MS²Rescore: Data-driven rescoring dramatically boosts immunopeptide identification rates

Arthur Declercq[1,2], Robbin Bouwmeester[1,2], Aurélie Hirschler[3], Christine Carapito[3], Sven Degroeve[1,2], Lennart Martens[1,2 §], and Ralf Gabriels[1,2]

[1] VIB-UGent Center for Medical Biotechnology, VIB, Belgium
[2] Department of Biomolecular Medicine, Ghent University, Belgium
[3] Laboratoire de Spectrométrie de Masse BioOrganique (LSMBO), Université de Strasbourg, CNRS

[§] To whom correspondence should be addressed
Tel: +32 9 264 93 58
Email: lennart.martens@vib-ugent.be
Address: Technologiepark 75, 9052 Ghent, Belgium

ORCID IDs:
Arthur Declercq: https://orcid.org/0000-0002-9376-1399/
Robbin Bouwmeester: https://orcid.org/0000-0001-6807-7029/
Aurélie Hirschler: https://orcid.org/0000-0001-5066-6263
Christine Carapito: https://orcid.org/0000-0002-0079-319X
Sven Degroeve: https://orcid.org/0000-0001-8349-3370/
Lennart Martens: https://orcid.org/0000-0003-4277-658X/
Ralf Gabriels: https://orcid.org/0000-0002-1679-1711/

Abbreviations

CE, collision energy

Da, dalton

FDR, false discovery rate

HCD, higher-energy collision-induced dissociation

HLA, human leukocyte antigen

LC-MS, liquid chromatography-mass spectrometry

MS2, tandem mass spectrometry

MHC, major histocompatibility complex

MGF, mascot generic format

PCC, Pearson correlation coefficient

PSM, peptide-spectrum match

USI, universal spectrum identifier

Abstract

Immunopeptidomics aims to identify Major Histocompatibility Complex-presented peptides on every cell that can be used in anti-cancer vaccine development. However, existing immunopeptidomics data analysis pipelines suffer from the non-tryptic nature of immunopeptides, complicating their identification. Previously, peak intensity predictions by MS²PIP and retention time predictions by DeepLC, have been shown to improve tryptic peptide identifications when rescoring peptide-spectrum matches with Percolator. However, as MS²PIP was tailored towards tryptic peptides, we have here retrained MS²PIP to include non-tryptic peptides. Interestingly, the new models not only greatly improve predictions for immunopeptides, but also yield further improvements for tryptic peptides. We show that the integration of new MS²PIP models, DeepLC, and Percolator in one software package, MS²Rescore, increases spectrum identification rate and unique identified peptides with 46% and 36% compared to standard Percolator rescoring at 1% FDR. Moreover, MS²Rescore also outperforms the current state-of-the-art in immunopeptide-specific identification approaches. Integration of immunopeptide MS²PIP models, DeepLC, and Percolator into MS²Rescore thus allows substantial improved identification of novel epitopes from existing immunopeptidomics workflows.

Keywords: bioinformatics; immunopeptidomics; machine learning; proteomics; mass spectrometry; peptide identification

Introduction

The immune system is a complex, yet remarkable system that protects us from both invaders from outside the body, i.e. pathogens, as well as from inside the body, i.e. malignancies (1). Increased understanding of the immune system allowed for great medical achievements such as vaccination, which is currently available for over 29 diseases, enabled the eradication of smallpox, and prevents over 3 million deaths each year (2). However, many diseases such as *Mycobacterium tuberculosis* or malignancies lack effective vaccines due to improper T-cell activation. A key issue in developing effective vaccines for these diseases is the lack of accurately identified major histocompatibility complex (MHC)-presented epitopes or immunopeptides. These epitopes are presented on the cell surface and enable T-cells to discern healthy cells from infected or malignant cells. While much effort has recently been invested in accurate prediction of these epitopes *in silico* (3), these are mostly limited to viruses as these contain fewer potential protein antigens (4). Moreover, these tools are not yet sufficiently precise to confidently identify epitopes (5, 6). Therefore, experimental immunopeptidomics workflows, such as epitope detection through liquid chromatography-mass spectrometry (LC-MS), are still the best way to accurately identify these immunopeptides (7).

While immunopeptidomics workflows have been readily developed and applied (8), acquisition of immunopeptides through LC-MS suffers from some major problems. First, the acquisition of immunopeptide spectra is hampered due to the low abundance of immunopeptides and even more so of neo-epitopes. Infrequently occurring epitopes are still very challenging to identify through LC-MS, despite enrichment efforts and sample preprocessing (9). Second, in contrast to standard proteomics experiments where proteins are usually digested with trypsin before LC-MS, immunopeptides are captured through immune purification with antibodies followed by acidic elution resulting in mostly non-tryptic peptides. The non-tryptic nature of immunopeptides results in one less positive charge due to the missing arginine or lysine at the peptide's C-terminus, causing many immunopeptides to be singly charged during MS acquisition. These singly charged peptides are much harder to analyze because, during fragmentation of the peptide, the charge resides on one of the fragments, leaving the other uncharged and thus lost (10). Moreover, most contaminants are singly charged as well, making identifications of immunopeptides much harder (11). The non-tryptic nature of immunopeptides hampers not only the acquisition but also the identification of immunopeptide spectra. To match each acquired spectrum with the peptide from which the spectrum originated, proteomics database search engines such as SEQUEST (12), X!Tandem (13), Andromeda (14), or PEAKS DB(15) generate *in silico* spectra for all potentially matching peptides. The complete list of peptides that could be in the sample is called the search space. It is important to note that spectra from peptides that are not included in this search space cannot be identified, even though they were acquired. In standard shotgun proteomics experiments, *in silico* tryptic digestion of relevant proteins yields a broad yet representable search space. In immunopeptidomics, however, the search space tends to be two orders of magnitude larger due to (i) seemingly random cleavage from the protein of origin, (ii) the variable length of MHC class I bound peptides, 8-11 amino acids, and MHC-II peptides, 6-24 amino acids (16) and (iii) the potential occurrence of conformational cis- and trans-spliced immunopeptides,

which are non-linear peptides that originate from the same or different proteins, respectively (17). Additionally, sequence variants and non-canonical protein sequences are often considered as well, even further increasing the search space. Such a search space expansion leads to considerably more ambiguity between candidate peptide-spectrum matches (PSMs) (18), lower PSM scores, drastically elevated false discovery rate (FDR) score thresholds, and ultimately in fewer identified immunopeptides (19). Furthermore, because tryptic peptides have been the longtime standard in proteomics, search engines as well as bioinformatics tools that aid in identifying LC-MS spectra, are tailored towards tryptic peptides, making them less accurate or not applicable at all for immunopeptidomics.

The high need for neo- and xeno-epitope discovery led to the development of many bioinformatics tools to improve or validate identifications in immunopeptidomics. On the one hand, motif deconvolution tools have been developed that leverage binding motifs of immunopeptides to validate immunopeptide identifications. On the other hand, full pipelines have been developed to improve immunopeptide identification. For example, MHCquant (20), which is a recent computational workflow designed specifically for neo-epitope identification, and PEAKS DB (15). Even though PEAKS DB is not specifically designed for immunopeptides, it is highly interesting due to its *de novo* assisted database searches, which tend to work well for large search spaces. Even though these tools can help with immunopeptide identification, they do not use all available information, such as retention time and fragment ion intensity patterns. Previously, it has been proven that integrating retention time predictions in standard proteomics workflows can improve identification rates (21). Similarly, adding peak intensity predictions to post-processing tools such as Percolator can also improve identification rates drastically (22), which has already been proven to work for immunopeptides as well by efforts such as Prosit (23, 24). Similarly, tools such as DeepLC (25) and MS²PIP (26–28) can provide accurate retention time predictions and peak intensity predictions, respectively, to aid in post-processing. Indeed, when combined with Percolator, identification rates at a fixed FDR have been proven to substantially increase (22). However, currently, DeepLC and MS²PIP are solely trained on tryptic peptides. This absence of lysine and arginine at the C-terminus is less of a problem for DeepLC as the effect on retention time is small and is accounted for through feature encoding (29). However, this is not the case for MS²PIP, as alterations in peptide composition, as well as fragmentation patterns and labeling methods heavily alter peak intensity patterns (26). Therefore, we here present greatly improved MS²PIP models that include immunopeptides, and non-tryptic peptides in general. Moreover, we have integrated MS²PIP and DeepLC with Percolator into the free and open-source MS²Rescore software tool, which enables improved rescoring of peptide identifications from various proteomics search engines. Altogether, we show that well-adapted fragmentation spectrum and retention time predictions integrated into MS²Rescore drastically increase immunopeptide identification rates and outperform existing post-processing methods.

Experimental Procedures

Training and evaluation of new MS²PIP spectrum prediction models for immunopeptides

To train and test new MS²PIP models, five publicly available immunopeptide data sets, and one publicly available chymotrypsin-digestion data set were downloaded from PRIDE Archive (30, 31). Similarly, for evaluation on representative unseen data, four distinct data sets were downloaded: (i) a data set containing HLA-I immunopeptides, (ii) a data set containing HLA-II immunopeptides, (iii) the data set with tryptic peptides that was previously used to evaluate the existing MS²PIP HCD models, and (iv) a data set containing chymotrypsin-digested peptide data. The corresponding ProteomeXchange project identifiers as well as the number of unique peptides and HLA patterns for each data set are listed in Supplementary Table S1.

As tandem mass spectrometry (MS2) fragmentation patterns are highly dependent on the instrument, instrument settings, fragmentation method, and any applied labeling methods (26), all MS²PIP train, test, and evaluation data must originate from experiments with the same experimental parameters. Unlabeled higher-energy collision-induced dissociation (HCD) data from Quadrupole-Orbitrap instruments was used, as this makes the newly trained models widely applicable and plenty of training data is readily available on public repositories. For each PRIDE Archive project, the raw mass spectrometry files were converted to Mascot Generic Format (MGF) files using ThermoRawFileParser (v1.3.4) (32). The corresponding identification files were converted to MS²PIP input files using custom Python scripts and were further filtered to retain unique combinations of peptide sequence, modifications, and precursor charge at 1% FDR. Next, all spectra were combined into one MGF file. The universal spectrum identifier (USI) (33) was used as unique identifier for each PSM to ensure reproducibility and a one-on-one mapping between peptide identifications and spectra. Data from each PRIDE Archive project was either used as train/test data or as evaluation data to ensure fully independent data sets, except for the chymotrypsin data, where the same project was used to provide both training/testing data (70%) and evaluation data (30%). This split was made after selecting unique peptide-modification-charge combinations, to ensure no overlap in samples between both splits.

Similarly to the 2019 MS²PIP models, new models were trained with the XGBoost machine learning algorithm (34). The Hyperopt (35)' package (v0.2.5) was used in combination with four-fold cross-validation for hyperparameter optimization, allowing 400 boosting rounds and early stopping fixed at 10 rounds. Hyperparameters were optimized for each training data set separately, as well as for b- and y-ion models. All selected hyperparameters are shown in Supplementary Table S2. To evaluate each model, the Pearson correlation coefficient (PCC) was calculated between observed and predicted b- and y-ion peak intensities for each spectrum. The model performances were further analyzed by peptide length and precursor charge. Ultimately, three models were trained: the *immunopeptide* model solely trained on immunopeptides, the *immunopeptide-chymotrypsin* model trained on immunopeptides supplemented with chymotrypsin-digested peptides, and the *non-tryptic immunopeptide model* solely trained on non-tryptic immunopeptides. Ultimately, two models were

integrated into MS²PIP: (i) the *immunopeptide* model and (ii) the *immuno-chymotrypsin* model. The former can be used for immunopeptide peak intensity predictions and the latter for tryptic and more general non-tryptic peptide predictions. For further analysis into rescoring immunopeptide PSMs, only the *immunopeptide* model was used.

To compare MS²PIP predictions with Prosit predictions, the same evaluation data sets were used as mentioned above. Prosit (v1.1.2) was downloaded from GitHub (https://github.com/kusterlab/prosit) and the hcd_hla and irt_prediction models were downloaded from Figshare (https://figshare.com/projects/prosit/35582). MS²PIP predictions were acquired for the general proteomics and chymotrypsin evaluation data with the immuno-chymotrypsin model, and for the HLA-I and HLA-II evaluation data with the immunopeptide model. Peptides that were not included in the Prosit output were filtered out of the MS²PIP predictions. The performance was measured in both PCC and spectral angle to ensure a thorough comparison. Only correlations for singly charged fragment ions were taken into account, as the newly trained MS²PIP models only predict intensities for these ions.

Evaluation of MS²Rescore on HLA class I peptides and comparison with Prosit rescoring

To validate the capacity of the new MS²PIP models to improve immunopeptide identification rates, the new models were implemented with DeepLC (v0.1.36) and Percolator (v3.5) into MS²Rescore. MS²Rescore calculates various meaningful features based on (i) the search engine output, (ii) the DeepLC-predicted and the observed retention times, and (iii) the MS²PIP-predicted and the observed MS2 peak intensities. These features are then passed to Percolator for PSM rescoring. Search engine features were selected based on the previous publication by Granholm et al. (36) and replicated for use with MaxQuant search results (37). MS²PIP features were used as first described by Silva et al (22). All features generated by MS²Rescore are listed in Supplementary Table S3.

MS²Rescore was validated on a large-scale HLA class I data set (38) which was also used to validate the recently published Prosit rescoring effort for immunopeptides (PXD021398) (24). This allows both an evaluation of the improved identification rates due to the new MS²PIP models, and a straight-forward comparison with Prosit rescoring. First, the msms.txt identification files for the projects' two MaxQuant searches (alkylated and non-alkylated samples), the Prosit-rescored Percolator output files, and the raw mass spectrometry files were downloaded from PRIDE Archive. The mass spectrometry files were then further processed with ThermoRawFileParser (v1.3.4) (32) and the PSMs for each of the two MaxQuant searches were rescored separately. Two rescoring methods were evaluated: (i) Using only search engine features, replicating a normal Percolator run, and (ii) using the full MS²Rescore feature set, including search engine-, MS²PIP-, and DeepLC-features. Additionally, these rescoring methods were compared with the original MaxQuant results and with the downloaded Prosit-rescoring results.

Each rescoring method was evaluated at varying FDR thresholds in terms of identification rate and number of unique identified peptides. The contribution of the different feature sets in MS²Rescore was visualized using Percolator's model weights, and the distributions of retention time difference and MS²PIP prediction correlations were compared between decoy PSMs, accepted target PSMs, and rejected target PSMs.

Additionally, as reported by Wilhelm et al. (24), sequence motif patterns for HLA pattern C*12:03 were further analyzed with GibbsCluster (v2.0) (39), for the gained and lost peptides compared to rescoring with only search engine features.

Evaluation of MS²Rescore across collision energy settings and peptide abundances

To further analyze MS²Rescore performance for various experimental collision energy settings, replicate LC-MS/MS runs were performed on HL60 cells at collision energy values of 25, 27, 30, 32 and 35 NCE (Supplementary Methods). The resulting spectra were searched with the Andromeda search engine (MaxQuant v1.6.14.0) against the human UniProtKB-SwissProt (14-09-2020; 20 388 sequences, Taxonomy ID 9606) database without any enzyme specificity. A minimal peptide length of seven amino acids was required. Oxidation (M) was set as variable modification with a maximum of three modifications per peptide. Mass tolerances were set at 5 ppm and 20 ppm for MS1 and MS2 spectra respectively. False discovery rate was kept at 100% with the use of a decoy strategy for downstream rescoring with (i) only search engine features, and (ii) the full MS²Rescore feature set. Furthermore, for all LC-MS/MS runs at all collision energy settings, precursor intensities were obtained from the MaxQuant msms.txt file to assess any differences in the performance of MS²Rescore between low and high abundant peptides.

Evaluation of MS²Rescore on HLA class II peptides

To validate MS²Rescore for HLA class II peptides, another set of raw mass spectrometry files were downloaded from PRIDE Archive (PXD015408). As the uploaded search engine results were already filtered at 5% FDR, the spectra were re-analyzed with PEAKS DB (v10.5) (15) with the same search parameters that were used in the original publication, i.e. no enzyme specificity, precursor error tolerance of 10 ppm, fragment ion tolerance of 0.01 Da, with oxidation (M), deamidation (NQ) and trioxidation (C) as variable modifications, searched against the UniProtKB-SwissProt database (01-2021, 22 235 sequences, Taxonomy ID 10090). The mzIdentML identification file as well as the corresponding MGF files were exported from PEAKS DB and were rescored with MS²Rescore with (i) only search engine features, and (ii) the full MS²Rescore feature set, as described above for the evaluation on HLA class I peptides.

General data processing and data visualization

All plots, unless specified, were generated in Jupyter notebooks (v6.4.0) using Python (v3.8.3) with the Matplotlib (v3.4.2) (40), Seaborn (v0.11.0) (41), UpSetPlot (v0.6.0), and spectrum_utils (v0.3.5) (42) libraries.

Results

Newly trained MS²PIP models accurately predict immunopeptide spectrum peak intensities

In order to improve the identification rate of immunopeptides by leveraging peak intensity predictions, new models for MS²PIP were trained specifically for immunopeptides. Despite using different training set compositions, all newly trained models drastically improve predictions for both HLA-I and HLA-II data in comparison with the tryptic 2019 HCD model (Figure 1A). Surprisingly, even for standard tryptic shotgun proteomics data the predictions from the new models are slightly better, largely due to the portion of tryptic peptides within the immunopeptide training data. Indeed, when these peptides are left out of the training data, accuracy drops in comparison with the 2019 HCD model. While both immunopeptide models are well suited to predict peak intensities for tryptic- and immunopeptides, the performance on chymotrypsin-digested peptides is not as high. Only when chymotrypsin-digested peptides are included in the training data, performance on chymotrypsin data increases. This difference in performance between models trained with or without chymotrypsin-digested peptides, is mainly seen for longer peptides and peptides with higher precursor charge states (Supplementary Figure S1). As models that are solely trained on immunopeptides have not seen longer peptides with higher charge-states, these models can not accurately model how these will behave in the mass spectrometer. Thus, even though immuno- and chymotrypsin-digested peptides are both considered non-tryptic, they are still very different. Overall, immunopeptide peak intensity predictions are drastically improved by all the newly trained models, with the immunopeptide model showing the highest accuracy (median PCC of 0.94). The exact median PCCs are listed in Supplementary Table S4. Examples of a prediction with median PCC values with the immunopeptide model and the corresponding, less accurate, 2019 model prediction are shown in Figure 1B and Figure 1C.

MS²Rescore drastically improves immunopeptide identification rates

Ultimately, the goal of these newly trained models is to improve immunopeptide identification rates by providing more accurate peak intensity predictions. Therefore, (i) identification results without rescoring, (ii) rescoring with solely search engine features (replicating a normal Percolator run), (iii) rescoring with MS²Rescore (including DeepLC and the new MS²PIP models), and (iv) rescoring with the recently published Prosit models were compared in terms of the total amount of identifications as well as the number of unique identifications based on sequence. Overall, rescoring with both MS²Rescore and Prosit substantially improved the spectrum identification rate in comparison with rescoring with search engine features alone or not rescoring, and this at both 1% and 0.1% FDR. Indeed, MS²Rescore achieves an identification rate of 11,1%, out of 18 million spectra, compared to 7.6% for traditional rescoring (an increase of 46%), and only 1.9% for the MaxQuant search results, all at 1% FDR (Figure 2A-B). Moreover, 83% of the identified spectra at 1% FDR are retained when restricting the threshold to 0.1% FDR. Thus, providing peak intensities and retention time predictions to Percolator substantially increases the number of identified immunopeptides. This is clearly illustrated by analyzing the Percolator weights for each separate feature as well as the combined

absolute weights for search engine features, MS²PIP features and DeepLC features (Supplementary Figure S2). Similarly, the number of unique identified immunopeptides increases by 36% when adding MS²PIP and DeepLC Features for the 1% FDR, and even more so for 0.1% FDR where the number of unique identified peptides reaches nearly 300% of the number of traditional Percolator identification results (Figure 2C-F). These gains are consistent across all 95 HLA class I alleles included in the data (Supplementary Figure S3-4), showing that the newly trained MS²PIP model, and therefore MS²Rescore, is generalizable across different HLA types. In fact, MS²Rescore allows for a substantial increase in identification rate for HLA types with initially fewer identifications (e.g., A0101, A0204, B4402…), indicating that MS²Rescore especially improves the peptide identification coverage for harder-to-identify HLA alleles.

The power of providing these predictions to Percolator is further illustrated when visualizing the distributions for decoy PSMs, rejected target PSMs, and accepted target PSMs. Indeed, the distributions for decoy and rejected target PSMs are highly similar for both the retention time error as well as the PCC, while the accepted target PSMs accumulate around low retention time errors and high PCCs (Figure 3A-B). The accepted target PSMs are clearly separable from the decoy and rejected target PSMs using only the PCC and retention time error distributions (Figure 3C-D). Furthermore, while both metrics correlate with the search engine score, a large amount of decoy and rejected target PSMs can only be separated from the target PSMs by also including PCC or retention time error information (Figure 3E-F). This clearly illustrates how Percolator achieves its much-improved separation between true and false target PSMs when provided with peak intensity and retention time prediction features. Furthermore, PSMs that previously would have been incorrectly accepted below a 1% FDR because of a high search engine score alone, are now rejected due to a low PCC, a high retention time error, or both. This most likely accounts for the small percentage of identified peptides that are lost after rescoring.

MS²Rescore outperforms the current state-of-the-art

The integration of MS²PIP-, DeepLC-, and search engine-based features in MS²Rescore has proven to substantially increase the identification rate of immunopeptides, and furthermore, outperforms the recently published Prosit rescoring method (24). In comparison with Prosit, MS²Rescore gains 5% and 35% more identifications at 1% FDR and 0.1% FDR, respectively. This trend continues for the number of unique identified peptides with a respective increase of 8% and 57% (Figure 2). Indeed, over 32 000 unique peptides were identified below the 0.1% FDR threshold by MS²Rescore while Prosit rescoring only identified these peptides at the 1% FDR threshold, highlighting the gain in confidence in terms of unique identified immunopeptides (Supplementary Figure S5). MS²Rescore thus substantially increases the identification rate, especially for more stringent FDR thresholds.

Peptides for which Prosit cannot predict MS2 peak intensities, i.e. unmodified (C), cysteinylated (C), and acetylation (N-terminus), were left out of the Prosit rescoring output. That is why MS²Rescore includes more PSMs in the unfiltered data set at 100% FDR, especially for the noIAA sample (Supplementary Figure S6A-B). For a second, thorough comparison, these PSMs were left out of the

MS²Rescore output as well. However, this seemed to have a rather negligible impact on the number of identifications at 1% and 0.1% FDR (Supplementary Figure S6C-D), and thus the difference in identification rates cannot be attributed to these filtered peptides.

Furthermore, a comparison of the peptide spectrum prediction accuracy of the newly trained MS²PIP models and Prosit indicates that higher performance of MS²Rescore cannot be attributed to improved peak intensity predictions. Indeed, depending on the correlation metric that is used, either MS²PIP or Prosit performs slightly better than the other on the evaluation data (Supplementary Figure S7). It is important to note, however, that for Prosit the correct CE value should be selected for optimal performance. To determine if the difference in rescoring between MS²Rescore and Prosit is driven by a difference in features, all MS²Rescore features that do not have a close counterpart in Prosit rescoring were removed for a separate run. This includes all DeepLC retention time features and various search engine features (Supplementary Table S3). As MS²Rescore and Prosit both include various similar peak intensity prediction features, these were retained. With this reduced feature set, the performance of MS²Rescore is slightly lower than for Prosit rescoring (Supplementary Figure S6A-D), which confirms that the retention time features and additional search engine features result in improved MS²Rescore performance over Prosit rescoring.

Similarly to Prosit, the identified sequence motif for HLA type C*12:03 was highly similar to the motif reported in the original publication of the data set (24, 38), while the peptides that were removed by MS²Rescore compared to search engine rescoring showed quite different, less conserved sequence motifs (Supplementary Figure S8).

MS²Rescore generalizes well across collision energy settings and peptide abundances

Because MS²PIP does not account for CE in its predictions, MS²PIP and consequently MS²Rescore could potentially be biased towards spectra obtained with certain CE values. Therefore, search results of replicate mass spectrometry runs with CE settings varying from 25 to 35 were post-processed with MS²Rescore. For each CE value, MS²Rescore shows a significant increase in identification rate. However, for larger (less optimal) CE values, the overall identification rate decreases (Supplementary Figure S9A). This is most likely due to a reduced quality in fragmentation spectra, which is reflected in the decreased explained ion current, and in line with the b- and y-ion MS²PIP PCC distributions for accepted PSMs when using suboptimal CE values (Supplementary Figure S9D-F). Most interestingly, the relative gain in unique identified peptides for MS²Rescore increases for higher (and therefore less optimal) CE values, approaching a 60% increase for CE 35, by slightly shifting the feature weights away from fragmentation features in favor of DeepLC retention time features (Supplementary Figure S9B-C). Consequently, MS²Rescore is able to recover peptides that would otherwise be lost due to lower-quality fragmentation spectra.

A similar effect is observed for low abundant peptides. Indeed, the largest relative gain achieved by MS²Rescore in terms of number of unique identified peptides is seen for the lowest precursor

intensities (Supplementary Figure S10A-B), where traditional rescoring fails to recover most identifications. MS²Rescore is thus not only able to increase the amount of identifications for immunopeptides in general, it can recover peptides previously lost due to low precursor intensities and thus lower-quality spectra or nonoptimal instrument settings.

MS²Rescore is unbiased to different HLA classes

To evaluate the performance of MS²Rescore on HLA class II peptides, another publicly available data set was re-analyzed. However, while for the HLA class I data set human immunopeptides and MaxQuant search engine (14) results were used, for this HLA class II data set, mouse data was searched with PEAKS DB (15). As was the case for HLA class I peptides, MS²Rescore significantly increases the identification rate for HLA class II peptides with 15% and 57% for the 1% and 0.1% FDR threshold, respectively (Supplementary Figure S11). These increases are, however, slightly lower than for the HLA class I data set. Moreover, where previously conventional rescoring showed a significant increase in comparison to search engine rescoring, here the gain in comparison with no rescoring is lower for both identification rate as well as number of unique identified peptides. This is likely due to (i) the less extensive search engine features that are calculated for the PEAKS DB pipeline in MS²Rescore (Supplementary Table S3), and (ii) to the fact PEAKS DB is likely better equipped to identify immunopeptides than MaxQuant due to its *de novo*-assisted database search (15). Nevertheless, the full MS²Rescore feature set, including peak intensity and retention time predictions, still results in a significantly higher identification rate. Altogether, these results show that MS²Rescore generalizes well across HLA class I and class II immunopeptides, across different species, and can boost the performance from different search engines.

Discussion

By training new peak intensity prediction models, we were able to greatly enhance immunopeptide identification rate through PSM rescoring. While all newly trained models greatly enhance peak intensity predictions for immunopeptides, the model trained solely on immunopeptides performed best. Even though the immuno-chymotrypsin model contained the same immunopeptide train set, the addition of the chymotrypsin-digested peptides did lower the performance slightly. Similarly, not including chymotrypsin-digested peptides in the training data resulted in lower accuracies for the chymotrypsin-digested peptides. Indeed, immunopeptides are generally much smaller and consequently carry a lower charge state. Therefore, immunopeptide trained models are not able to predict the behavior of longer and higher charged peptides in the mass spectrometer. While both immuno- and chymotrypsin-digested peptides are considered non-tryptic, their properties can be very different, leading to reduced accuracy of peak intensity prediction models when applied on a different type of non-tryptic peptides. Therefore, a general non-tryptic peak intensity prediction model seems not ideal, as it could lower performance for subsets like immunopeptides. Surprisingly, while immuno- and chymotrypsin-digested peptides are antagonistic, immuno- and tryptic peptides seem synergistic in terms of training data. This comes as no surprise as almost 50% of the immunopeptide training data consists of tryptic peptides. Immunopeptides are thus not necessarily non-tryptic. However, the actual occurrence of tryptic peptides in immunopeptidomics samples is most likely much lower. This unrepresentatively sized tryptic portion most likely originated from the tryptic bias in current immunopeptidomics workflows. Indeed, in previous studies, tryptic MHC peptide coverage could rise to 70% (43). By training new, non-tryptic models of MS²PIP, we take a first step in decreasing this tryptic bias, to ultimately be able to analyze an unbiased immunopeptide landscape.

Moreover, by integrating the new immunopeptide model with retention time predictions and search engine features into MS²Rescore, we greatly enhanced the ability of Percolator to rescore immunopeptide PSMs, resulting in a much-improved immunopeptide identification workflow. Furthermore, rescoring drastically increases the number of unique identified peptides, which is of crucial importance for the discovery of potential neo-epitopes for cancer vaccination or xeno-epitopes for anti-bacterial, and to a lesser extent, anti-viral vaccines. Moreover, while previously almost no identifications were found at a more confident 0.1% FDR threshold, MS²Rescore allows a lowering of the FDR threshold to 0.1%, while retaining 83% of the peptides identified at 1% FDR. This illustrates the large increase in confidence of the identified PSMs MS²Rescore introduces. Besides the increase in both PSM confidence and identification rate, MS²Rescore has shown to be unbiased with regard to HLA patterns and CE settings. Most importantly, the relative identification gain introduced by MS²Rescore is even larger for HLA patterns that initially had fewer identifications, showing that MS²Rescore is able to increase the view on the immunopeptide landscape for traditionally harder-to-identify HLA patterns. Moreover, MS²Rescore is able to recover peptide identifications that would have been lost due to lower quality spectra by making use of DeepLC retention time features, and can therefore recover substantial additional identifications for low abundant peptides. This potentially enables the recovery of biologically relevant neo- or xeno-epitopes that occur less frequently in the

sample. Furthermore, MS²Rescore is able to gain immunopeptide identifications regardless of the search engine used, for both HLA class I and class II peptides, and across different species.

Additionally, MS²Rescore with DeepLC and the new immunopeptide MS²PIP models shows an improved identification rate over the recently published Prosit effort, especially for lower FDR thresholds. As Prosit has shown to provide more accurate predictions compared to previous MS²PIP models, it is unlikely that MS²Rescore's higher performance can be attributed to superior peak intensity predictions. Indeed, the peak intensity prediction accuracies of the new MS²PIP models and of Prosit are highly similar for immunopeptides (Supplementary Figure S7) even when Prosit has been optimized for the right CE. These negligible differences in peak intensity prediction correlations are therefore likely not the reason for the higher performance of MS²Rescore in favor of Prosit. Instead, it is more likely that the main difference in rescoring performance is the result of the generation of more relevant MS²PIP-, DeepLC and search engine-derived features. Indeed, when the majority of the search engine features and all DeepLC retention time features were omitted, reflecting the more limited Prosit feature set, the performance of MS²Rescore drops as well (Supplementary Figure S6). By providing a more extensive feature set, MS²Rescore creates a unique feature space that allows Percolator to separate true from false identifications much better than when provided with limited features without retention time or peak intensity information (Figure 3). The combination of all these calculated features is therefore likely to be driver of MS²Rescore's superior performance.

MS²Rescore is freely available under the permissive Apache 2.0 open-source license on GitHub (https://github.com/compomics/ms2rescore) and can easily be installed locally through the cross-platform PyPI Python package, as well as with a standalone windows install script. Both a command line interface and a graphical user interface are available, various identification files from different search engines are accepted, and both MS²PIP and DeepLC can handle a variety of modifications, eliminating the need to filter identification files before rescoring. Altogether, these new models show great promise to greatly extend the immunopeptide landscape in existing and future immunopeptidomics experiments.

Acknowledgments

R.B. acknowledges funding from the Vlaams Agentschap Innoveren en Ondernemen under project number HBC.2020.2205.; R.G. received funding from the Research Foundation Flanders (FWO) [1S50918N]. S.D. and L.M. acknowledge funding from the European Union's Horizon 2020 Programme (H2020-INFRAIA-2018-1) [823839]; L.M. acknowledges funding from the Research Foundation Flanders (FWO) [G028821N] and from Ghent University Concerted Research Action [BOF21/GOA/033]. A.D. received funding from the Research Foundation Flanders (FWO) [1SE3722]. NanoLC-MS/MS instruments were supported by the French Proteomic Infrastructure (ProFI FR2048; ANR-10-INBS-08-03).

Author contributions

AD, SD, LM, and RG conceived the project. AD developed the new MS²PIP models, carried out the implementation in MS²Rescore, and performed all validations. RG developed the initial MS²Rescore Python package. RB assisted with the integration and validation of the DeepLC models into MS²Rescore. AH and CC conducted the collision energy ramp experiments. AD and RG analyzed the results. AD and RG wrote the manuscript, and all authors contributed to the final version.

Conflict of interest

The authors declare that they have no conflict of interest.

Data availability

MS²Rescore is available at https://github.com/compomics/ms2rescore. All additional code used in this work is available at compomics/ms2rescore-immunopeptidomics-manuscript (github.com). The data used for training and evaluation of the newly trained models, the models themselves as well as the MS²Rescore output is available on Zenodo at https://doi.org/ 10.5281/zenodo.6532013. Additional mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE (31) partner repository with the dataset identifier PXD033868.

References

1. Sattler, S. (2017) in *Advances in Experimental Medicine and Biology* (Springer New York LLC), pp 3–14.

2. Centers For Disease Control And Prevention (2021) Recommended Vaccines By Disease.

3. Raoufi, E., Hemmati, M., Eftekhari, S., Khaksaran, K., Mahmodi, Z., Farajollahi, M. M., and Mohsenzadegan, M. (2020) Epitope Prediction by Novel Immunoinformatics Approach: A State-of-the-art Review. *International Journal of Peptide Research and Therapeutics* 26, 1155–1163

4. Mayer, R. L., and Impens, F. (2021) Immunopeptidomics for next-generation bacterial vaccine development. *Trends in Microbiology*,

5. Larsen, M. V, Lundegaard, C., Lamberth, K., Buus, S., Lund, O., and Nielsen, M. (2007) Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction. *BMC Bioinformatics* 8, 1–12

6. Zhang, H., Lundegaard, C., and Nielsen, M. (2009) Pan-specific MHC class I predictors: a benchmark of HLA class I pan-specific prediction methods. *BIOINFORMATICS ORIGINAL PAPER* 25, 83–89

7. Bassani-Sternberg, M., Pletscher-Frankild, S., Jensen, L. J., and Mann, M. (2015) Mass spectrometry of human leukocyte antigen class i peptidomes reveals strong effects of protein abundance and turnover on antigen presentation. *Molecular and Cellular Proteomics* 14, 658–673

8. Solleder, M., Guillaume, P., Racle, J., Michaux, J., Pak, H. S., Müller, M., Coukos, G., Bassani-Sternberg, M., and Gfeller, D. (2020) Mass spectrometry based immunopeptidomics leads to robust predictions of phosphorylated HLA class I ligands. *Molecular and Cellular Proteomics* 19, 390–404

9. Faridi, P., Purcell, A. W., and Croft, N. P. (2018) In Immunopeptidomics We Need a Sniper Instead of a Shotgun. *Proteomics* 18,

10. Pfammatter, S., Bonneil, E., Lanoix, J., Vincent, K., Hardy, M.-P. P., Courcelles, M., Perreault, C., and Thibault, P. (2020) Extending the Comprehensiveness of Immunopeptidome Analyses Using Isobaric Peptide Labeling. *Analytical Chemistry* 92, 9194–9204

11. Purcell, A. W., Ramarathinam, S. H., and Ternette, N. (2019) Mass spectrometry–based identification of MHC-bound peptides for immunopeptidomics. *Nature Protocols* 14, 1687–1707

12. Eng, J. K., McCormack, A. L., and Yates, J. R. (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* 5, 976–989

13. Craig, R., and Beavis, R. C. (2004) TANDEM: Matching proteins with tandem mass spectra. *Bioinformatics* 20, 1466–1467

14. Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V., and Mann, M. (2011) Andromeda: A peptide search engine integrated into the MaxQuant environment. *Journal of Proteome Research* 10, 1794–1805

15. Zhang, J., Xin, L., Shan, B., Chen, W., Xie, M., Yuen, D., Zhang, W., Zhang, Z., Lajoie, G. A., and Ma, B. (2012) PEAKS DB: De novo sequencing assisted database search for sensitive and accurate peptide identification. *Molecular and Cellular Proteomics* 11,

16. Jiang, J., Natarajan, K., and Margulies, D. H. (2019) in *Advances in Experimental Medicine and Biology* (Springer New York LLC), pp 21–62.

17. Faridi, P., Li, C., Ramarathinam, S. H., Vivian, J. P., Illing, P. T., Mifsud, N. A., Ayala, R., Song, J., Gearing, L. J., Hertzog, P. J., Ternette, N., Rossjohn, J., Croft, N. P., and Purcell, A. W. (2018) A subset of HLA-I peptides are not genomically templated: Evidence for cis- and trans-spliced peptide ligands. *Sci Immunol* 3,

18. Colaert, N., Degroeve, S., Helsens, K., and Martens, L. (2011) Analysis of the resolution limitations of peptide identification algorithms. *Journal of Proteome Research* 10, 5555–5561

19. Verheggen, K., Ræder, 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* (John Wiley & Sons, Ltd)

20. Bichmann, L., Nelde, A., Ghosh, M., Heumos, L., Mohr, C., Peltzer, A., Kuchenbecker, L., Sachsenberg, T., Walz, J. S., Stevanović, S., Rammensee, H. G., and Kohlbacher, O. (2019) MHCquant: Automated and Reproducible Data Analysis for Immunopeptidomics. *Journal of Proteome Research* 18, 3876–3884

21. Dorfer, V., Maltsev, S., Winkler, S., and Mechtler, K. (2018) CharmeRT: Boosting Peptide Identifications by Chimeric Spectra Identification and Retention Time Prediction. *Journal of Proteome Research* 17, 2581–2589

22. C Silva, A. S., Bouwmeester, R., Martens, L., Degroeve, S., C. Silva, A. S., Bouwmeester, R., Martens, L., Degroeve, S., Silva, A. S. C., Bouwmeester, R., Martens, L., Degroeve, S., and Wren, J. (2019) Accurate peptide fragmentation predictions allow data driven approaches to

replace and improve upon proteomics search engine scoring functions. *Bioinformatics* 35, 1401–1403

23. Li, K., Jain, A., Malovannaya, A., Wen, B., and Zhang, B. (2020) DeepRescore: Leveraging Deep Learning to Improve Peptide Identification in Immunopeptidomics. *Proteomics* 20,

24. Wilhelm, M., Zolg, D. P., Graber, M., Gessulat, S., Schmidt, T., Schnatbaum, K., Schwencke-Westphal, C., Seifert, P., Krätzig, N. de A., Zerweck, J., Knaute, T., Bräunlein, 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. *Nature Communications 2021 12:1* 12, 1–12

25. Bouwmeester, R., Gabriels, R., Hulstaert, N., Martens, L., and Degroeve, S. (2021) DeepLC can predict retention times for peptides that carry as-yet unseen modifications. *Nature Methods*, 1–7

26. Gabriels, R., Martens, L., and Degroeve, S. (2019) Updated MS^2PIP web server delivers fast and accurate MS^2 peak intensity prediction for multiple fragmentation methods, instruments and labeling techniques. *Nucleic Acids Research* 47, W295–W299

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, W326-30

28. Degroeve, S., and Martens, L. (2013) MS2PIP: A tool for MS/MS peak intensity prediction. *Bioinformatics* 29, 3199–3203

29. Ruiz Cuevas, M. V., Hardy, M.-P., Holly, J., Bonneil, É., Durette, C., Courcelles, M., Lanoix, J., Côté, C., Staudt, L. M., Lemieux, S., Thibault, P., Perreault, C., and Yewdell, J. (2020) Most Non-Canonical Proteins Uniquely Populate the Proteome or Immunopeptidome. *SSRN Electronic Journal*,

30. Martens, L., Hermjakob, H., Jones, P., Adamsk, M., Taylor, C., States, D., Gevaert, K., Vandekerckhove, J., and Apweiler, R. (2005) PRIDE: The proteomics identifications database. *Proteomics*,

31. Perez-Riverol, Y., Csordas, A., Bai, J., Bernal-Llinares, M., Hewapathirana, S., Kundu, D. J., Inuganti, A., Griss, J., Mayer, G., Eisenacher, M., Pérez, E., Uszkoreit, J., Pfeuffer, J., Sachsenberg, T., Yılmaz, Ş., Tiwary, S., Cox, J., Audain, E., Walzer, M., Jarnuczak, A. F., Ternent, T., Brazma, A., and Vizcaíno, J. A. (2019) The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Research* 47, D442–D450

32. 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. *Journal of Proteome Research* 19, 537–542

33. Deutsch, E. W., Perez-Riverol, Y., Carver, J., Kawano, S., Mendoza, L., Van Den Bossche, T., Gabriels, R., Binz, P. A., Pullman, B., Sun, Z., Shofstahl, J., Bittremieux, W., Mak, T. D., Klein, J., Zhu, Y., Lam, H., Vizcaíno, J. A., and Bandeira, N. (2021) Universal Spectrum Identifier for mass spectra. *Nature Methods* 18, 768–770

34. Chen, T., and Guestrin, C. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York, NY, USA).

35. Bergstra, J., Yamins, D., and Cox, D. D. (2013) in *30th International Conference on Machine Learning, ICML 2013*, pp 115–123.

36. Granholm, V., Kim, S., Navarro, J. C. F., Sjölund, E., Smith, R. D., and Käll, L. (2014) Fast and accurate database searches with MS-GF+percolator. *Journal of Proteome Research* 13, 890–897

37. Tyanova, S., Temu, T., and Cox, J. (2016) The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nature Protocols 2016 11:12* 11, 2301–2319

38. 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. *Nature Biotechnology* 38, 199–209

39. Andreatta, M., Alvarez, B., and Nielsen, M. (2017) GibbsCluster: Unsupervised clustering and alignment of peptide sequences. *Nucleic Acids Research* 45, W458–W463

40. Hunter, J. D. (2007) Matplotlib: A 2D graphics environment. *Computing in Science and Engineering* 9, 90–95

41. Waskom, M. (2021) seaborn: statistical data visualization. *Journal of Open Source Software* 6, 3021

42. Bittremieux, W. (2020) Spectrum-utils: A Python Package for Mass Spectrometry Data Processing and Visualization. *Analytical Chemistry* 92, 659–661

43. Chen, R., Fauteux, F., Foote, S., Stupak, J., Tremblay, T. L., Gurnani, K., Fulton, K. M., Weeratna, R. D., Twine, S. M., and Li, J. (2018) Chemical Derivatization Strategy for Extending the Identification of MHC Class i Immunopeptides. *Analytical Chemistry* 90, 11409–11416

604

605

Figure captions

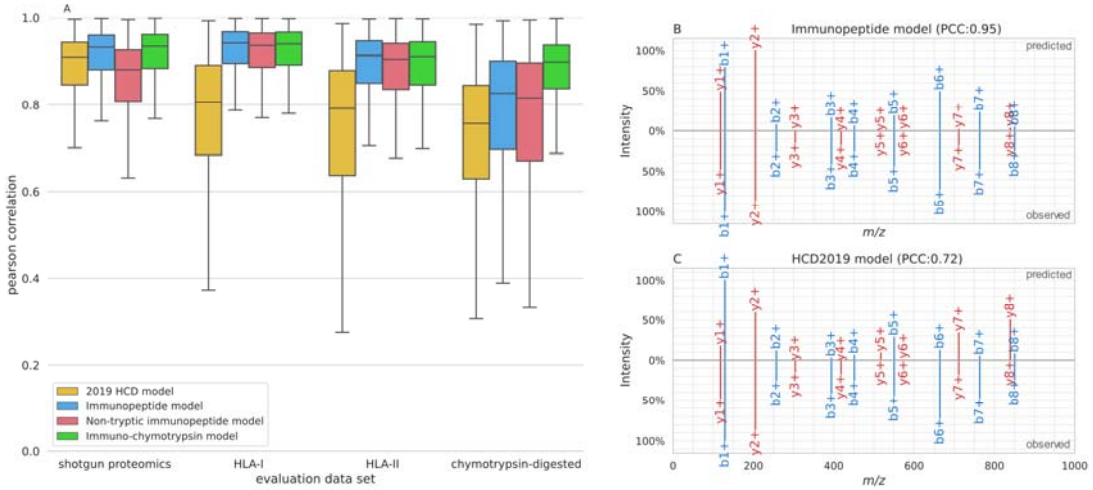

Figure 1

Performance evaluation of the new non-tryptic MS²PIP models. Boxplots comparing Pearson correlation distributions between predicted and observed peak intensity values on spectrum level for the 2019 HCD model, immunopeptide model, non-tryptic immunopeptide model, and immuno-chymotrypsin model (A). Performance is evaluated on general tryptic proteomics data, HLA-I immunopeptide data, HLA-II immunopeptide data, and chymotrypsin data. Predictions for immunopeptide “KQHGVNVSV” by new immunopeptide MS²PIP model (top) compared to observed MS2 spectrum (mzspec:PXD005231:20160513_TIL1_R1:scan:10909) (bottom) (B). This peptide was selected for visualization as its PCC lies close to the median for the immunopeptide model on all HLA-I evaluation data. Predictions by the 2019 HCD model for the same immunopeptide (C) as visualized in B.

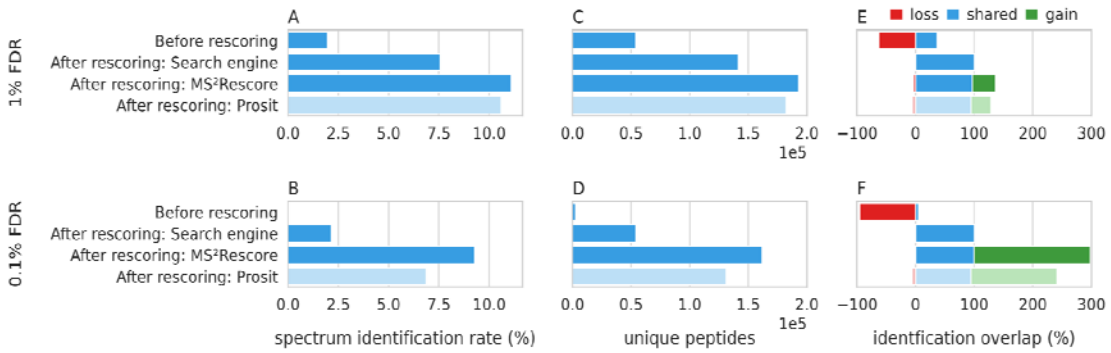

Figure 2

Percentage of identified spectra and unique identified peptides using different rescoring methods (PXD021398). Bar charts showing the spectrum identification rate out of 18.375.659 spectra (A, B), showing the total number of unique identified peptides in terms of sequence (C, D), and showing the shared (blue), gained (green) and lost (red) number of unique (by sequence) identified immunopeptides in relation to rescoring with only search engine features. for the 1% FDR (E) and 0.1%FDR (F) threshold. All results are shown for the 1% FDR (A, C, E) and 0.1% FDR (B, D, F) thresholds.

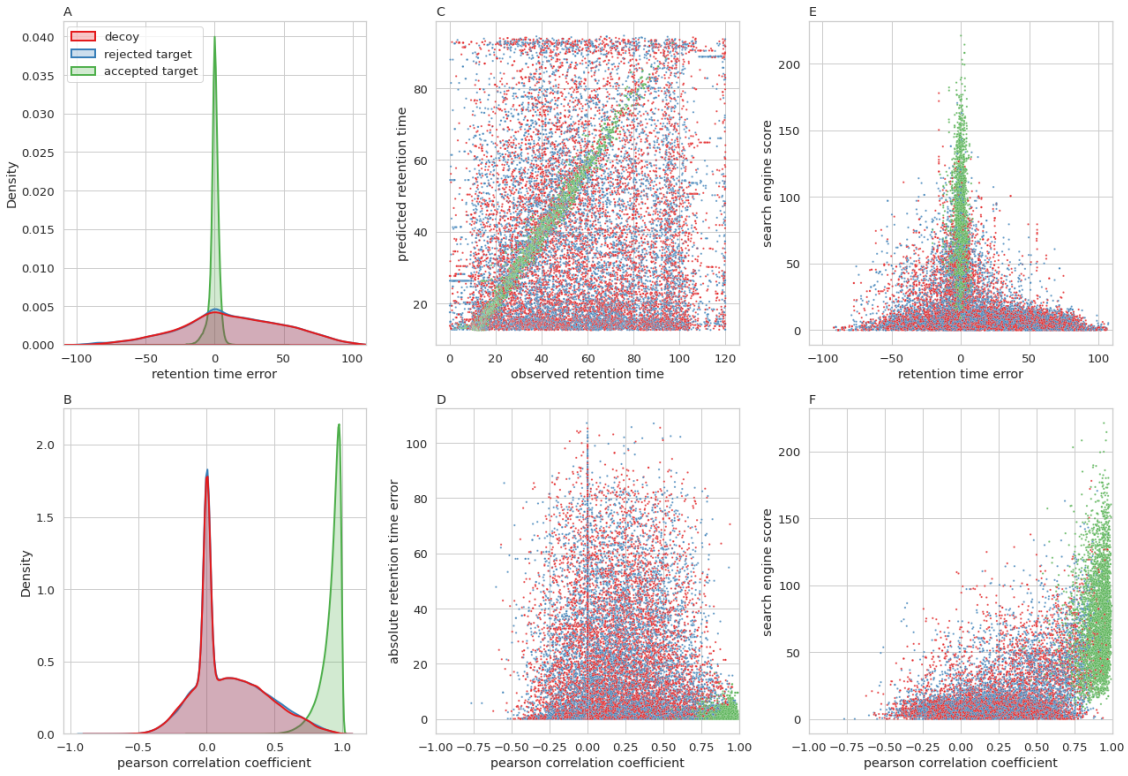

Figure 3

Distributions of MS²PIP-, DeepLC and search engine-based features. Density plots showing the distribution of the smallest retention time error between observed and predicted retention time (A) and showing the Pearson correlation between observed and predicted peak intensities (B) for each unique combination of sequence and modification for decoys (red) rejected targets, q value > 0.01(blue) and accepted targets, q value < 0.01 (green) (note the rejected targets distribution coincides with the decoy distribution).

Scatterplots showing the relation between observed and predicted retention time (C), between Pearson correlation and retention time error (D), between retention time error and search engine score (E), and between Pearson correlation and search engine score (F) for decoys (red), rejected targets (blue), and accepted targets (green) for all PSMs from “GN20170531_SK_HLA_C0102_R1_01”. Note that all retention time errors are the smallest retention time error for each precursor (see rt_diff_best table S3), all Pearson correlation coefficients are calculated for each full log2 transformed spectrum, and all zero values are caused by the fact that either the observed or predicted intensities for a given ion type are all zero.