the open-source code available from the Prosit team (see https://github.com/kusterlab/prosit).
For timing comparisons of Prosit prediction on CPU against GPU, spectral prediction on CPU was run on Intel Sky Lake processors, while the GPU predictions were run on an NVIDIA Tesla K40m Graphics Card.
All Prosit spectral predictions were generated using the 2020 HCD model and iRT predictions using the 2019 model. For users who have very large datasets and easy access to GPU servers, we also provide the modified version of the original Prosit code, including a converted singularity image so that Prosit can be run on a high performance computing cluster, a change to the MSP export code so that Prosit predicted iRT values were included, and an option so that m/z values of all fragment ions were not calculated by Prosit. We found it was much more efficient to calculate the m/z values of the fragment ions in the inSPIRE pipeline and greatly reduced the required prediction time, particularly if a large number of predicted spectra were required.

Results

inSPIRE
We developed inSPIRE to be a flexible rescoring pipeline, which provides the power of Prosit prediction for users without specialized computational hardware and can be applied to a vast number of tandem MS proteomics datasets generated with HCD or CID fragmentation. Although it is optimized for HLA-I immunopeptidomics, inSPIRE can also be applied to standard proteomics experiments. inSPIRE provides flexibility through compatibility with commonly used search engines, *i.e.*, MaxQuant, PEAKS, or Mascot, as well as compatibility with open data formats, *i.e.*, mgf and mzML formats. For HLA-I

immunopeptidomics, the inSPIRE-affinity variant can be employed, which allows integration of NetMHCpan predictions of HLA-I-peptide binding affinity, and potentially others in future releases.

When using Prosit, inSPIRE is subject to the limitations of the Prosit predictor and will filter out PSMs where the peptides are of length less than 7 or greater than 30. If the sample contains unmodified cysteines (non-carbamidomethylated) or variable modifications other than the oxidation of methionine these PSMs will be filtered out by inSPIRE. Unmodified cysteines and a wider range of variable modifications are supported if the user selects MS^{2}PIP as their spectral predictor (inSPIRE-MS^{2}PIP supports a maximum of 9 unique modifications). However, we did not prioritize the development of the MS^{2}PIP pipeline given that there already exists a fully open-source rescoring pipeline, which utilizes MS^{2}PIP prediction in MS^{2}Rescore.

inSPIRE provides multiple pipelines to fulfil different user requirements. The core functionality is provided via the "core" pipeline (though individual steps may be run independently), which enables MS2 spectral rescoring (**Fig. 1**). The first subsection of the "core" functionality, "prepare", formats the search engine output for Prosit or MS^{2}PIP (and NetMHCpan if required). The required Prosit or MS^{2}PIP predictions are then generated via the "predictSpectra" pipeline. For Prosit, this entailed the conversion of the GPU only models available from the Prosit team to a version that could be run on an ordinary CPU (see for Tool Implementation and statistical analysis details). We found that execution of the "predictSpectra" pipeline on the CPU was effective and timing even compared favorably to execution of the original Prosit code when we removed the calculation of m/z values for all possible fragment ions (**Fig. S1**). We also validated that the predictions from the inSPIRE CPU implementation did not differ from the online Prosit model by running the spectral prediction pipelines for both tools on 13,054 unique peptide-charge combinations (the peptides identified by MaxQuant in the HLA-A02:01 immunopeptidome). We found that the predicted iRT values and MS2 spectra agreed to near machine single-precision with a mean spectral angle between predicted spectra of 0.9999997 of a mean difference in iRT of the order of 10^{-5} (**File S2**).

NetMHCpan prediction is not currently integrated within inSPIRE due to license restrictions, but if the user wishes to employ the inSPIRE-affinity variant they could generate the predictions independently (see instructions in the README on GitHub and **File S1**). The final part of the inSPIRE core pipeline, "rescore" utilizes all available data for improved rescoring. This process generates all required features from search results, spectral predictions and NetMHCpan predicted binding affinities. Once all features are generated, inSPIRE calls Percolator to rescore the PSMs. These results are then benchmarked against Percolator rescoring without spectral features and a HTML report is provided to the user (see the examples provided in **File S3**). This report provides details of varying feature importance, feature distributions, and performance of the inSPIRE pipeline against Percolator with classical features (**Fig. 1**). If inSPIRE-affinity was used, or binding affinity was selected as a validation technique, this report also shows the percentage of NetMHCpan predicted binders that are identified.

In addition to the core functionality, inSPIRE also provides a calibration pipeline, which allows calibration of the collision energy setting passed to Prosit. The inSPIRE calibration pipeline is a simple pipeline as described by Wilhelm *et al*. (29), where the highest scoring unmodified PSMs are considered based on search engine score and spectral angles against Prosit predictions for each collision energy between 20 and 40 (inclusive) are generated. The collision energy that provides the highest mean spectral angle is selected as the recommended collision energy setting for further analysis.

In inSPIRE, we have introduced a number of changes to the feature set and feature selection approaches compared to other spectral rescoring pipelines such as Prosit rescoring and MS²Rescore. For example, rather than providing features matching y- and b-ions, we distinguished between the dominant ion series (the series with greater predicted coverage) and the lesser ion series. While this is unlikely to impact tryptic proteome digestion datasets, where the y-series is generally dominant, we found it a more useful distinction for HLA-I immunopeptidome rescoring, where there is more variation in which ion series is dominant. We also found that considering m/z error on the MS2 fragment ions was a useful feature. We provide a full description of all features used by inSPIRE in **Table S1**.

Compared to other pipelines, another major change was the use of features from a Prosit-*delta* predictor (see Experimental Procedures section). In our ground truth datasets, we found that swapping the position of certain pairs of adjacent amino acids in the true peptide sequence led to a very small change in the PSM spectral angle. We termed this change the 'Prosit-*delta*'. In early development of inSPIRE, peptides that had a small Prosit-*delta* were often misassigned in the ground truth datasets, resulting in

incorrect (although similar) peptide sequences. While generating Prosit predicted spectra and spectral angles for swaps of every pair of adjacent amino acid residues would massively increase the computational load of the pipeline, we aimed to use a less intensive predictor to estimate the sensitivity of Prosit at each fragmentation site of the peptide. Therefore, the aim of including these Prosit-*delta* predictor features was to identify the sensitivity of the Prosit MS2 spectral prediction to minor changes in amino acid residue positions. These *delta* predictions are not available for the inSPIRE-MS^{2}PIP pipeline (see Experimental Procedures section for full details on the technical aspects of the Prosit-*delta* predictor).

Additional to its above-described flexibility, inSPIRE provides several options to allow for manual feature inclusion or exclusion by the user. For example if the user had a very small dataset where some features in the standard inSPIRE feature set could lead to the introduction of bias, they can simply add a list of "excludeFeatures" to the inSPIRE config file. Furthermore, if the user was particularly interested in certain sequence identifications and wished to examine their MS2 spectra more closely, inSPIRE provides a plotting tool, which generates pair plots in pdf format and compare the experimental MS2 spectrum to the Prosit predicted MS2 spectrum. An example of these plots for PSMs of varying quality is provided in **File S4**. All that is required is to select the rows of interest from the inSPIRE final assignments or provide a csv file with the peptides of interest along with their source file and scan number. This functionality may be of particular interest to users who wish to use inSPIRE, for example, for epitope target discovery in immunopeptidomics.

inSPIRE boosts PSM identifications in HLA-I immunopeptidomes and tryptic proteome digestions

We focused our initial benchmarking of inSPIRE against the Prosit rescoring pipeline and the MS^{2}Rescore pipeline with MaxQuant search results, as well as comparing it to a baseline rescoring without the use of spectral prediction. For comparison between the pipelines, we attempted to provide as fair a comparison between tools as possible; thereby, we reran the final Percolator rescoring with the same command line arguments used for all pipelines (see Experimental Procedures for details). However, one area of difference, which we could not correct for, was the fact that the current release of MS^{2}Rescore dropped PSMs with duplicate scan numbers, meaning that chimeric spectra could not be discovered. This feature might be changed in the next release of MS^{2}Rescore (personal communication), when we would expect an increase in PSMs identified.

We applied all pipelines – *i.e.* Prosit rescoring, inSPIRE, inSPIRE-affinity, MS^{2}Rescore, inSPIRE-MS^{2}PIP and inSPIRE-MS^{2}PIP-affinity – to HLA-I immunopeptidomes derived from K562-A*02:01 (**Fig. 2A-E, Fig. S2A,B**) and K562-B*07:02 (**Fig. 2F-J, Fig. S2C,D**) cell lines, and tryptic proteome digestions derived from K562 cell lines (**Fig. 2K-M**). Since the Prosit rescoring web server allowed only a single MS file per search, we analyzed a single MS file for both HLA-I immunopeptidomes (**Fig. 2A-J**) and tryptic proteome digestions (**Fig. 2K-M**).

In our initial MaxQuant analysis, we used a reference database informed by RNA sequencing of K562 cell lines, which consisted of 43,578 entries. Then, we repeated the analysis using the full Gencode reference database, which consisted of 392,583 entries. This strategy allowed evaluation of the impact of the reference database size on the PSM yield of inSPIRE compared to the other pipelines in the range of estimated FDRs 1-5% (**Fig. 2A-M**). We focused our analysis of the peptides identified on PSMs identified at 1% FDR as this is the most commonly employed FDR threshold in recent proteomics and immunopeptidomics studies (**Fig. 2, Fig. S2-S5**).

It has already been demonstrated that the Prosit rescoring pipeline and MS^{2}Rescore significantly increase PSM yield over baseline rescoring without spectral prediction (29, 30). Similarly, we observed a significant impact of inSPIRE, with more than a 150% increase in PSMs discovered at 1% FDR for all immunopeptidome datasets as compared to the baseline rescoring (**Fig. S2**).

For rescoring pipelines using spectral prediction applied to HLA-I immunopeptidomes, inSPIRE identified a slightly higher number of PSMs (4 – 6% increase) compared to the Prosit Rescoring pipeline and inSPIRE-affinity showed the highest PSM yield (8 – 9% increase on the Prosit Rescoring pipeline). The increase in PSMs identified between Prosit Rescoring and inSPIRE using MS^{2}PIP (3 - 6% increase) was similar to the increase of inSPIRE over Prosit Rescoring. In each case, MS^{2}Rescore identified the fewest PSMs at 1% FDR (**Fig. 2A,C,F,H**). In the case of the immunopeptidome dataset, this difference was unlikely to be explained entirely by the dropping of chimeric MS2 spectra; it may be more related to the fact that MS^{2}Rescore uses 100 features in its rescoring as opposed to the 40 features used by

Prosit Rescoring and the 41- 42 features used by inSPIRE and inSPIRE-affinity. This larger feature set may be less suitable when rescoring small immunopeptidome datasets as there is a greater risk of overfitting with a large number of features and a smaller dataset, leading to a reduced number of PSMs identified when cross-validation is applied within Percolator. In contrast to the performances on HLA-I immunopeptidomes, the performance of inSPIRE, Prosit Rescoring and inSPIRE-MS^{2}PIP was very similar on the tryptic proteome digestion dataset using both RNA-informed and full Gencode reference databases, with a marginal improvement in PSM yield by inSPIRE over Prosit Rescoring and a marginal increase by Prosit Rescoring over inSPIRE-MS^{2}PIP (**Fig. 2K,L**). Again, fewer PSMs were identified by MS^{2}Rescore, although, in this case, the difference could almost entirely be explained by the removal of chimeric MS2 spectra in the MS^{2}Rescore pipeline. The number of unique scans identified at 1% FDR was very similar for all pipelines, with all identifying approximately 30,000 unique scans.

To validate the assignments of each pipeline, we initially computed the percentage of peptides, identified at 1% FDR for each pipeline, which were predicted to bind the cognate HLA-I allele by NetMHCpan among the peptides identified in the HLA-I immunopeptidomes. This percentage was high and similar across all pipelines (**Fig. 2B,D,G,I**). As second validation step, we computed the percentage of peptides, identified using the Gencode reference database by each pipeline that were also identified using RNA-informed reference database. The analysis of this metrics also pointed toward a reliable peptide identification in both HLA-I immunopeptidomes and tryptic proteome digestions (**Fig. 2E,J,M**).

By examining the incremental PSMs discovered by competing pipelines, *i.e.* the PSMs exclusively discovered by one pipeline but not the other, we observed greater variation in these two validation metrics. We performed such "head-to-head" analysis for inSPIRE against baseline rescoring (**Fig. S2**), inSPIRE against Prosit Rescoring (**Fig. S3**), inSPIRE-affinity against inSPIRE (**Fig. S4**) and inSPIRE-MS^{2}PIP against MS^{2}Rescore (**Fig. S5**). The best performance on each metric was invariably observed in the pool of PSMs shared between pipelines. However, the incremental PSMs from the pipeline which identified the greater number of PSMs at 1%FDR generally showed higher values for the two validation metrics over the competing pipeline that identified fewer PSMs. Overall, we found that peptides exclusively identified by inSPIRE variants showed a higher percentage of peptides predicted to be HLA-I binders compared to those peptides that were exclusively identified by the baseline, Prosit rescoring and MS^{2}Rescore. Furthermore, in the latter comparisons, peptides exclusively identified by inSPIRE variants using the Gencode reference database were more frequently identified using RNA-informed reference database (**Fig. S2, S3, S5**). Only two exceptions broke this homogenous pattern: (i) the percentage of peptides predicted to be HLA-I binders in the K562-B*07:02 HLA-I immunopeptidomes using RNA-informed reference database comparing inSPIRE against Prosit rescoring pipeline (**Fig. S3C**)**;** (ii) the percentage of peptides identified using the Genecode reference database that were also identified using the RNA-informed reference database in the K562-B*07:02 HLA-I immunopeptidomes comparing inSPIRE-MS^{2}PIP against MS^{2}Rescore (**Fig. S5B**). These exceptions may indicate a level of noise in our validation metrics (see the caveats described in the Experimental Procedures section). However, overall, the evidence across all incremental PSM comparisons (**Fig. S2-S5**) and pipelines (**Fig. 2**), indicated a consistent quality in the PSMs identified by Percolator at 1% FDR for each pipeline.

In addition to these independent metrics, we also examined MS2 coverage and spectral angle distribution for the incremental PSMs discovered by competing pipelines (**Fig. S2-S5**), which could provide some insight into the features prioritized by each pipeline. Not surprisingly, we found that the PSMs identified by inSPIRE only had significant higher spectral angle distribution compared to the baseline rescoring pipeline which does not use features from Prosit (**Fig. S2**). Furthermore, PSMs exclusively identified by inSPIRE but not Prosit Rescoring typically had a greater MS2 coverage but a lower spectral angle than those exclusively identified by Prosit Rescoring but not by inSPIRE (**Fig. S3**).

In our comparison of inSPIRE-affinity to the standard inSPIRE pipeline we noted that the incremental PSMs identified by inSPIRE-affinity showed higher mean spectral angle and MS2 coverage over inSPIRE standard, despite the added importance of binding affinity (**Fig. S4**). Therefore, inSPIRE-affinity did not only identify peptides with higher HLA-I-peptide binding affinities, but also with overall better spectral features. This suggested that the MS2 spectral and HLA-I-peptide binding affinity prediction features worked effectively in concert rather than one aspect being solely prioritized over the other. The MS^{2}Rescore pipeline did not compute spectral angle between the experimental and MS^{2}PIP predicted MS2 spectra. Therefore, it should not come as a surprise that the mean spectral angle was greater for

the PSMs identified exclusively by inSPIRE-MS^{2}PIP than for PSMs identified exclusively by MS^{2}Rescore (**Fig. S5**).

To study the impact of inSPIRE on PSM yield as compared to Prosit rescoring on a wide variety of HLA-I alleles, we performed rescoring on 12 mono-allelic HLA-I cell lines from the large HLA-I immunopeptidome dataset published by Sarkizova *et al.* (55). For this analysis, we focused on acquiring data using diverse HLA-I alleles, and included datasets where peptide sequence motifs were less well understood (*e.g.*, the HLA-G alleles). This strategy allowed testing of the effect of inSPIRE-affinity in such challenging settings. Overall, we found that the resulting peptide sequence motifs from Prosit rescoring compared to inSPIRE rescoring were extremely similar (**Fig. S6**). However, with regards to peptide identification, we observed a 0.1-7.6% increase (mean = 3.1%) in PSMs identified at 1% FDR with the inSPIRE pipeline over the Prosit rescoring pipeline, and 0.6-10.6% increase (mean = 4.2%) over Prosit rescoring when using the inSPIRE-affinity pipeline (**Fig. S7**).

We then compared the percentage of peptides predicted by NetMHCpan to be either binders or strong binders of the cognate HLA-I complex and identified at 1% FDR across different HLA-I alleles with a broad range of NetMHCpan performance (**Fig. S8**). In this analysis, the variation in the percentage of peptides predicted to be HLA-I binders was larger between HLA-I alleles than between pipelines. In addition, in those HLA-I alleles for which NetMHCpan prediction reported a low NetMHCpan's PPV, *i.e*. where the HLA-I-peptide binding affinity was not efficiently predicted by NetMHCpan, inSPIRE-affinity showed a similar percentage of peptides predicted to be HLA-I binders than the other pipelines. This further indicates that inSPIRE-affinity did not blindly assign peptides based on predicted HLA-I-peptide binding affinity alone, particularly when the HLA-I-peptide binding affinity prediction was less reliable.

inSPIRE shows high specificity and stable performance on ground truth datasets of varying size

Although the validation analyses performed so far suggested a high performance of inSPIRE and inspire-affinity, we wished to further verify that the increased PSM yield observed by applying inSPIRE pipelines was due to an improved sensitivity of inSPIRE compared to the other pipelines, rather than the result of spurious identifications. To this end, we applied inSPIRE, Prosit Rescoring, and the baseline rescoring to ground truth datasets of synthetic peptide libraries of pathogen-derived 9, 10 and 15 amino acid long peptides (**File S5**). The ground truth dataset construction followed the approach described in Cormican and colleagues (53), and is explained in the Experimental Procedures section. The pipelines' benchmarking on a ground truth dataset containing PSMs with characteristics similar to HLA-I immunopeptidomes could let us estimate the precision – *i.e.* number of correctly identified peptides over the number of identified peptides – and recall – *i.e.* number of correctly identified peptides over the number of correct peptides – of a given method. The computation of precision and recall (PR) is a standard strategy for performance evaluation of binary predictors and has also been applied to proteomics in other contexts (56, 57). Optimal performance in terms of PR would show a tool achieving close to the maximum possible recall while maintaining high precision until very low scoring thresholds lead to a steep drop. The maximum possible recall for each rescoring pipeline was the fraction of the true PSMs correctly identified by the initial search engine at any identification cut off. Hence, a lower limit on the recall indicates that there were more incorrect assignments in the original database search. Within the immunopeptidomics field, we observed that implementing Percolator with a standard feature set on small datasets could lead to a lower precision (13). Therefore, we tested inSPIRE performance in ground truth datasets with increasing size, from a mean of just under 3,000 total PSMs using 2 RAW files to over 12,000 PSMs from 8 RAW files (**Fig. 3**). To remove the effects of different performance on different RAW files, we performed rescoring on 4 sets of 2 RAW files (**Fig. 3A**), 2 sets of 4 RAW files (**Fig. 3B**) and a single set of 8 RAW files (**Fig. 3C**) and calculated PR across all sets (see Experimental Procedures for more details). In the case of inSPIRE, we observed stable performance on all ground truth datasets, even when rescoring was performed on a small number of PSMs (**Fig. 3A**).

For the baseline rescoring, we observed very similar performance no matter the size of the dataset, achieving 19-20% recall at 99% precision for any dataset size. The pipelines using spectral prediction saw a steady increase in performance with dataset size. The Prosit rescoring pipeline increased from 32% recall at 99% precision when rescoring on 2 RAW files (**Fig. 3A**), to 36% recall at 99% precision when rescoring on 4 RAW files (**Fig. 3B**), to 38% recall at 99% precision when rescoring on all 8 RAW files (**Fig. 3C**). Similarly, with inSPIRE, at 99% precision, we observed the recall of 36% when rescoring on 2 RAW files (**Fig. 3A**), 40% when rescoring on 4 RAW files (**Fig. 3B**), and 41% when rescoring on 8

RAW files (**Fig. 3C**). Therefore, under all conditions, we observed a performance improvement of inSPIRE over Prosit rescoring (**Fig. 3A-C**), which was in line with the increase in PSMs observed on the HLA-I immunopeptidome datasets (**Fig. 2**). Hence, the results on the ground truth datasets provided further validation of the results on the HLA-I immunopeptidome datasets. To note, both Prosit-rescoring and inSPIRE obtained on average 98% precision at their respective estimated 1% FDRs across all datasets, indicating a slight underestimation of the FDR for both tools in these experimental conditions (**Fig. 3A-C**).

inSPIRE is performant on larger scale datasets and across search engines

In contrast to Prosit rescoring pipeline, inSPIRE supports multiple MS files in a single run, and can be combined with various database search engines (**Fig. 1**). To estimate how inSPIRE would perform on results from larger datasets – *e.g*. derived from multiple MS files - and different search engines, we tested inSPIRE on larger datasets of HLA-I immunopeptidomes and tryptic proteome digestions than those investigated in **Fig. 2**. Indeed, we applied inSPIRE to search results from three MS files of K562-B*07:02-derived HLA-I immunopeptidomes (**Fig. 4A,B**). As reference database, we use both RNA-informed and Gencode reference databases, thereby evaluating the impact of the reference database size on inSPIRE's PSM yield. Rescoring with inSPIRE increased the PSM yield at 1% FDR for all search engines by 31-33% for PEAKS DB, 225-281% for Mascot, and 120-127% for MaxQuant compared to the baseline Percolator rescoring. Interestingly the larger increase in PSMs using inSPIRE with PEAKS DB and MaxQuant was observed when using the RNA-informed rather than Gencode reference database. The best performance came from the rescoring of PEAKS DB search results with a 15-18% increase over MaxQuant results (**Fig. 4A,B**).

As with the performance using a single technical replicate (**Fig. 2E,J,M**), we observed a high and comparable percentage of peptides identified at 1% FDR when searching the Gencode reference database, which were also found in the search results of the RNA-informed reference database, with a minimum of 98.2% for Mascot search results after rescoring with inSPIRE and a maximum of 99.0% for the PEAKS DB baseline (**Fig. 4C**).

We performed the same analysis on six MS files from K562 cell tryptic proteome digestions using MaxQuant, Mascot and PEAKS DB. inSPIRE rescoring improved the PSM yield of all search engines compared to the baseline Percolator rescoring, which was even more pronounced than the HLA-I immunopeptidome datasets. As with the HLA-I immunopeptidome datasets, the best identification rate was achieved with inSPIRE rescoring of PEAKS DB search results (**Fig. 4D,E**).

While the percentage of peptides identified using the Gencode reference database identified also using the RNA-informed reference database was high and consistent for results with Mascot and PEAKS with and without rescoring, a slight decrease of this percentage was observed by applying the inSPIRE rescoring to MaxQuant results (from 98.9% to 97.2% of identified peptides; **Fig. 4F**).

The most remarkable variation in search engine performance from HA-I immunopeptidome to tryptic proteome digestion searches came from Mascot. Indeed, this search engine identified the fewest PSMs on the HLA-I immunopeptidome datasets although showed performance similar to PEAKS DB on the tryptic proteome digestion search results (**Fig. 4A,B,D,E**). This is in line with results obtained with other approaches (13).

More generally, inSPIRE rescoring was particularly impactful relative to the original search engine choice in the enlarged HLA-I immunopeptidome search space. Indeed, rescoring of MaxQuant search results in the tryptic proteome digestion search space still provided fewer identifications than PEAKS DB and Mascot baseline results. In opposite, in the HLA-I immunopeptidome results, even Mascot, the lowest performing search engine in this case, identified more PSMs after rescoring than PEAKS DB without rescoring at 1% FDR (**Fig. 4A-B**).

To understand the impact of the search engines on the pool of identified peptides, we analyzed the overlap among peptides identified using the three search engines with and without inSPIRE (**Fig. S9**). In the HLA-I immunopeptidome dataset, we observed that inSPIRE rescoring led to a particularly large increase in the number of shared identified peptides among the search engines (**Fig. S9A,B**). In the tryptic proteome digestion dataset, the impact was less striking since even without inSPIRE rescoring the majority of the identified peptides were discovered by all three search engines (**Fig. S9C,D**).

Furthermore, we investigated the impact of MS1 intensity on the ability of the search engines and inSPIRE rescoring to identify PSMs in HLA-I immunopeptidomes (**Fig. S10A-C**) and in tryptic proteome

digestions (**Fig. S10D-F**). In HLA-I immunopeptidomes, PSMs identified by inSPIRE only showed significantly lower MS1 intensity distributions compared to PSMs identified by both inSPIRE and the search engines and inSPIRE. This suggested that the use of inSPIRE allowed the detection of lower intensity PSMs in HLA-I immunopeptidomics (**Fig. S10A-C**). Such differences were, however, absent when analyzing tryptic proteome digestion samples (**Fig. S10D-F**).

Insight into inSPIRE optimization of spectral prediction features by modelling amino acid pair switch (Prosit-*delta*).

Beyond some improvements to the feature set and the integration of NetMHCpan predictions, the inSPIRE pipeline employs a novel approach to PSM rescoring, namely the prediction of the sensitivity of the Prosit MS2 spectrum prediction in case of switch of adjacent amino acid residue pairs. This switch (or permutation) had previously been noted by Collaert *et al*. (15) as a difference that traditional search engines struggled to detect. This sensitivity, or lack thereof, is represented in the examples of Prosit MS2 spectrum prediction of the peptides MATYGWNLVK and AIKVLRGFKK identified in the synthetic peptide library samples (**Fig. 5A-B**). For these peptides, we challenged Prosit MS2 spectrum prediction by switching the position of two adjacent amino acids and computed the difference in the spectral angle between the true and the modified peptides, which we named 'Prosit-*delta*' value. In the case of the peptide MATYGWNLVK, the position switch between alanine (A) and threonine (T) in the true peptide MATYGWNLVK, which resulted in the modified peptide MTAYGWNLVK, led to a large Prosit-*delta* value, with the spectral angle dropping from 0.92 for the original sequence to 0.61 for the modified sequence (**Fig. 5C**). In contrast, for the peptide AIKVLRGFKK, the switch in position between the phenylalanine (F) and lysine (K), which resulted in the theoretical peptide AIKVLRGKFK, led to a small Prosit-*delta* value (spectral angle drops from 0.88 to 0.86) as we saw only minor differences in predicted MS2 spectra between the original and the modified peptides (**Fig. 5D**).

In early development of inSPIRE, we noticed that misassigned peptide sequences in the synthetic peptides' ground truth datasets often occurred when a similar peptide sequence was found in the constructed reference database (data not shown); in particular, this often happened when a peptide sequence differed from the true peptide sequence without impacting on the spectral angle, *i.e*. with a small Prosit-*delta* (see representative example in **Fig. 5B,D**). Hence, we hypothesized that the distribution of these Prosit-*delta* values for each position in the sequence could be a useful feature in rescoring, and that sequences where the Prosit spectral angle was less sensitive to minor changes in amino acid position should be assigned with less confidence than those where the spectral angle was more sensitive. To avoid the heavy computational burden of generating Prosit predicted spectra for all modified sequences, we developed a model to predict the Prosit-*delta* values as described in the Experimental Procedures. To make the model as independent as possible of the datasets benchmarked with inSPIRE, we used large publicly available HLA-I immunopeptidome data from Paes *et al*. (39) and Bassani-Sternberg *et al.* (40), as well as the synthetic immuno-peptide dataset used to train the Prosit model (29) as training and test data (**Table S2**). Peptide length and charge state distribution of the training data reflected those distributions typically observed in HLA-I immunopeptidome datasets (**Fig. S11A,B**). The peptide sequence motifs in the training dataset were evenly distributed, thereby confirming that we were not training the Prosit-*delta* predictor on peptides which were biased toward some specific sequence motifs (**Fig. S11C-F**).

The resulting Prosit-*delta* predictor was an ensemble learning based model, which primarily focused on the local features of the permutation site, although further features such as precursor charge state, were also considered (**Fig. S12**). Interestingly, collision energy was one of the most important features, which highlights the need for careful calibration of Prosit before usage. The full feature set used by the Prosit-*delta* predictor is described in detail in **Table S3**. Application of the trained Prosit-*delta* predictor to K562-A*02:01 and -B*07:02 HLA-I immunopeptidomes and tryptic proteome digestion dataset resulted in good performance (**Fig. 5E,F**).

To understand the impact of these Prosit-*delta* predictions on inSPIRE performance, we rescored the search results of the HLA-I immunopeptidome and tryptic proteome digestion (see **Fig. 2** and **Fig. S2-S4**) with the Prosit-*delta* features excluded. We found that the Prosit-*delta* features had little to no impact when applied to the tryptic proteome digestion datasets, where the enzyme specificity reduced the search space and made the search engine more sensitive to changes in sequence (**Fig. 5G**). On the HLA-I immunopeptidomes, the Prosit-*delta* implementation had an impact when the RNA-informed

reference database was used, although the most impact was observed when the Gencode reference database was used (**Fig. 5H, Fig. S13**). For the search of the K562-B*07:02 HLA-I immunopeptidome using the Gencode reference database, the increase in PSM yield over Prosit rescoring was almost entirely due to the Prosit-*delta* features (**Fig. 5H**). As with the tryptic proteome digestions, the Prosit-*delta* features had little impact when applied to inSPIRE-affinity (**Fig. 5H**). This could be explained by the fact that peptide sequence motifs driving the HLA-I-peptide binding motifs – and, hence, the HLA-I-peptide binding affinity prediction – are already very sensitive to minor changes in peptide sequence. Therefore, in inSPIRE-affinity, the impact of Prosit-*delta* features might be attenuated by the impact of HLA-I-peptide binding affinity prediction.

To validate these latter results, we again analyzed the percentage of peptides predicted by NetMHCpan to be HLA-I binders as well as the percentage of peptides identified using the Gencode reference database that were also identified using the RNA-informed reference database. As observed in the previous analyses, the various inSPIRE pipelines resulted in a high and comparable peptide percentage (**Fig. S14**), which hinted toward a reliable peptide identification.

Discussion

The integration of MS2 spectral prediction with rescoring strategies is a fruitful area to boost MS identification performance, and could find in the inSPIRE pipeline a versatile, high performing, user-friendly and open-source tool. The ability of inSPIRE to generate Prosit predictions on a standard CPU architecture significantly lowers the entry barrier for researchers, thereby "bringing Prosit to the people". The standard implementation of inSPIRE has demonstrated similar performance to the Prosit Rescoring pipeline for search results of tryptic proteome digestions and improved performance for HLA-I immunopeptidomes. The integration of NetMHCpan prediction of HLA-I-peptide binding affinity in inSPIRE-affinity pipeline raises the performance even further by optimizing inSPIRE for the analysis of HLA-I immunopeptidomes. In this study, the increased PSM identification rate of inSPIRE over the Prosit rescoring and MS^{2}Rescore pipelines has further been validated by investigation of the peptides identified. We have observed consistency in the percentage of identified peptides predicted to bind to the cognate HLA-I complex and the percentage of peptides identified from the Gencode reference database that were validated by RNA sequencing evidence. Furthermore, we have demonstrated the improved recall of inSPIRE over Prosit rescoring at 99% precision on ground truth datasets. In comparison to Prosit rescoring, the inSPIRE pipeline also brings significant benefits in terms of data volume and flexibility across multiple search engines. We have demonstrated that inSPIRE can provide increased PSM identification rates in each of these scenarios. The ability to apply inSPIRE to PEAKS DB, in particular, allows for significant improvement over rescoring of MaxQuant search results. Finally, we provide a detailed documentation and a step-by-step user guide to achieve easy access to inSPIRE for both the coding-experienced and -inexperienced user.

The addition of the Prosit-*delta* predictor boosts identification rates and can open potential avenues for other features based on meta-analysis, *i.e.,* features considering not only the match between experimental and predicted MS2 spectrum of a given peptide, but the uniqueness and sensitivity of the prediction. These features showed a greater impact on analysis confronted with larger search spaces such as the full Gencode database. This suggests the Prosit-*delta* features may assist with the challenges raised by the expansion of database size through the increased interest in noncanonical peptide identification in proteomics and immunopeptidomics. For example, the impact of the database size on method performance has been demonstrated for post-translational spliced peptides (13), and spectral prediction features as a solution for the search space size problem in proteogenomics have been proposed (31).

As suggested by Verbruggen *et al*. (31), we found the spectral prediction features far more impactful when dealing with the larger immunopeptidome search space compared to tryptic proteome digestion search spaces. Interestingly, this was not always the case when comparing results from the full Gencode reference database to the results from the RNA-informed reference databases. In fact, the increase in PSMs identified by applying inSPIRE and Prosit rescoring compared to the baseline, which was observed on the K562-B*07:02 HLA-I immunopeptidomes, showed to be larger when the RNA-informed reference database rather than the full Gencode reference database was used. This could point toward a limitation on the potential improvement in peptide yield on rescoring, when the true peptide is not

selected by the original search engine at any confidence threshold, and, hence, cannot be found through rescoring. This limitation may also contribute to the lack of variation between inSPIRE and Prosit Rescoring pipelines for the search results of the tryptic proteome digestions. In this case, the number of MS2 scans left unidentified in the rescored search results was similar to the number of decoy hits in the search results. Hence, any great increase in identification rate on the tryptic proteome digestion datasets would have to be treated with a certain degree of suspicion.

In conclusion, we speculate that the application of rescoring pipelines using MS2 spectral features will become the standard approach to tackle large search space problems in proteogenomics. We, here, provide a fully open-source tool, inSPIRE, which can aid flexible MS analysis pipeline development in a user-friendly manner in the future.

Data and software availability

Our MS proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE (58) partner repository with the dataset identifier PXD031709 (13), PXD031812 (53), and PXD034056.

The MS proteomics data published by Paes *et al*. (39) are available at the PRIDE repository with the dataset identifier PXD015489.

The RNA sequencing data used for the analysis of our MS proteomics data have been deposited in the NCBI Sequence Read Archive database with the accession code PRJNA721129 (13).

The inSPIRE software has been implemented with Python and is available at GitHub (https://github.com/QuantSysBio/inSPIRE)

The Prosit-*delta* training software has been implemented with Python and is available at GitHub (https://github.com/QuantSysBio/prosit-delta).

The modified version of the Prosit spectral prediction code is also available at GitHub (https://github.com/QuantSysBio/qsb-modified-prosit).

The models downloaded by the inSPIRE pipeline at run time have been deposited on figshare (https://figshare.com/articles/software/inSPIRE_Models/20368035).

The RNA sequencing data generated Peas *et al.* (39) are available upon request to the authors.

Analyses were carried out in Python 3.8.

Figures have been generated in Python using the Plotly library and Logomaker for the sequence logo plots (59). Postprocessing was done with Adobe Illustrator v26.2.

MS analysis was carried out with MaxQuant version 1.16.17, Mascot v2.7.01, PEAKS X Pro 10.6. Rescoring was carried out with Percolator version 3.0.5. Preprocessing of MS RAW files for Mascot was performed with Mascot-Distiller version 2.8.0.1 and RAW files were converted to mgf/mzML format for inSPIRE input using ms-convert GUI (ProteoWizard version 3.0.9134) and ThermoRawFileParser version 1.4.0 for input in Prosit-*delta* training pipeline (51).

Acknowledgments

We thank: (i) H. Urlaub and L. Welp (MPI-NAT), X. Yang and S. Lynham (KCL) for MS assistance, (ii) the Gesellschaft fuer wissenschaftliche Datenverarbeitung mbH Goettingen (GWDG) for support and access to the GWDG GPU-cluster; (iii) the Proteomics Facility at MPI-NAT for computational infrastructure support; (iv) the Percolator and MS^{2}Rescore teams for their support via GitHub; (v) A. Sette and J. Sidney (LJIAI) for providing the synthetic peptide library; N.C. Chiam (MPI-NAT) for designing the inSPIRE logo.

The study was in part supported by: (i) MPI-NAT collaboration agreement 2020, Cancer Research UK [C67500; A29686] and National Institute for Health Research (NIHR) Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust and King's College London and/or the NIHR Clinical Research Facility to MM; (ii) European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 945528) to JL. JAC and YH are supported by the International Max-Planck Research School (IMPRS) for Genome Science. WTS is supported by the European Union's Framework Programme for Research and Innovation Horizon Europe (2021-2027) under the Marie Skłodowska-Curie Grant Agreement No. 101065466.

Conflict of Interest

The Authors have no competing interests to declare.

References

1. Abelin, J. G., Keskin, D. B., Sarkizova, S., Hartigan, C. R., Zhang, W., Sidney, J., Stevens, J., Lane, W., Zhang, G. L., Eisenhaure, T. M., Clauser, K. R., Hacohen, N., Rooney, M. S., Carr, S. A., and Wu, C. J. (2017) Mass Spectrometry Profiling of HLA-Associated Peptidomes in Mono-allelic Cells Enables More Accurate Epitope Prediction. *Immunity* 46, 315-326
2. Aebersold, R., and Mann, M. (2016) Mass-spectrometric exploration of proteome structure and function. *Nature* 537, 347-355
3. Ouspenskaia, T., Law, T., Clauser, K. R., Klaeger, S., Sarkizova, S., Aguet, F., Li, B., Christian, E., Knisbacher, B. A., Le, P. M., Hartigan, C. R., Keshishian, H., Apffel, A., Oliveira, G., Zhang, W., Chen, S., Chow, Y. T., Ji, Z., Jungreis, I., Shukla, S. A., Justesen, S., Bachireddy, P., Kellis, M., Getz, G., Hacohen, N., Keskin, D. B., Carr, S. A., Wu, C. J., and Regev, A. (2022) Unannotated proteins expand the MHC-I-restricted immunopeptidome in cancer. *Nat Biotechnol* 40, 209-217
4. Verheggen, K., Raeder, H., Berven, F. S., Martens, L., Barsnes, H., and Vaudel, M. (2020) Anatomy and evolution of database search engines-a central component of mass spectrometry based proteomic workflows. *Mass Spectrom Rev* 39, 292-306
5. Kall, L., Storey, J. D., MacCoss, M. J., and Noble, W. S. (2008) Posterior error probabilities and false discovery rates: two sides of the same coin. *J Proteome Res* 7, 40-44
6. Cravatt, B. F., Simon, G. M., and Yates, J. R., 3rd (2007) The biological impact of mass-spectrometry-based proteomics. *Nature* 450, 991-1000
7. Caron, E., Kowalewski, D. J., Chiek Koh, C., Sturm, T., Schuster, H., and Aebersold, R. (2015) Analysis of Major Histocompatibility Complex (MHC) Immunopeptidomes Using Mass Spectrometry. *Mol Cell Proteomics* 14, 3105-3117
8. Barbosa, C. R. R., Barton, J., Shepherd, A. J., and Mishto, M. (2021) Mechanistic diversity in MHC class I antigen recognition. *Biochem J* 478, 4187-4202
9. Liepe, J., Holzhutter, H. G., Bellavista, E., Kloetzel, P. M., Stumpf, M. P., and Mishto, M. (2015) Quantitative time-resolved analysis reveals intricate, differential regulation of standard- and immuno-proteasomes. *Elife* 4, e07545
10. Mishto, M., Liepe, J., Textoris-Taube, K., Keller, C., Henklein, P., Weberruss, M., Dahlmann, B., Enenkel, C., Voigt, A., Kuckelkorn, U., Stumpf, M. P., and Kloetzel, P. M. (2014) Proteasome isoforms exhibit only quantitative differences in cleavage and epitope generation. *Eur J Immunol* 44, 3508-3521
11. Mansurkhodzhaev, A., Barbosa, C. R. R., Mishto, M., and Liepe, J. (2021) Proteasome-Generated cis-Spliced Peptides and Their Potential Role in CD8(+) T Cell Tolerance. *Front Immunol* 12, 614276
12. Goodenough, E., Robinson, T. M., Zook, M. B., Flanigan, K. M., Atkins, J. F., Howard, M. T., and Eisenlohr, L. C. (2014) Cryptic MHC class I-binding peptides are revealed by aminoglycoside-induced stop codon read-through into the 3' UTR. *Proc Natl Acad Sci U S A* 111, 5670-5675
13. Mishto, M., Horokhovskyi, Y., Cormican, J. A., Yang, X., Lynham, S., Urlaub, H., and Liepe, J. (2022) Database search engines and target database features impinge upon the identification of post-translationally cis-spliced peptides in HLA class I immunopeptidomes. *Proteomics*, e2100226
14. Ruiz Cuevas, M. V., Hardy, M. P., Holly, J., Bonneil, E., Durette, C., Courcelles, M., Lanoix, J., Cote, C., Staudt, L. M., Lemieux, S., Thibault, P., Perreault, C., and Yewdell, J. W. (2021) Most non-canonical proteins uniquely populate the proteome or immunopeptidome. *Cell Rep* 34, 108815
15. Colaert, N., Degroeve, S., Helsens, K., and Martens, L. (2011) Analysis of the resolution limitations of peptide identification algorithms. *J Proteome Res* 10, 5555-5561
16. Krug, K., Carpy, A., Behrends, G., Matic, K., Soares, N. C., and Macek, B. (2013) Deep coverage of the Escherichia coli proteome enables the assessment of false discovery rates in simple proteogenomic experiments. *Mol Cell Proteomics* 12, 3420-3430
17. Kall, L., Canterbury, J. D., Weston, J., Noble, W. S., and MacCoss, M. J. (2007) Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat Methods* 4, 923-925
18. Ma, K., Vitek, O., and Nesvizhskii, A. I. (2012) A statistical model-building perspective to identification of MS/MS spectra with PeptideProphet. *BMC Bioinformatics* 13 Suppl 16, S1
19. Searle, B. C. (2010) Scaffold: a bioinformatic tool for validating MS/MS-based proteomic studies. *Proteomics* 10, 1265-1269
20. The, M., MacCoss, M. J., Noble, W. S., and Kall, L. (2016) 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

21. Granholm, V., Noble, W. S., and Kall, L. (2012) A cross-validation scheme for machine learning algorithms in shotgun proteomics. *BMC Bioinformatics* 13 Suppl 16, S3
22. Giese, S. H., Sinn, L. R., Wegner, F., and Rappsilber, J. (2021) Retention time prediction using neural networks increases identifications in crosslinking mass spectrometry. *Nat Commun* 12, 3237
23. Bichmann, L., Nelde, A., Ghosh, M., Heumos, L., Mohr, C., Peltzer, A., Kuchenbecker, L., Sachsenberg, T., Walz, J. S., Stevanovic, S., Rammensee, H. G., and Kohlbacher, O. (2019) MHCquant: Automated and Reproducible Data Analysis for Immunopeptidomics. *J Proteome Res* 18, 3876-3884
24. AS, C. S., Bouwmeester, R., Martens, L., and Degroeve, S. (2019) Accurate peptide fragmentation predictions allow data driven approaches to replace and improve upon proteomics search engine scoring functions. *Bioinformatics* 35, 5243-5248
25. Elias, J. E., Gibbons, F. D., King, O. D., Roth, F. P., and Gygi, S. P. (2004) Intensity-based protein identification by machine learning from a library of tandem mass spectra. *Nat Biotechnol* 22, 214-219
26. Degroeve, S., and Martens, L. (2013) MS2PIP: a tool for MS/MS peak intensity prediction. *Bioinformatics* 29, 3199-3203
27. Degroeve, S., Maddelein, D., and Martens, L. (2015) MS2PIP prediction server: compute and visualize MS2 peak intensity predictions for CID and HCD fragmentation. *Nucleic Acids Res* 43, W326-330
28. Gessulat, S., Schmidt, T., Zolg, D. P., Samaras, P., Schnatbaum, K., Zerweck, J., Knaute, T., Rechenberger, J., Delanghe, B., Huhmer, A., Reimer, U., Ehrlich, H. C., Aiche, S., Kuster, B., and Wilhelm, M. (2019) Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat Methods* 16, 509-518
29. Wilhelm, M., Zolg, D. P., Graber, M., Gessulat, S., Schmidt, T., Schnatbaum, K., Schwencke-Westphal, C., Seifert, P., de Andrade Kratzig, N., Zerweck, J., Knaute, T., Braunlein, E., Samaras, P., Lautenbacher, L., Klaeger, S., Wenschuh, H., Rad, R., Delanghe, B., Huhmer, A., Carr, S. A., Clauser, K. R., Krackhardt, A. M., Reimer, U., and Kuster, B. (2021) Deep learning boosts sensitivity of mass spectrometry-based immunopeptidomics. *Nat Commun* 12, 3346
30. Declercq, A., Bouwmeester, R., Hirschler, A., Carapito, C., Degroeve, S., Martens, L., and Gabriels, R. (2022) MS(2)Rescore: Data-driven rescoring dramatically boosts immunopeptide identification rates. *Mol Cell Proteomics*, 100266
31. Verbruggen, S., Gessulat, S., Gabriels, R., Matsaroki, A., Van de Voorde, H., Kuster, B., Degroeve, S., Martens, L., Van Criekinge, W., Wilhelm, M., and Menschaert, G. (2021) Spectral Prediction Features as a Solution for the Search Space Size Problem in Proteogenomics. *Mol Cell Proteomics* 20, 100076
32. Gabriel, W., The, M., Zolg, D. P., Bayer, F. P., Shouman, O., Lautenbacher, L., Schnatbaum, K., Zerweck, J., Knaute, T., Delanghe, B., Huhmer, A., Wenschuh, H., Reimer, U., Medard, G., Kuster, B., and Wilhelm, M. (2022) Prosit-TMT: Deep Learning Boosts Identification of TMT-Labeled Peptides. *Anal Chem*
33. Zolg, D. P., Gessulat, S., Paschke, C., Graber, M., Rathke-Kuhnert, M., Seefried, F., Fitzemeier, K., Berg, F., Lopez-Ferrer, D., Horn, D., Henrich, C., Huhmer, A., Delanghe, B., and Frejno, M. (2021) INFERYS rescoring: Boosting peptide identifications and scoring confidence of database search results. *Rapid Commun Mass Spectrom*, e9128
34. Goloborodko, A. A., Levitsky, L. I., Ivanov, M. V., and Gorshkov, M. V. (2013) Pyteomics--a Python framework for exploratory data analysis and rapid software prototyping in proteomics. *J Am Soc Mass Spectrom* 24, 301-304
35. Levitsky, L. I., Klein, J. A., Ivanov, M. V., and Gorshkov, M. V. (2019) Pyteomics 4.0: Five Years of Development of a Python Proteomics Framework. *J Proteome Res* 18, 709-714
36. Jurtz, V., Paul, S., Andreatta, M., Marcatili, P., Peters, B., and Nielsen, M. (2017) NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *J Immunol* 199, 3360-3368
37. Reynisson, B., Alvarez, B., Paul, S., Peters, B., and Nielsen, M. (2020) 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
38. Nielsen, M., Lundegaard, C., Blicher, T., Lamberth, K., Harndahl, M., Justesen, S., Roder, G., Peters, B., Sette, A., Lund, O., and Buus, S. (2007) NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. *PLoS One* 2, e796
39. Paes, W., Leonov, G., Partridge, T., Chikata, T., Murakoshi, H., Frangou, A., Brackenridge, S., Nicastri, A., Smith, A. G., Learn, G. H., Li, Y., Parker, R., Oka, S., Pellegrino, P., Williams, I., Haynes, B. F., McMichael, A. J., Shaw, G. M., Hahn, B. H., Takiguchi, M., Ternette, N., and Borrow, P. (2019) Contribution of proteasome-catalyzed peptide cis-splicing to viral targeting by CD8(+) T cells in HIV-1 infection. *Proc Natl Acad Sci U S A* 116, 24748-24759

40. Bassani-Sternberg, M., Chong, C., Guillaume, P., Solleder, M., Pak, H., Gannon, P. O., Kandalaft, L. E., Coukos, G., and Gfeller, D. (2017) Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allostery regulating HLA specificity. *PLoS Comput Biol* 13, e1005725
41. Hughes, C. S., Moggridge, S., Muller, T., Sorensen, P. H., Morin, G. B., and Krijgsveld, J. (2019) Single-pot, solid-phase-enhanced sample preparation for proteomics experiments. *Nat Protoc* 14, 68-85
42. Gutman, I., Gutman, R., Sidney, J., Chihab, L., Mishto, M., Liepe, J., Chiem, A., Greenbaum, J., Yan, Z., Sette, A., Kosaloglu-Yalcin, Z., and Peters, B. (2022) Predicting the Success of Fmoc-Based Peptide Synthesis. *ACS Omega* 7, 23771-23781
43. Li, S., Sullivan, N. L., Rouphael, N., Yu, T., Banton, S., Maddur, M. S., McCausland, M., Chiu, C., Canniff, J., Dubey, S., Liu, K., Tran, V., Hagan, T., Duraisingham, S., Wieland, A., Mehta, A. K., Whitaker, J. A., Subramaniam, S., Jones, D. P., Sette, A., Vora, K., Weinberg, A., Mulligan, M. J., Nakaya, H. I., Levin, M., Ahmed, R., and Pulendran, B. (2017) Metabolic Phenotypes of Response to Vaccination in Humans. *Cell* 169, 862-877 e817
44. Chiu, C., McCausland, M., Sidney, J., Duh, F. M., Rouphael, N., Mehta, A., Mulligan, M., Carrington, M., Wieland, A., Sullivan, N. L., Weinberg, A., Levin, M. J., Pulendran, B., Peters, B., Sette, A., and Ahmed, R. (2014) Broadly reactive human CD8 T cells that recognize an epitope conserved between VZV, HSV and EBV. *PLoS Pathog* 10, e1004008
45. Weiskopf, D., Angelo, M. A., Grifoni, A., O'Rourke, P. H., Sidney, J., Paul, S., De Silva, A. D., Phillips, E., Mallal, S., Premawansa, S., Premawansa, G., Wijewickrama, A., Peters, B., and Sette, A. (2016) HLA-DRB1 Alleles Are Associated With Different Magnitudes of Dengue Virus-Specific CD4+ T-Cell Responses. *J Infect Dis* 214, 1117-1124
46. Weiskopf, D., Angelo, M. A., de Azeredo, E. L., Sidney, J., Greenbaum, J. A., Fernando, A. N., Broadwater, A., Kolla, R. V., De Silva, A. D., de Silva, A. M., Mattia, K. A., Doranz, B. J., Grey, H. M., Shresta, S., Peters, B., and Sette, A. (2013) Comprehensive analysis of dengue virus-specific responses supports an HLA-linked protective role for CD8+ T cells. *Proc Natl Acad Sci U S A* 110, E2046-2053
47. Weiskopf, D., Angelo, M. A., Bangs, D. J., Sidney, J., Paul, S., Peters, B., de Silva, A. D., Lindow, J. C., Diehl, S. A., Whitehead, S., Durbin, A., Kirkpatrick, B., and Sette, A. (2015) The human CD8+ T cell responses induced by a live attenuated tetravalent dengue vaccine are directed against highly conserved epitopes. *J Virol* 89, 120-128
48. Weiskopf, D., Cerpas, C., Angelo, M. A., Bangs, D. J., Sidney, J., Paul, S., Peters, B., Sanches, F. P., Silvera, C. G., Costa, P. R., Kallas, E. G., Gresh, L., de Silva, A. D., Balmaseda, A., Harris, E., and Sette, A. (2015) Human CD8+ T-Cell Responses Against the 4 Dengue Virus Serotypes Are Associated With Distinct Patterns of Protein Targets. *J Infect Dis* 212, 1743-1751
49. Weiskopf, D., Bangs, D. J., Sidney, J., Kolla, R. V., De Silva, A. D., de Silva, A. M., Crotty, S., Peters, B., and Sette, A. (2015) Dengue virus infection elicits highly polarized CX3CR1+ cytotoxic CD4+ T cells associated with protective immunity. *Proc Natl Acad Sci U S A* 112, E4256-4263
50. Weiskopf, D., Angelo, M. A., Sidney, J., Peters, B., Shresta, S., and Sette, A. (2014) Immunodominance changes as a function of the infecting dengue virus serotype and primary versus secondary infection. *J Virol* 88, 11383-11394
51. Hulstaert, N., Shofstahl, J., Sachsenberg, T., Walzer, M., Barsnes, H., Martens, L., and Perez-Riverol, Y. (2020) ThermoRawFileParser: Modular, Scalable, and Cross-Platform RAW File Conversion. *J Proteome Res* 19, 537-542
52. Frankish, A., Diekhans, M., Ferreira, A. M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J. M., Sisu, C., Wright, J., Armstrong, J., Barnes, I., Berry, A., Bignell, A., Carbonell Sala, S., Chrast, J., Cunningham, F., Di Domenico, T., Donaldson, S., Fiddes, I. T., Garcia Giron, C., Gonzalez, J. M., Grego, T., Hardy, M., Hourlier, T., Hunt, T., Izuogu, O. G., Lagarde, J., Martin, F. J., Martinez, L., Mohanan, S., Muir, P., Navarro, F. C. P., Parker, A., Pei, B., Pozo, F., Ruffier, M., Schmitt, B. M., Stapleton, E., Suner, M. M., Sycheva, I., Uszczynska-Ratajczak, B., Xu, J., Yates, A., Zerbino, D., Zhang, Y., Aken, B., Choudhary, J. S., Gerstein, M., Guigo, R., Hubbard, T. J. P., Kellis, M., Paten, B., Reymond, A., Tress, M. L., and Flicek, P. (2019) GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res* 47, D766-D773
53. Cormican, J. A., Soh, W. T., Mishto, M., and Liepe, J. (2022) iBench: A ground truth approach for advanced validation of mass spectrometry identification method. *Proteomics*, e2200271
54. Chen, T. Q., and Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System. *Kdd'16: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 785-794
55. Sarkizova, S., Klaeger, S., Le, P. M., Li, L. W., Oliveira, G., Keshishian, H., Hartigan, C. R., Zhang, W., Braun, D. A., Ligon, K. L., Bachireddy, P., Zervantonakis, I. K., Rosenbluth, J. M., Ouspenskaia, T., Law, T., Justesen, S., Stevens, J., Lane, W. J., Eisenhaure, T., Lan Zhang, G., Clauser, K. R., Hacohen, N., Carr, S. A., Wu, C. J., and Keskin, D. B. (2020) A large peptidome dataset

improves HLA class I epitope prediction across most of the human population. *Nat Biotechnol* 38, 199-209
56. Collatz, M., Mock, F., Barth, E., Holzer, M., Sachse, K., and Marz, M. (2021) EpiDope: a deep neural network for linear B-cell epitope prediction. *Bioinformatics*
57. Cox, J., Hein, M. Y., Luber, C. A., Paron, I., Nagaraj, N., and Mann, M. (2014) Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol Cell Proteomics* 13, 2513-2526
58. Perez-Riverol, Y., Csordas, A., Bai, J., Bernal-Llinares, M., Hewapathirana, S., Kundu, D. J., Inuganti, A., Griss, J., Mayer, G., Eisenacher, M., Perez, E., Uszkoreit, J., Pfeuffer, J., Sachsenberg, T., Yilmaz, S., Tiwary, S., Cox, J., Audain, E., Walzer, M., Jarnuczak, A. F., Ternent, T., Brazma, A., and Vizcaino, J. A. (2019) The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res* 47, D442-D450
59. Tareen, A., and Kinney, J. B. (2020) Logomaker: beautiful sequence logos in Python. *Bioinformatics* 36, 2272-2274

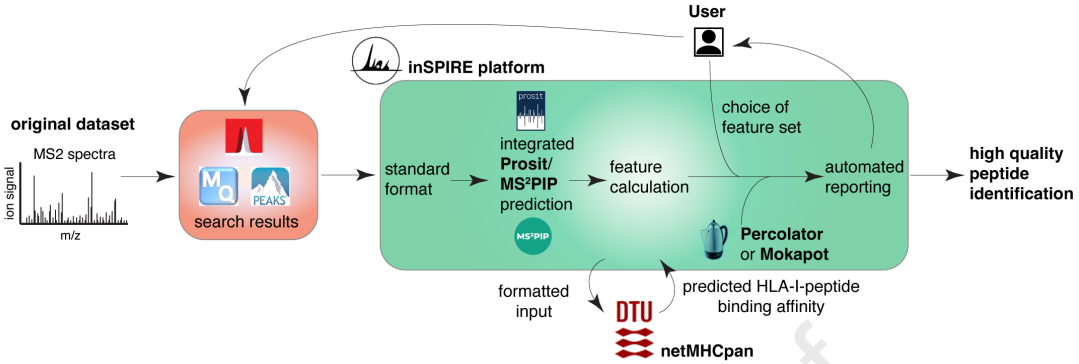
Figure 1. Schematic of the inSPIRE pipeline. Shown is an overview of the flow of execution for the inSPIRE pipeline. inSPIRE takes inputs from Mascot, PEAKS DB, or MaxQuant and interacts with Prosit, MS^{2}PIP, and NetMHCpan, providing formatted inputs and using their respective predictions for rescoring. inSPIRE calls Percolator internally to execute the PSM rescoring. The user provides a configuration file for inSPIRE and must work with *Prosit* (and optionally NetMHCpan) to provide the predictions to inSPIRE.

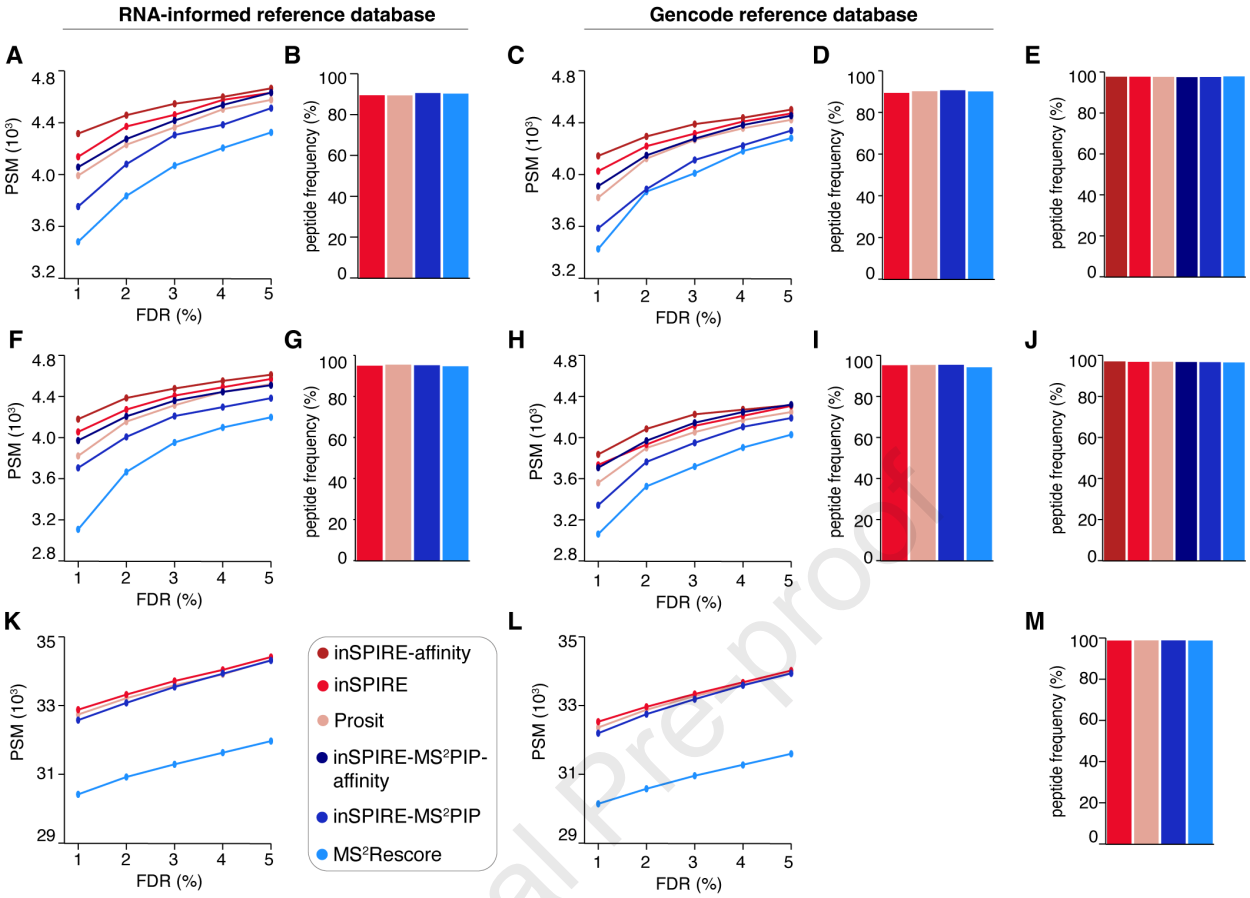
Figure 2. PSM identification by inSPIRE compared to the Prosit rescoring pipeline on tryptic proteome digestions and HLA-I immunopeptidomes. (**A-M**) Analysis of the PSMs identified in HLA-I immunopeptidomes of K562-A*02:01 (**A-E**) and K562-B*07:02 (**F-J**) cell lines, as well as the tryptic proteome digestion of the K562 cell line (**K-M**) by different pipelines. In (**A,B,F,G,K**) the RNA-informed and in (**C,D,H,I,L**) the full Gencode reference databases have been used. (**A,C,F,H,K,L**) Number of PSMs identified by applying the pipelines on MaxQuant search results. (**B**,**D,G,I**) Comparison of the percentage of identified peptides also predicted to bind the HLA-A*02:01 (**B,D**) and HLA-B*07:02 (**G,I**) complexes. Only peptides with lengths between 8-14 residues were included to allow the NetMHCpan HLA-I-peptide binding prediction. (**E,J,M**) The percentage of peptides identified by each pipeline on search results using the Gencode reference database which were also found at any confidence level in the MaxQuant search of the RNA-informed database of the cognate cell line.

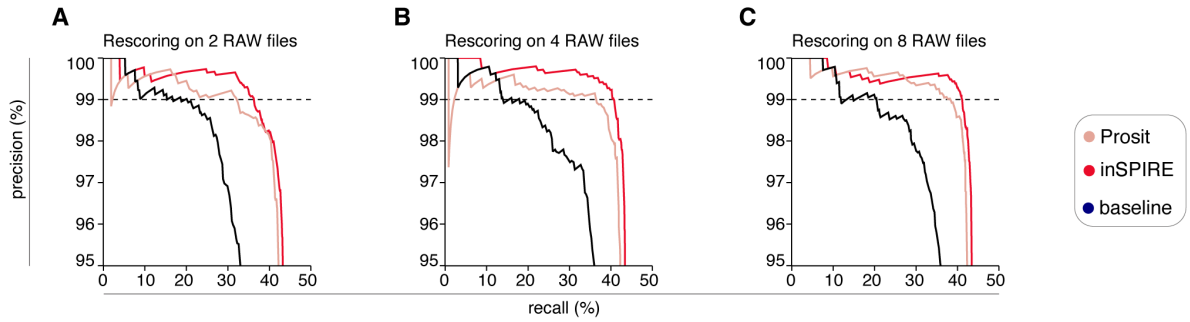
Figure 3. Performance of inSPIRE compared to the Prosit rescoring pipeline on synthetic peptides' ground truth dataset. (**A-C**) Shown is the precision (number of correctly identified peptides over number of identified peptides) against the recall (number of correctly identified peptides over number of correct peptides) of different rescoring pipelines. (**A**) Performance of different rescoring pipelines on ground truth data from two MS files (mean of 2,108 target PSMs, 824 decoy PSMs). (**B**) Performance of different rescoring pipelines on ground truth data from four MS files (mean 4,216 target PSMs, 1,647 decoy PSMs). (**C**) Performance of different rescoring pipelines on ground truth data from eight MS files (8,631 target PSMs, 3,293 decoy PSMs). Dash line represents a precision of 0.99, which approximately corresponds to 1% FDR.

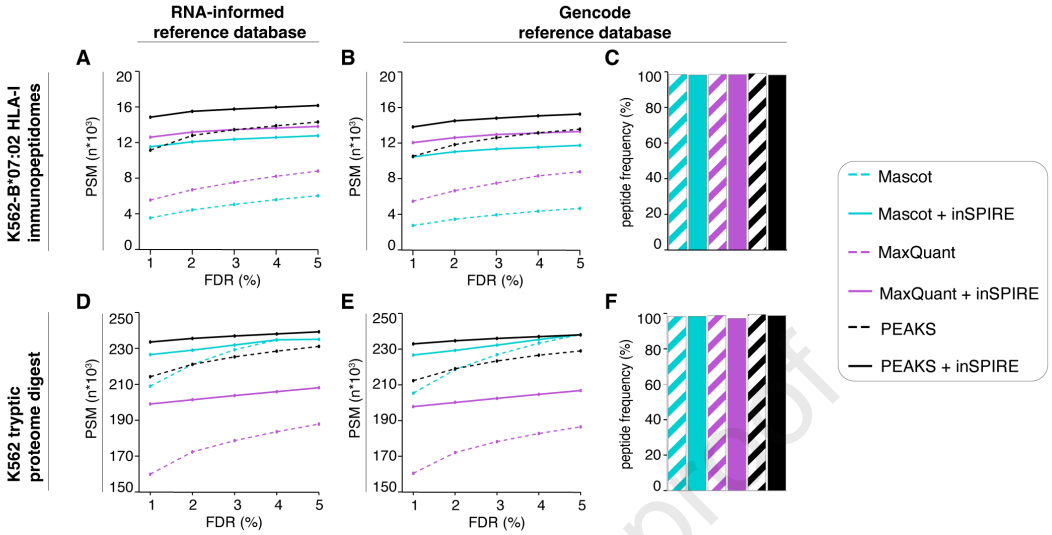
Figure 4. inSPIRE increases the PSM yield preferentially benefiting MaxQuant searches. (**A-F**) Analysis of the PSMs identified in either the HLA-I immunopeptidomes of K562-B*07:02 cell line (1 biological and 3 technical replicates; **A-C**) or the tryptic proteome digestion of the K562 cell line (3 biological and 2 technical replicates; **D-F**) by different pipelines. In (**A,D**) the RNA-informed and in (**B**,**C,E,F**) the full Gencode reference databases have been used. (**A,B,D,E**) Number of PSMs identified by the three search engines with or without inSPIRE rescoring. (**C,F**) The percentage of peptides identified by each pipeline on search results using the Gencode reference database, which were also found at any confidence level in the MaxQuant search of the RNA-informed database for the two datasets.

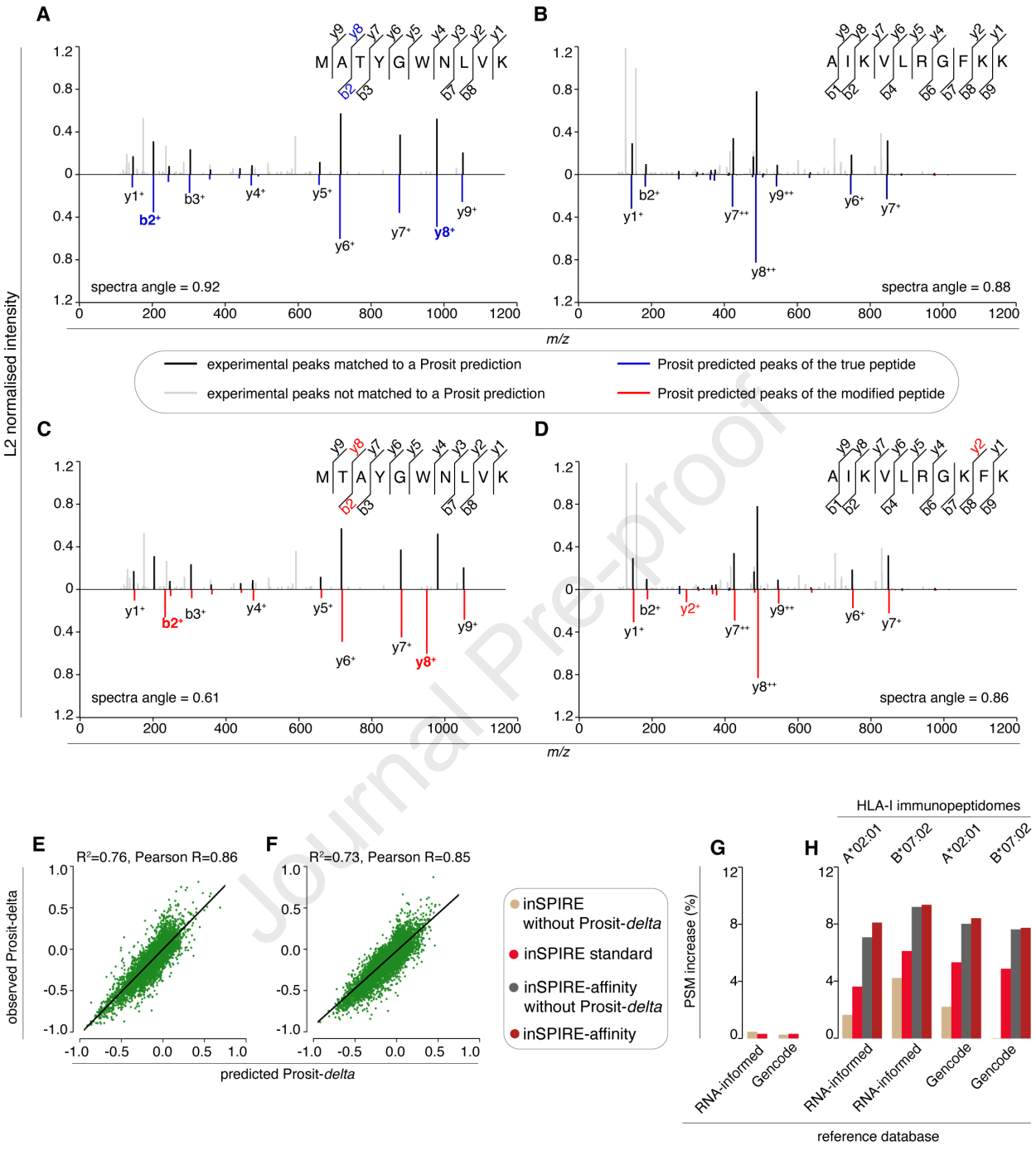
Figure 5. Prosit-*delta* feature and its impact on inSPIRE's PSM yield in HLA-I immunopeptidomics. (**A-D**) Prosit predicted MS2 spectra compared to experimentally measured MS2 spectra in representative cases wherein a switch of an amino acid residue pair results in either a large (**A,C**) or small (**B,D**) Prosit-*delta*. In each case, the pair plot is shown for the identified peptide (**A,B**) and for the peptide produced by the permutation of two adjacent amino acids in the original sequence (**C,D**). The peptide in (**A**) was identified in the synthetic peptide library sample SPL4-2 (scan number 11061). The peptide in (**B**) was identified in the synthetic peptide library sample SPL3-2 (scan number 15058). (**E**) Scatter plot of observed Prosit-*delta* values of PSMs in the search results of the K562-A*02:01 and -B*07:02 HLA-I immunopeptidomes against our model's predictions. The data is down-sampled to 10% of the original data to make the figure clearer. (**F**) Scatter plot of observed Prosit-*delta* values of PSMs in the search results of the K562 tryptic proteome digestion against our model's predictions. The data is down-sampled to 10% of the original data to make the figure clearer. (**G,H**) Relative increase in the number of identified PSMs by inSPIRE using different features as compared to Prosit rescoring pipeline. The analysis refers to the K562 tryptic proteome digestion (**G**) and the K562-B*07:02 and -A*02:01 HLA-I immunopeptidomes (**H**) searched by MaxQuant using either the RNA-informed or the full Gencode reference databases, and 1% FDR.











- inSPIRE (*in silico* Spectral Predictor Informed REscoring)
- inSPIRE is a flexible and performant open-source rescoring pipeline
- inSPIRE allows large scale rescoring with multiple MS search files
- inSPIRE can be applied to various search engines
- inSPIRE has better performance than the original Prosit rescoring pipeline

Author contribution

JAC, MM and JL developed the project, performed and/or supervised the data analysis and data generation and wrote the manuscript. YH performed the RNA sequencing data analysis and database generation. WTS performed the performed the trypsin digestions, measured the cognate samples and proofread the manuscript.

inSPIRE (*in silico* Spectral Predictor Informed REscoring) is a flexible and performant open-source rescoring pipeline built on Prosit or MS^{2}PIP MS spectral prediction. inSPIRE is compatible with several search engines, allows large scale rescoring with data from multiple MS search files, enables Prosit prediction without specialized GPU hardware, increases sensitivity to minor differences in amino acid residue position, and can be applied to various MS sample types, including tryptic proteome digestions but is specifically optimized for immunopeptidomes.