A Flexible Learning Infrastructure for Proteomics

Christopher S Wilkins[1], Justice Sefas[1], Aivett Bilbao[1], Richard D Smith[1], Ljiljana Pasa-Tolic[2], Samuel H Payne[1], Jared B Shaw[2]
[1]Biological Sciences Division, Pacific Northwest National Laboratory, [2]Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Overview

- FLIP is a modular, computational framework for developing and optimizing tandem mass spectrometry (MS/MS) scoring models.
- FLIP enables rapid optimization for new MS/MS methods and experimental conditions that change fragmentation propensities.
- Designed to be reused and expanded for integration with new MS/MS identification software.

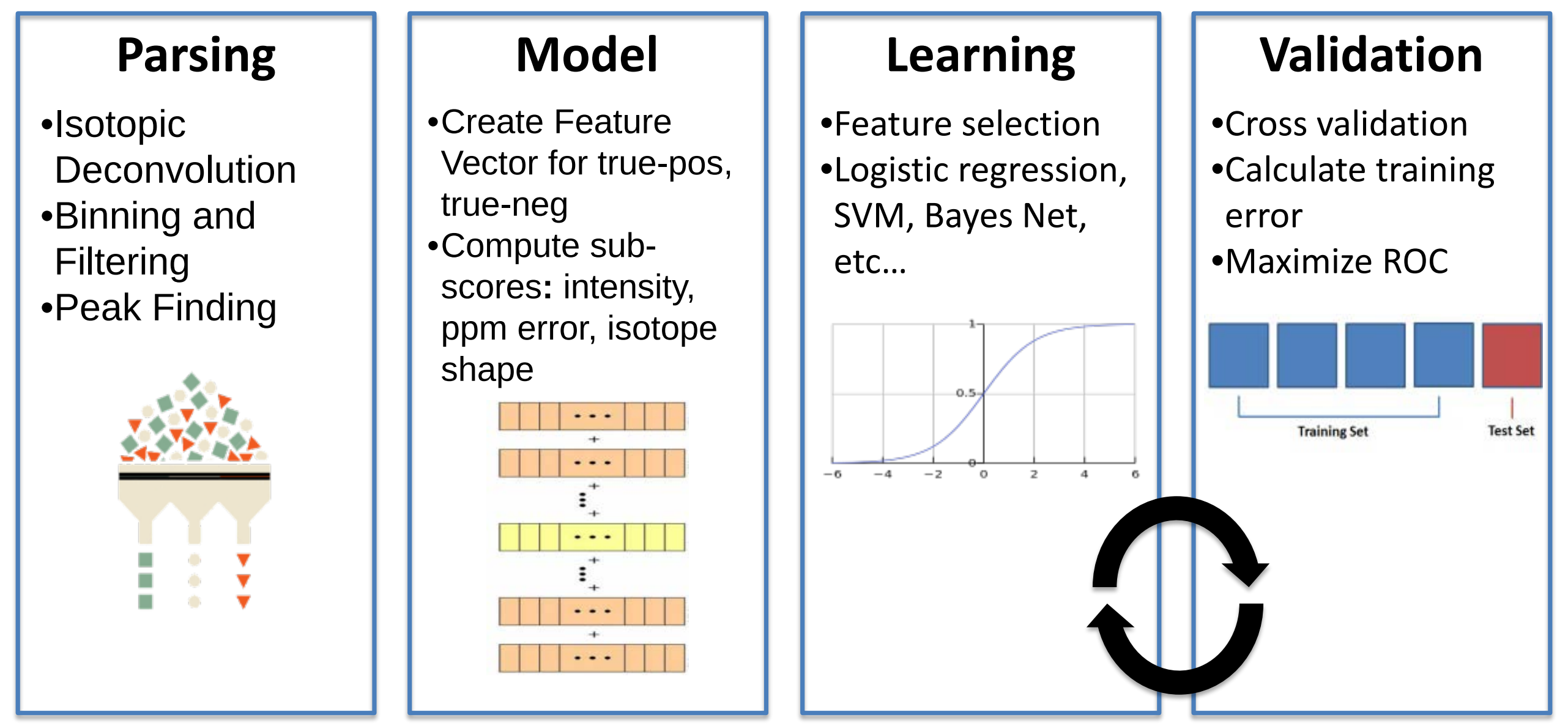
Introduction

- The development of new MS/MS technologies is driven by the necessity to achieve more complete and confident peptide/protein characterization in proteomics experiments.[1]
- Mechanisms of fragmentation and the resulting product ion types vary greatly between MS/MS methods.
- Additionally, experimental conditions, such as the presence TMT/iTRAQ, can significantly change MS/MS fragmentation propensities.
- Development of MS/MS scoring models has typically been considered part of the software development process and is rarely a user customizable component of peptide identification tools.[2]
- Scoring models are usually hard-coded, hindering adaptation and optimization for new MS/MS methods.
- FLIP's modular software architecture enables rapid learning and validation of scoring models for MS/MS data.
- FLIP is trained for new fragment types (UVPD) on both top down and bottom up data.

Methods

Software development

- FLIP is composed of four independent modules written in C#: parsing, modeling, learning, and cross validation (Figure 1). Each module can be replaced without recompilation of the entire software package.
- Required input: raw MS/MS data, true-positive and negative PSMs in community standard formats.
- A classifier is used to weight fragment ion features, such as mass error, isotopic fit, and intensity, that best separate the true-positive and true-negative training data.
- By default, FLIP supports both logistic regression and support vector machine models through the Accord.NET machine learning framework.
- The trained model is written to a tab-separated file, which serves as input for MS/MS scoring.



Parsing
- Isotopic Deconvolution
- Binning and Filtering
- Peak Finding

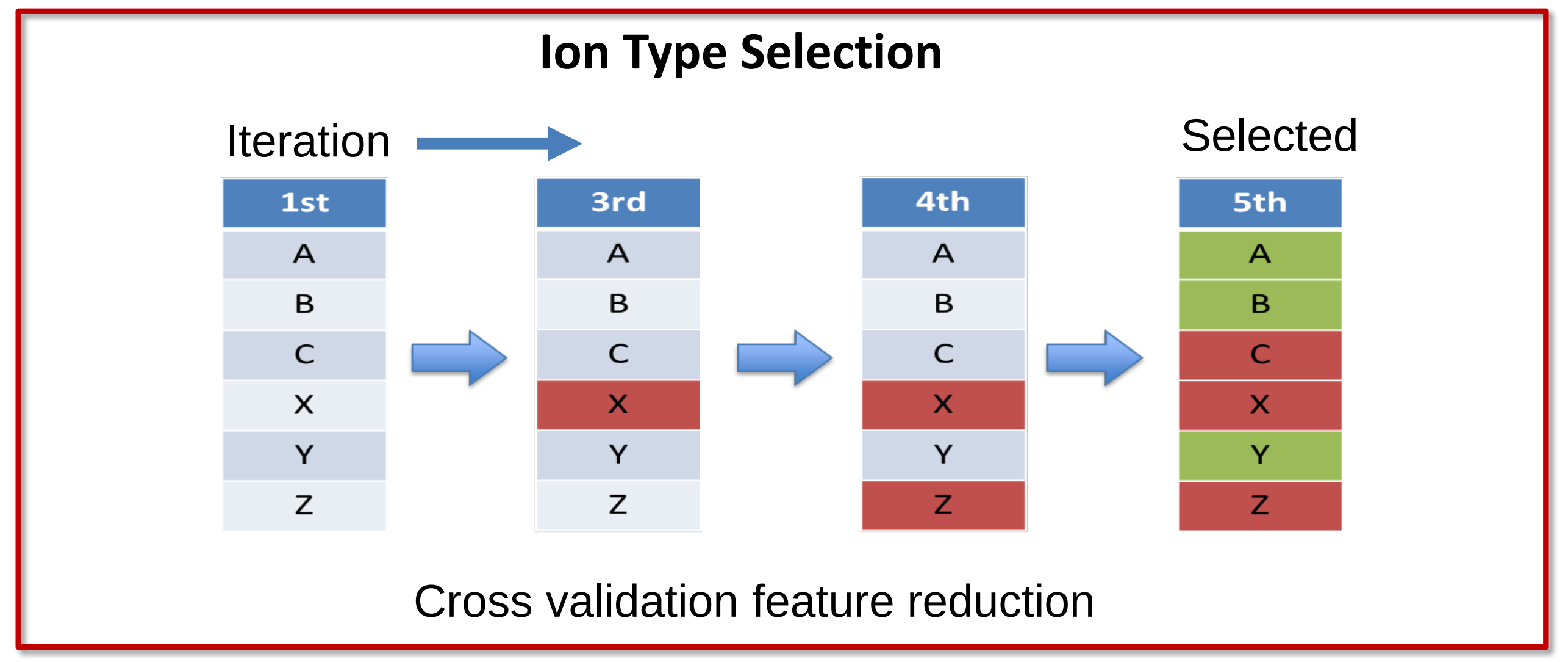
Model
- Create Feature Vector for true-pos, true-neg
- Compute sub-scores: intensity, ppm error, isotope shape

Learning
- Feature selection
- Logistic regression, SVM, Bayes Net, etc...

Validation
- Cross validation
- Calculate training error
- Maximize ROC

Figure 1. The FLIP workflow. Each box represents a set of algorithms that can be replaced via the Dependency injection pattern without recompilation of the FLIP codebase.

- To select ions for scoring, FLIP starts with a very large set of possible fragment ions and performs multiple rounds of 10-fold cross-validation. Each round reduces the number of product ions used for training.



Ion Type Selection

Cross validation feature reduction

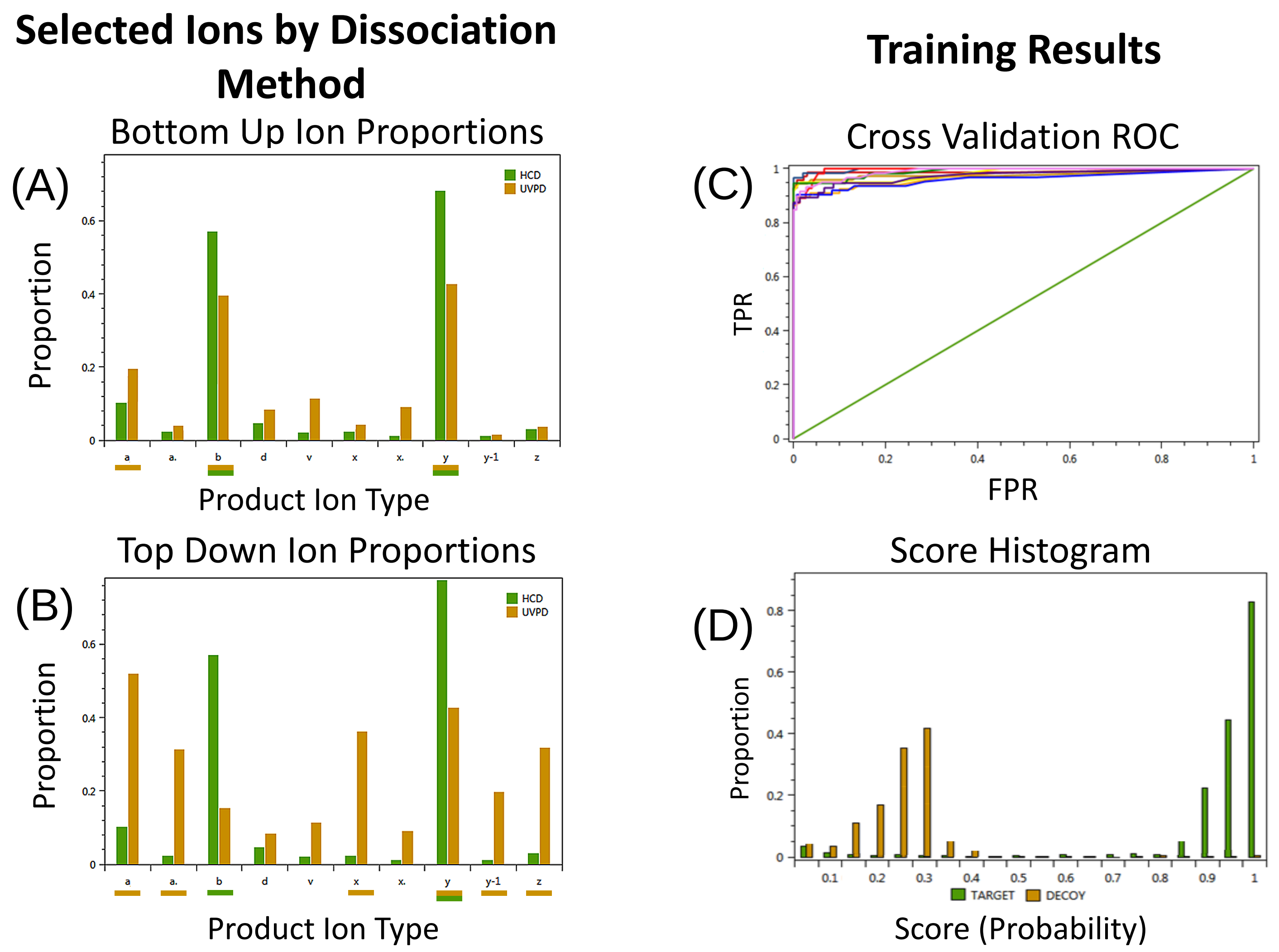
Figure 2. Cross validation feature reduction is used to remove the features for the lowest weighted ion each iteration of 10-fold cross validation. Area under the ROC curve is calculated each iteration to determine a stopping point for the product ion selection.

Mass spectrometry

- Experiments were performed using a Thermo Fisher Scientific Q Exactive HF modified to enable UVPD in a similar fashion previously published.[3] 193 nm photons were generated by a Coherent Excistar XS 500Hz excimer laser.
- Peptide and protein reversed phase separations were performed using 70 cm C18 and 50 cm C2 columns, respectively, with a Waters nanoACQUITY UPLC system at 300 nL/min.

Results

- Our goal was to create a flexible framework capable of adapting to many types of proteomics data. Here we use FLIP to train scoring models for top-down and bottom-up proteomics experiments utilizing UVPD and HCD.



Selected Ions by Dissociation Method

Training Results

(A) Bottom Up Ion Proportions

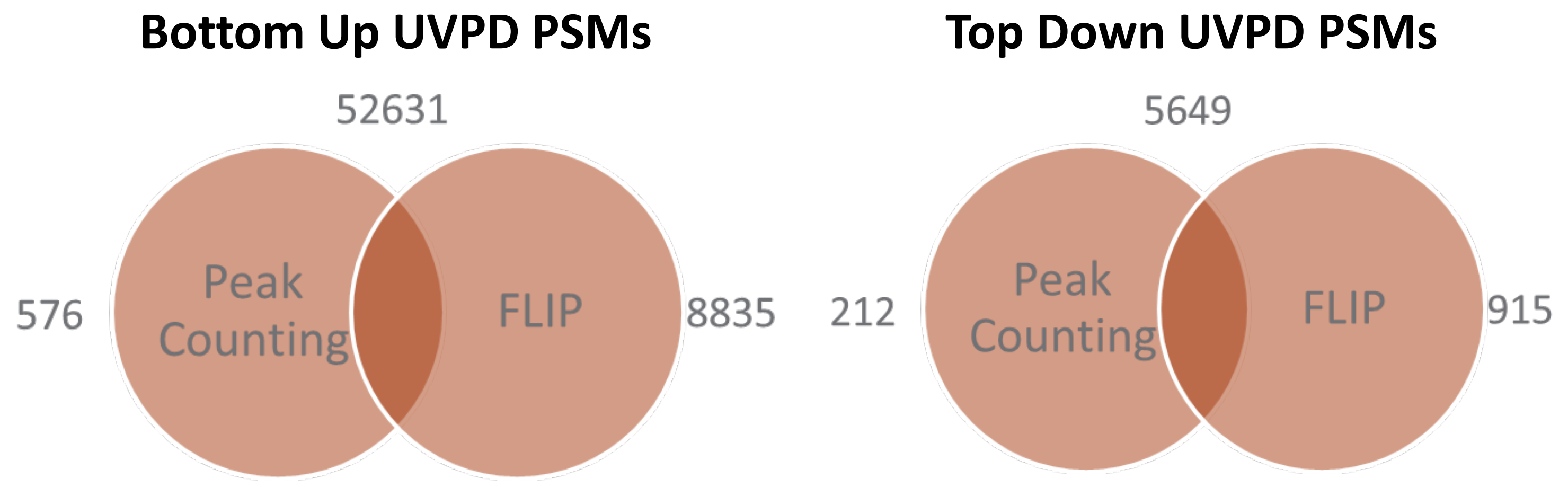
(B) Top Down Ion Proportions

(C) Cross Validation ROC

(D) Score Histogram

Figure 3. Proportion of each ion type found by FLIP for bottom up (A) and top down (B) HCD and UVPD experiments. Lines under the ion names indicate which ions were selected during training for each dissociation method. ROC curve for the final round of 10-fold cross validation (C) and score histogram calculated by FLIP using a logistic regression scoring model (D).

- We created bottom up MS/MS spectra for 6 bacterial organisms, which resulted in over 100,000 unique peptides, and top down MS/MS spectra for 3 bacterial organisms resulting in 6400 proteins. The scoring models were evaluated with Hela peptides with MSGF+, and a Fibrobacter succinogenes sample using MSPathFinder.



Bottom Up UVPD PSMs

Top Down UVPD PSMs

Figure 4. Number of PSMs at <1% false discovery rate found with our database search tool MSPathFinder. MSPathFinder was run with a trained FLIP scoring model and a scoring model that counted the number of A, B, C, X, Y, Z ions.

- We observed that FLIP was able to define effective models with significant differences in the numbers and types of fragment ions found for each of these data types, increasing the number of confident identifications (<1% false discovery rate) by 15% for bottom up and 12% for top down, when compared to a peak counting scoring model.

Conclusions

- FLIP is a universal tool for creating scoring models for many types of MS/MS spectra.
- FLIP allows developers to quickly adapt their informatics tools to data with new experimental conditions and fragmentation properties.
- FLIP does not require the user to manually determine fragment ions for training.
- This tool is available as part of the Informed Proteomics software package on Github at http://github.com/PNNL-Comp-Mass-Spec/Informed-Proteomics

Acknowledgments

This research is part of the "High Resolution and Mass Accuracy Capability" development project at EMSL, a Department of Energy (DOE) Office of Science User Facility sponsored by the Office of Biological and Environmental Research and located at the Pacific Northwest National Laboratory (PNNL), a multiprogram national laboratory operated by Battelle for the DOE under contract DE-AC05-76RL01830. This project was also supported by the Environmental Research (OBER) Pan-omics program at PNNL.

References:

1) Shaw, J. B. et. al. *J. Am. Chem. Soc.* **2013**, *135*, 12646–12651.
2) Kim, S. et. al. *Mol. Cell. Proteomics* **2010,** 9, 2840-2852
3) Fort, K. L. et. al. *Anal. Chem.* **2016**, 88, 2303–2310.

CONTACT: **Christopher Wilkins**
Biological Sciences Division
Pacific Northwest National Laboratory
E-mail: christopher.wilkins@pnnl.gov

www.pnnl.gov