Protein Annotator: Freeware for the Generation of UniProt Formatted XML Files Containing Custom User-Specified Features

NORTHWESTERN UNIVERSITY

Joseph B. Greer, Ryan T. Fellers, Richard D. LeDuc, Bryan P. Early, Alexandra J. van Nispen, Paul M. Thomas, Neil L. Kelleher

Northwestern University, Evanston, IL

Background

- ✓ The National Resource for Translational and Developmental Proteomics (NRTDP) has developed a beta-version tool that allows users to create custom UniProt formatted XML files
- ✓ It allows users to add unannotated point and range features
- ✓ Exports UniProt formatted XML files that ProSightPC can use to create databases

Introduction

Protein Annotator is a free application allowing researchers to create UniProt formatted files with specific, custom features. When searching top-down data against shotgun-annotated databases, protein entries often do not contain the specific features such as custom PTMs, ADCs, glycans, disulfide bonds, or endogenous cleavages that researchers are interested in. This significantly limits their ability to search for novel proteoforms. Additionally, searching for a specific proteoform is difficult when an entry is heavily modified. Both situations require a candidate database created from custom entries. Unfortunately, manually creating UniProt formatted files containing custom entries is tedious and prone to syntax errors. Protein Annotator allows researchers to create custom, valid UniProt files – making it easier to perform targeted proteoform searches.

Application Workflow

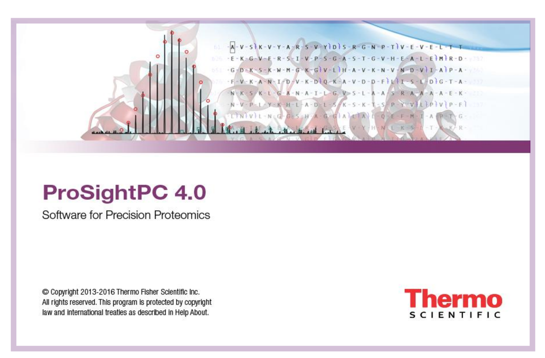
XML file

Input

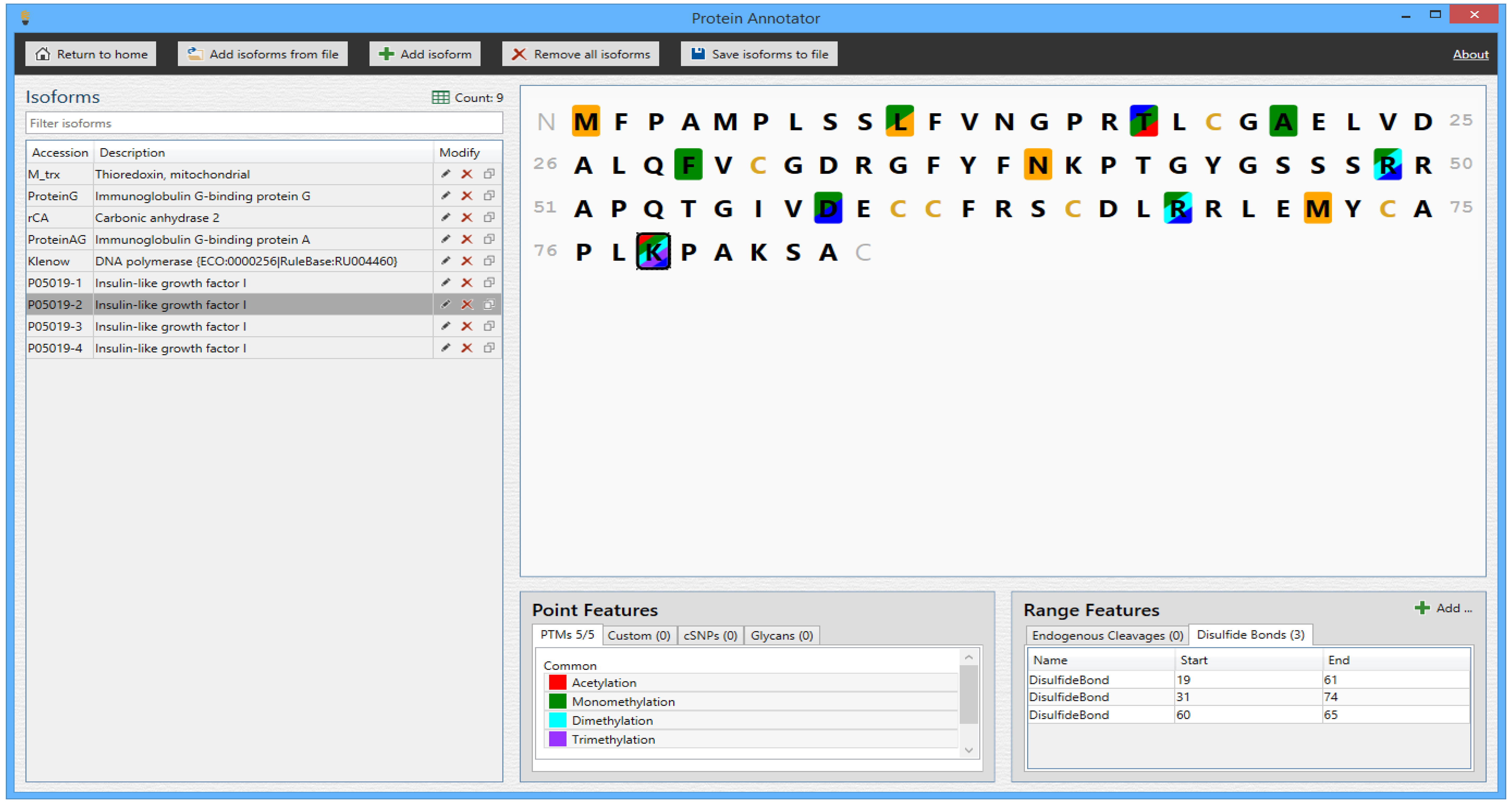
- UniProt formatted FASTA, text or XML file

Output

- UniProt formatted XML containing user defined custom features that can be used as input in tools such as ProSightPC 4.0



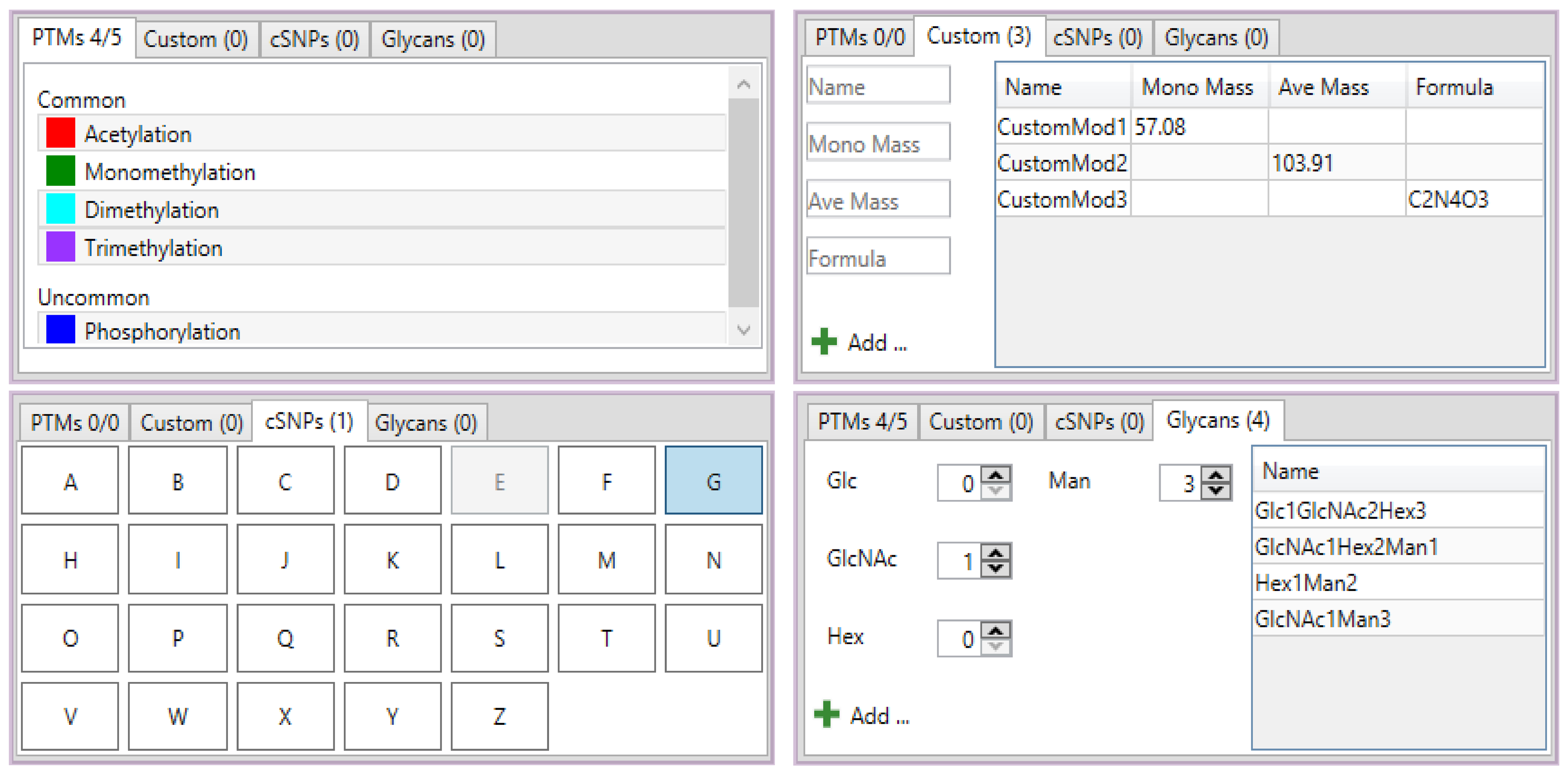
User Interface



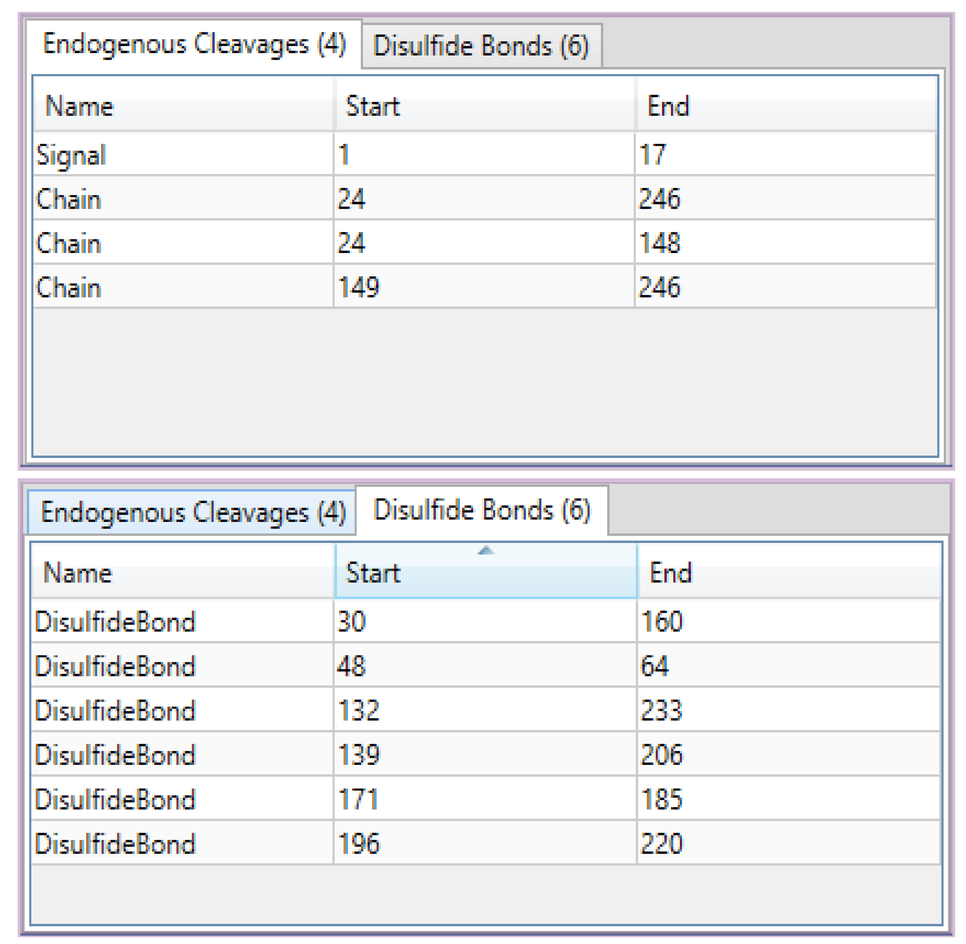
The user interface consists of four sections: the isoform table, the fragment map, the point feature panel, and the range feature panel. In this example, the input file contained entries associated with Thermo Fisher Scientific's Pierce Intact Protein Standard Mix.

http://proteinannotator.northwestern.edu

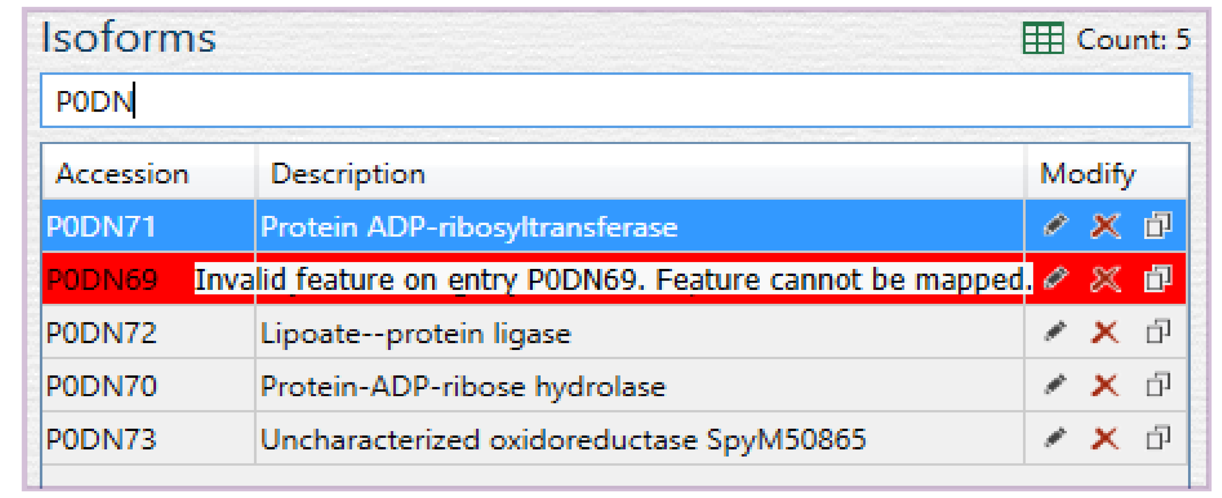
Point Features



Range Features



Isoform Table



Isoforms can be filtered and error messages are displayed.

Important Concepts

.FASTA — **FASTA file**
Text based file format. Application requires that FASTA files follow the UniProt format - entries contain a description line and the canonical protein sequence

.txt — **Text file**
Text based file format. Application requires that text files follow the UniProt format - entries contain the canonical protein sequence and annotated modifications

.XML — **XML file**
Extensible Markup Language based file format. Application requires that XML files follow the UniProt format - entries contain the canonical sequence and annotated modifications

Point Feature — **Point Feature**
Modification that only applies to a single amino acid such as PTMs and cSNPs

PTM — **Post-Translational Modification**
Processing event resulting from the addition or subtraction of a modifying group to one amino acid

cSNP — **Coding Single Nucleotide Polymorphism**
Single nucleotide variation at a specific position that changes the amino acid sequence

Range Feature — **Range Feature**
Modification that applies to a range of amino acids

Endogenous Cleavage — **Endogenous Cleavage**
Post translational cleavage event resulting in a new proteoform

Pierce Standard — **Pierce Intact Protein Standard Mix™**
Thermo Fisher Scientific developed mixture of six highly pure recombinant proteins used as an intact and top-down standard

Future Work

- Continue testing and work towards a stable public release
- Improve range feature visualization
- Expand point feature options

Acknowledgements

The authors would like to thank the members of the Top-Down Proteomics Development Team at Northwestern University. This research was supported by the National Institute of General Medical Sciences, and the National Institutes of Health under grant P41 GM108569. **Conflict Statement**: Some of the authors are involved with software commercialization.