

The UniDec Processing Pipeline for Rapid Analysis of Biotherapeutic Mass Spectrometry Data

Wilson Phung,¹ Corey E. Bakalarski,¹ Trent B. Hinkle,¹ Wendy Sandoval,¹ Michael T. Marty^{2,*}

¹Microchemistry, Proteomics, and Lipidomics Department, Genentech, Inc., South San Francisco, CA, 94080, USA

²Department of Chemistry and Biochemistry and the Bio5 Institute, University of Arizona, Tucson, AZ, 85721, USA

ABSTRACT: Recent advances in native mass spectrometry (MS) and denatured intact protein MS have made these techniques essential for biotherapeutic characterization. As MS analysis has increased in throughput and scale, new data analysis workflows are needed to provide rapid quantitation from large datasets. Here, we describe the UniDec Processing Pipeline (UPP) for the analysis of batched biotherapeutic intact MS data. UPP is built into the UniDec software package, which provides fast processing, deconvolution, and peak detection. The user and programming interfaces for UPP read a spreadsheet that contains the data file names, deconvolution parameters, and quantitation settings. After iterating through the spreadsheet and analyzing each file, it returns a spreadsheet of results and HTML reports. We demonstrate the use of UPP to measure correct pairing percentage on a set of bispecific antibody data and to measure drug-to-antibody ratios from antibody-drug conjugates. Moreover, because the software is free and open-source, users can easily build on this platform to create customized workflows and calculations. Thus, UPP provides a flexible workflow that can be deployed in diverse settings and for a wide range of biotherapeutic applications.

INTRODUCTION

Native and intact protein mass spectrometry (MS) have become indispensable tools for analysis of therapeutic antibodies and other therapeutic modalities.¹⁻³ By measuring the masses of intact antibodies, MS quickly reveals the distribution of proteoforms and can detect changes to the expected species. Importantly, as therapeutic modalities have become more complex, the mass distributions also confirm correct assembly of the products. For example, native/intact MS is useful for elucidating the correct pairing of bispecific antibodies and correct assembly of more complex structures.⁴⁻⁷ Over the past decade, the demand for bispecific antibody analysis has increased to thousands of samples per year. New modalities and antibody formats are often developed, leading to a variety of projects for intact mass analysis, which may include light chain ratio optimization, antibody pairing combinations, CDR swapping panels, and purification strategies. In developing and optimizing bispecific antibody pairing strategies, minimizing undesired species, including mis-paired antibodies and homodimers, is crucial in any bispecific therapeutic platform.

Native/intact MS is also useful for measuring the drug-to-antibody ratio for antibody-drug conjugates (ADCs) and in characterizing other covalently modified antibodies.⁸⁻¹² In designing ADCs, determining the drug payload is essential in improving potency as well as enhancing the functionality of the ADC as a whole.¹³ Thus, the ability to characterize bispecific antibodies and ADCs in a high-throughput manner is highly beneficial in evaluating these therapeutic strategies.

Unlike denatured intact protein MS, which has conventionally used a range of online injection strategies to enable

higher throughput analysis, native MS initially relied on manual injection with single-use borosilicate needles for each sample. However, recent work has advanced automated online injections and sample preparation.¹ For example, using online buffer exchange¹⁴ or online size-exclusion chromatography¹⁵ enables automated native MS analysis with very little user intervention at a rate of minutes per sample. Native MS is thus catching up with denatured intact protein MS in data acquisition throughput. At the same time, faster high-throughput methods are also being developed for denatured intact protein analysis, with rates as fast as a sample per second.^{16,17}

Together, these advances in data collection throughput have driven a need for higher throughput data analysis methods. In addition to computational time to process the data, manual work collecting metadata and organizing files can greatly increase the total time needed for analysis. Reporting/export options with limited customization or functionality further exacerbate the data turnaround time. As one reviewer noted, data analysis is currently the critical bottleneck for LC-MS workflows in biotherapeutic analysis.

A range of open source and commercial packages are available, which have been reviewed previously.¹⁸ Among the open-source options, UniDec has become widely used in academic and industrial settings due to its speed and flexibility.¹⁹ Prior publications have developed scoring methods,²⁰ algorithm improvements,²¹ and new modules to help support high-throughput data analysis with collections of related data.²² However, there was previously not a simple workflow for analysis of a large number of independent samples in the peer-reviewed literature, especially applied to biotherapeutic settings.

Here, we describe the UniDec Processing Pipeline (UPP), a new module in the UniDec software package designed to streamline analysis and reporting of large, independent data sets. We discuss the key components of UPP and demonstrate its use for rapid analysis of a dataset of bispecific antibody pairing and calculating drug-to-antibody ratios (DARs). We also discuss additional applications that could be built on this flexible open-source platform.

CODE AND SOFTWARE AVAILABILITY

UPP is part of the UniDec software package, which is distributed free and open source on GitHub: <https://github.com/michaelmarty/UniDec>. It has a modified BSD 3-clause license that permits unlimited use (including for commercial purposes and with modifications), unlimited numbers of downloads and installations, and very permissive redistribution, including allowing commercial redistribution with proper attribution (detailed in the license). Thus, UPP is readily customizable and can be deployed in a wide range of settings.

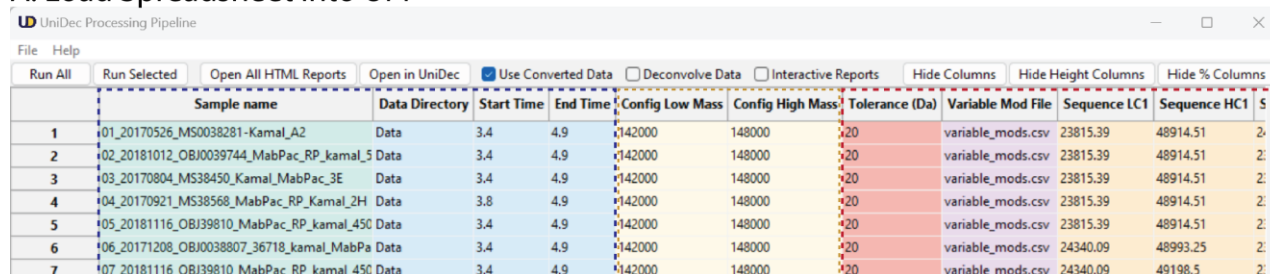
A compiled, stand-alone Windows graphical user interface (GUI) can be downloaded from GitHub: <https://github.com/michaelmarty/UniDec/releases>. Support for Mac and Linux operating systems is available through Python distribution described below. UPP can be run through the GUI (Figure 1) by selecting UPP from the main Launcher. Additional documentation and a wiki page

with video tutorials can also be found on the GitHub page (<https://github.com/michaelmarty/UniDec/wiki>).

UniDec is written primarily in Python with the core UniDec algorithm in C. All changes to implement UPP were added to the Python code and relied on the existing UniDec application programming interface (API). In addition to GitHub, UniDec is also available from the Python Packaging Index (PyPI, <https://pypi.org>) and can be installed with “pip install unidec”. After installing UniDec, the main GUI can be launched with the command: “python -m unidec.Launcher”. With Python, the UniDec GUI can be run on Linux and Mac computers. However, it can also be run through the command line and scripted. Binaries of the C code are provided for Linux and Mac, but users may need to run the compiling scripts on their own machine.

Finally, to facilitate use in high-throughput settings, a UniDec Docker image has been built. Freely available for download and deployment from DockerHub (<https://hub.docker.com/r/michaelmarty/unidec>), this image allows for instant access to UPP analysis using Docker or Singularity on any system, from personal laptops to high-performance computing clusters and cloud providers such as Amazon Web Services. Between the GUI for desktop use, the PyPI distribution for easy Python scripting and direct integration into custom data processing pipelines, and the container for large scale deployment, UPP is

A. Load Spreadsheet into UPP



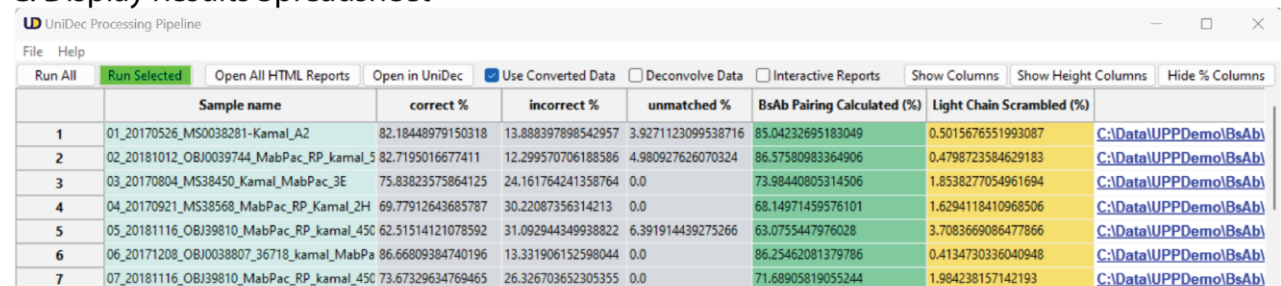
	Sample name	Data Directory	Start Time	End Time	Config Low Mass	Config High Mass	Tolerance (Da)	Variable Mod File	Sequence LC1	Sequence HC1	S
1	01_20170526_MS0038281-Kamal_A2	Data	3.4	4.9	142000	148000	20	variable_mods.csv	23815.39	48914.51	2
2	02_20181012_OBJ0039744_MabPac_RP_kamal_5	Data	3.4	4.9	142000	148000	20	variable_mods.csv	23815.39	48914.51	2
3	03_20170804_MS38450_Kamal_MabPac_3E	Data	3.4	4.9	142000	148000	20	variable_mods.csv	23815.39	48914.51	2
4	04_20170921_MS38568_MabPac_RP_Kamal_2H	Data	3.8	4.9	142000	148000	20	variable_mods.csv	23815.39	48914.51	2
5	05_20181116_OBJ39810_MabPac_RP_kamal_450	Data	3.4	4.9	142000	148000	20	variable_mods.csv	23815.39	48914.51	2
6	06_20171208_OBJ0038807_36718_kamal_MabPa	Data	3.4	4.9	142000	148000	20	variable_mods.csv	24340.09	48993.25	2
7	07_20181116_OBJ39810_MabPac_RP_kamal_450	Data	3.4	4.9	142000	148000	20	variable_mods.csv	24340.09	49198.5	2

File Information Deconvolution Parameters Match Settings

B. Batch Process



C. Display Results Spreadsheet



	Sample name	correct %	incorrect %	unmatched %	BsAb Pairing Calculated (%)	Light Chain Scrambled (%)	HTML Report
1	01_20170526_MS0038281-Kamal_A2	82.18448979150318	13.888397898542957	3.9271123099538716	85.04232695183049	0.5015676551993087	C:\Data\UPPDemo\BsAb\
2	02_20181012_OBJ0039744_MabPac_RP_kamal_5	82.7195016677411	12.299570706188586	4.980927626070324	86.57580983364906	0.4798723584629183	C:\Data\UPPDemo\BsAb\
3	03_20170804_MS38450_Kamal_MabPac_3E	75.83823575864125	24.161764241358764	0.0	73.98440805314506	1.8538277054961694	C:\Data\UPPDemo\BsAb\
4	04_20170921_MS38568_MabPac_RP_Kamal_2H	69.77912643685787	30.22087356314213	0.0	68.14971459576101	1.6294118410968506	C:\Data\UPPDemo\BsAb\
5	05_20181116_OBJ39810_MabPac_RP_kamal_450	62.51514121078592	31.092944349938822	6.391914439275266	63.0755447976028	3.7083669086477866	C:\Data\UPPDemo\BsAb\
6	06_20171208_OBJ0038807_36718_kamal_MabPa	86.66809384740196	13.331906152598044	0.0	86.25462081379786	0.4134730336040948	C:\Data\UPPDemo\BsAb\
7	07_20181116_OBJ39810_MabPac_RP_kamal_450	73.67329634769465	26.326703652305355	0.0	71.68905819055244	1.984238157142193	C:\Data\UPPDemo\BsAb\

Relative Intensities BsAb Pairing HTML Report

Figure 1: Overview of UPP showing the selected parts of the spreadsheet GUI and key steps of the workflow, including (A) loading the data into the GUI, (B) batch processing through the key steps, and (C) displaying the results with reports.

UniDec Report

File Name: 03_20170804_MS38450_Kamal_MabPac_3E.txt

Directory: C:\Data\UPPDemo\BsAb\Data

Symbol	Mass	Centroid	Height	Match	Matcherror	Label	FWHM	LowVal	FWHM	HighVal	FWHM	Error	Mean
0	○	145450.0	145456.284437	11.976420	145442.71	7.29	LC2 Mispair (Incorrect)	65.0	145425.0	145490.0	10.333417		
1	▽	145630.0	145636.850686	100.000000	145626.82	3.18	BsAb (Correct)	55.0	145610.0	145665.0	3.494541		
2	△	145825.0	145826.896428	21.421528	145810.93	14.07	LC1 Mispair (Incorrect)	65.0	145795.0	145860.0	4.392800		

*Click on a column header to sort.

The BsAb Pairing Calculated is: 72.98840766018084

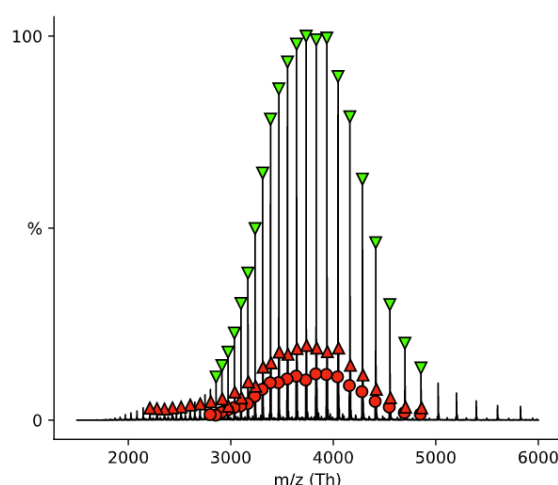
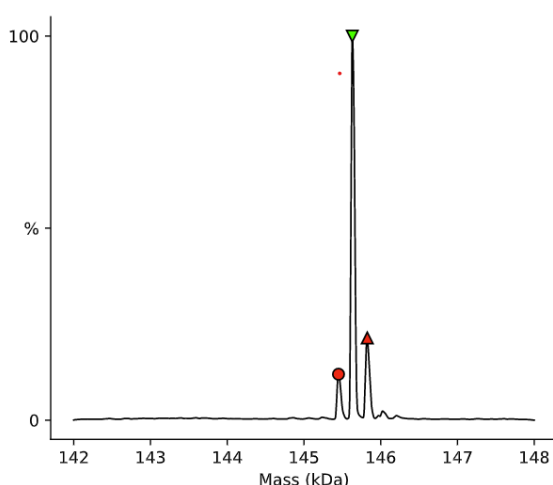


Figure 2: Example HTML report for a bispecific antibody data file. The report includes the table of peaks (*top*), deconvolved mass spectrum (*left*), and annotated *m/z* spectrum (*right*).

easily accessible and interoperable with a variety of systems.

COMPUTATIONAL DESIGN AND METHODS

Overall Design and GUI

The primary data structure in UPP is a simple data table, imported from either an Excel spreadsheet or CSV file into a Pandas DataFrame²³ (Figure 1A). UPP operates by iterating over each row of the table (Figure 1B), applying different options specified in the input to the deconvolution and analysis, and writing outputs from the analysis into a results spreadsheet, which can be displayed in the GUI (Figure 1C). As described below, these outputs can include peak intensities for specific mass combinations or DAR calculations, depending on the input values in the spreadsheet. The default for the software is to use peaks heights, but peak areas can also be returned by setting a “Quant Mode” column in the spreadsheet. Users can deconvolve and analyze a set of data automatically without viewing any spectra. However, HTML reports are generated for each file and linked in the results spreadsheet (Figure 2). Each of these individual reports is also concatenated into a larger combined HTML report viewable in any web browser. Examples of input

spreadsheets, HTML reports, and results spreadsheets are provided in the supporting information.

UPP consists of three main Python modules. First, the UPP.py file provides the GUI. The UPP GUI is a simple spreadsheet interface that allows spreadsheet files to be opened, saved, and manipulated (Figure 1). Users can select specific rows to run through UniDec or run the entire spreadsheet. A limited set of run options are present, but specific deconvolution settings are entered into the spreadsheet, not in the GUI, to enable automated analysis at scale.

Batch Processing Engine

Second, the batch.py file provides the core engine of the UPP workflow. The UniDecBatchProcessor object can either read a spreadsheet file or a Pandas DataFrame object. The batch processor is called by the GUI, but it can also be run through scripting or command line inputs (for example with a command “python -m unidec.batch file.xlsx”).

The engine iterates over each row in the spectrum and reads the values specified by different column keywords present in the spreadsheet (Figure 1B). The complete list of recognized column keywords is detailed in the help menu, and a copy of the help page is provided in the supporting information. The only required keyword is “Sample name”.

The “Sample name” column provides the location of the file to deconvolve. Note, the capitalization of the keywords should match exactly. An additional optional keyword of “Data Directory” can be provided to specify the data location if a full path is not provided in the “Sample name”. A wide range of file types are currently supported by UniDec, including text, csv, mzML (using `pymzml`²⁴), mzXML (using `pyteomics`^{25, 26}), and raw data formats from Agilent (using `multiplierz`²⁷), Waters, and Thermo. Note, the vendor raw data formats use libraries that are only available on Windows. We welcome support from other file types from anyone willing to contribute Python libraries for conversion.

For each row iteration, UniDec will open the specified file, process the data, run the deconvolution, perform peak picking, and generate an HTML report. The UniDec processing, deconvolution, and peak picking have been described previously,^{19, 28} and more information can be found in the online wiki and tutorial videos linked above. Briefly, UniDec uses a Bayesian deconvolution approach that combines smoothing of charge and/or mass distributions with a Richardson-Lucy deconvolution of peak shapes. Iterating between these two processes, it assigns the charge state distribution for each m/z data point. This matrix of m/z vs z is then transformed into a matrix of mass vs z , which is summed across the charge dimension to yield the zero-charge mass distribution. Peaks are then selected from this mass distribution based on a user-defined relative intensity threshold and a user-defined local mass window, where peaks are local maxima within the window that exceed the intensity threshold.

Each report has a sortable list of peaks, plots of both the deconvolved mass and raw data, and a list of parameters used. An example is provided in the supporting information and shown in Figure 2. The HTML format makes reports easily shareable, and they can be opened directly in a browser by clicking on the GUI. The individual report locations are added to the output results spreadsheet, which is saved at the end of the run. For simple deconvolution, the only required field is the file location, and the only output will be the “Reports” column. A global HTML report is also saved alongside the results spreadsheet. This global report concatenates the individual HTML reports into a combined document for easy browsing, and an example is provided in the supporting information for the DAR dataset.

Settings for the deconvolution can be adjusted by adding additional columns to the spreadsheet. For example, including “Start Time” and “End Time” as keywords will select specific time ranges from the data to analyze (assuming retention time is present in the original data format). All scans between the start time and end time will be summed together into a single spectrum. To deconvolve distinct time regions in a single data file, multiple rows can be added with different retention time settings in each row.

Various deconvolution settings can also be adjusted. For example, adding columns like “Config Low Mass” and “Config High Mass” can be used to set the minimum and maximum masses for the deconvolution. An external “Config File” location can also be added as a column to override the default parameters with a new config parameter set. In contrast with MetaUniDec²² and other UniDec batch processing features,¹⁹ UPP enables each row of the spreadsheet

to have different, customizable deconvolution parameters if needed. An example input spreadsheet is provided in the supporting information.

In addition to modifying config parameters in the spreadsheet, users can open the main UniDec GUI on any individual file to manually fine tune settings (either directly from the UPP GUI or by opening the file separately with the main UniDec GUI). During data conversion, a fresh config file is created for each file. However, if the “Use Converted Data” option is selected, the existing config file (with any manual changes) will be used. In all cases, the config file will still be overwritten by settings in the spreadsheet, so the spreadsheet must be updated to reflect the manual adjustments.

Matching Workflow

The third primary Python module used in UPP is `matchtools.py`, which provides an extensible library of modules that analyze the peaks that are detected in the deconvolution step. These libraries have been designed to provide a framework for custom peak analysis, and users are welcome to design their own recipes for analyzing the peaks and reporting the results back to the output spreadsheet. To demonstrate the potential for these recipes, we designed two analysis workflows. The first checks for correct combinations of masses from a list of provided masses and/or sequences. The second calculates DARs for antibody-drug conjugates (ADCs). Each recipe can be loaded into the system at runtime and can be activated by required keywords in the column names of the input data table.

Checking for Correct Pairing of Bispecific Antibodies

The goal of this recipe is to extract the peak intensities for predicted masses. Within this general framework, there are a number of ways to accomplish the overarching goal, depending on the column keyword and cell values provided. In the most basic case, users can provide the masses directly in cells and include either “Correct”, “Incorrect”, or “Ignore” in the column labels. Only the correct column is required. There can be multiple columns of each type, as long as they include the “Correct”, “Incorrect”, and “Ignore” keywords somewhere in the column header. For our bispecific antibody example, we set “LC1 Mispair (Incorrect)” and “LC2 Mispair (Incorrect)” as two possible incorrect species.

Beyond the basic case of directly providing masses, users can also match with combinations of masses/sequences. Here, columns are provided in the spreadsheet with the keyword “Sequence” plus some unique identifier. For example, we use “Sequence LC1” for the first light chain value. Currently, the values provided in each sequence cell can either be the mass of the species or the amino acid sequence of the protein, which UniDec will automatically use to calculate the mass. However, it would be possible in the future to convert SMILES, nucleic acid sequences, or other similar codes to mass if a suitable function can be provided in Python. Custom code could also be written to query a database based on identifiers in the cell and retrieve a mass value.

In our example of bispecific antibody analysis, we specify the predicted masses for “Sequence LC1”, “Sequence HC1”, “Sequence LC2”, and “Sequence HC2”. Additional columns can also be provided to apply fixed modifications and disulfide oxidation, which requires an amino acid sequence.

All of these adjust the masses that UniDec will generate to match with the detected peaks.

After defining sequences, users can then specify the sequence combinations under the correct, incorrect, or ignore columns. Here, the cell uses “Seq” with the unique identifier as a code to specify the “Sequence” species in a string with “+” separating the species. For example, the “LC1 Mispair (Incorrect)” column has a cell value of “SeqLC1+SeqHC1+SeqLC1+SeqHC2”. This combination tells the software to combine the masses of the columns with “Sequence LC1” + “Sequence HC1” + “Sequence LC1” + “Sequence HC2”. As a reminder, correct capitalization of the keywords is required. Also, it is possible to only have a single species (“SeqProtein” for example) in the cell. An example input file is provided in the supporting information.

For each correct, incorrect, or ignored column (whether defined directly or as sequence combinations), this recipe will calculate the potential mass for this species, apply any variable modifications, and generate a list of potential species. It will then match this list with peaks found in the data, subject to the defined tolerance. For each combination, it will return the peak intensities (as defined by the quant mode, both absolute and relative) to the results file. It will also sum all the correct, incorrect, and ignored peaks to generate the total peak intensities (both absolute and relative) of the correct, incorrect, and ignored species. Finally, it will calculate the percentage correct vs. incorrect after ignored species are removed and report which matches are found. In the HTML reports, peaks are colored based on their status of correct (green), incorrect (red), ignored (blue), or unknown (yellow), as shown in Figure 2 and in the example report in the supporting information. An example results spreadsheet is also provided in the supporting information. Overall, this workflow allows users to quickly extract the absolute and relative intensities of combinations of potential species.

Although we have illustrated this for bispecific antibodies, it would be straightforward to use this same workflow for measuring protein-ligand binding or covalent protein modifications. Here, users would specify “Sequence Protein” and “Sequence Ligand” with the necessary masses. Correct binding stoichiometries could be defined as “SeqProtein+SeqLigand”. Incorrect binding could be defined as

“SeqProtein”. UPP would then return the percentage of protein that is bound to the ligand versus unbound. Additional custom calculations or other binding stoichiometries could be added as needed.

Drug-to-Antibody Ratio Calculations

In addition to the correct pairing workflow described above, we also developed a matching workflow for calculating DARs. Here, the spreadsheet requires the “Protein Mass”, which can be either 1) the predicted mass, 2) the amino acid sequence of the protein, 3) or a “Seq” code word combination, using the same nomenclature described above. Fixed modifications can be applied in several ways, as described in the help documentation.

This workflow also requires the “Drug Mass” and “Max Drugs”, which specifies the maximum number of potential drug conjugations to consider. A “Min Drugs” column, specifying the minimum number of potential drug conjugations to consider, can also be supplied, but it will default to 0. UniDec then combines different numbers of the drug mass, ranging from the minimum to the maximum number of potential conjugations, with the total protein mass. These masses are matched with peaks from the spectrum to determine the peak intensities. The DAR is then calculated²⁹ from the peak intensities and added as a new column on the report. An example report is provided in the supporting information, and a screenshot of the outputs and select inputs is shown in Figure 3.

RESULTS AND DISCUSSION

Application to Bispecific Antibodies

To demonstrate the use of UPP, we first tested it against a dataset of bispecific antibodies containing 115 independent denatured LC/MS runs that had been previously published.⁴ Example data for this BsAb workflow and the DAR workflow described below are posted at MassIVE (MSV000092242, DOI: 10.25345/C52Z13069). In the spreadsheet (see example in the supporting information), we specified the file names and the data directory. Data was provided as Thermo Raw format and converted using the internal libraries in UniDec. The time range was specified to capture the antibody peak eluting from the column. Simple deconvolution settings were provided to limit the m/z and mass range and specify peak picking settings.

	Sample name	Protein Mass	Drug Mass	Max Drugs	Min Drugs	DAR	Reports
1	20221027_EMR_SEC_Native_OBJ42236_1_degly.raw	146050	471	15	0	0.0	C:\Data\UPPDemo\DR\20221027_EMR
2	20221027_EMR_SEC_Native_OBJ42236_2_biotin_A_degly.raw	146050	339	15	0	4.417904806936531	C:\Data\UPPDemo\DR\20221027_EMR
3	20221027_EMR_SEC_Native_OBJ42236_3_biotin_B_degly.raw	146050	339	15	0	4.131094423510705	C:\Data\UPPDemo\DR\20221027_EMR
4	20221027_EMR_SEC_Native_OBJ42236_4_DIG_C_degly.raw	146050	471	15	0	4.334316306977574	C:\Data\UPPDemo\DR\20221027_EMR
5	20221027_EMR_SEC_Native_OBJ42236_5_DIG_D_degly.raw	146050	471	15	0	4.140988051846604	C:\Data\UPPDemo\DR\20221027_EMR
6	20221027_EMR_SEC_Native_OBJ42236_6_degly.raw	144360	471	15	0	0.0	C:\Data\UPPDemo\DR\20221027_EMR
7	20221027_EMR_SEC_Native_OBJ42236_7_biotin_E_degly.raw	144360	339	15	0	3.9620650834689424	C:\Data\UPPDemo\DR\20221027_EMR
8	20221104_EMR_SEC_Native_OBJ42236_8_biotin_F_degly.raw	144360	339	15	0	3.6914268426750807	C:\Data\UPPDemo\DR\20221104_EMR
9	20221027_EMR_SEC_Native_OBJ42236_10_DIG_G_degly.raw	144360	471	15	0	3.9874336375126322	C:\Data\UPPDemo\DR\20221027_EMR
10	20221027_EMR_SEC_Native_OBJ42236_9_DIG_H_degly_2.raw	144360	471	15	0	4.057583950490218	C:\Data\UPPDemo\DR\20221027_EMR

Figure 3: Screenshot of the output and select inputs from the DAR calculation mode.

For the match settings, a match tolerance of 20 Da was chosen. Two files did not match within this tolerance and were expanded to 50 Da. A global fixed modification of -32 Da was applied to account for disulfides. Predicted masses were supplied for each of the four “Sequences”: LC1, LC2, HC1, and HC2. The “BsAb (Correct)” column was specified as SeqLC1+SeqHC1+SeqLC2+SeqHC2. The “LC1 Mispair (Incorrect)” column was SeqLC1+SeqHC1+SeqLC1+SeqHC2, and the “LC2 Mispair (Incorrect)” column was SeqLC2+SeqHC1+SeqLC2+SeqHC2. Species are annotated in Figure 4A and 4B.

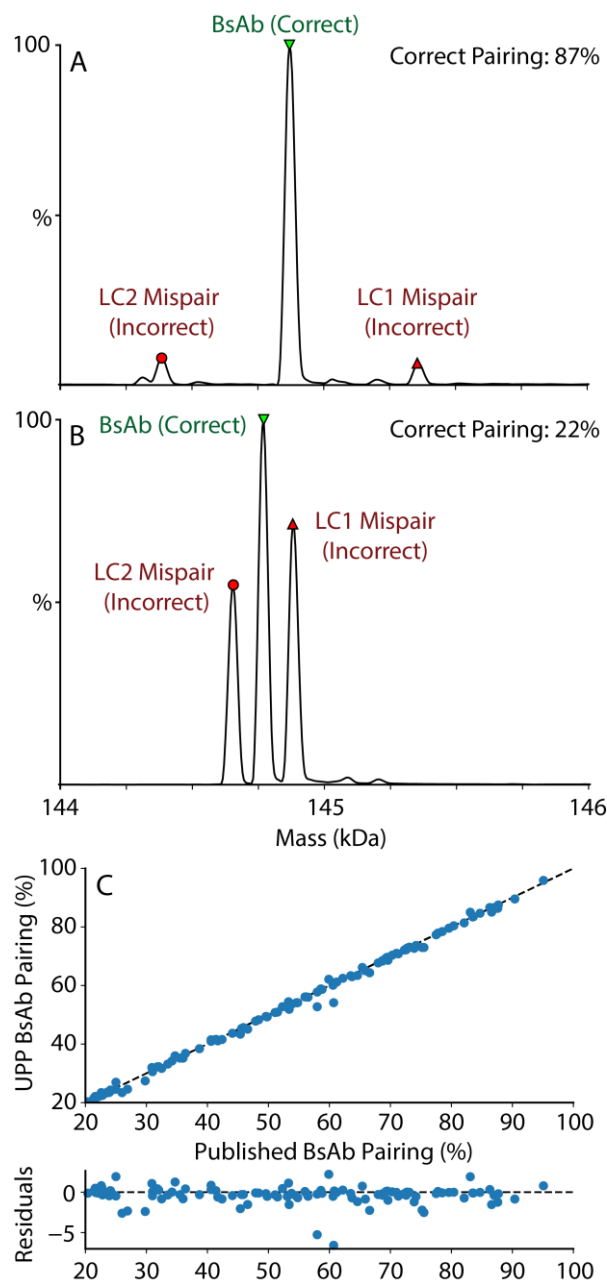


Figure 4: Deconvolved data for bispecific antibody (BsAb) with high (A) and low (B) correct pairing percentages. (C) Comparison of published BsAb pairing percent (ref. 4) versus the UPP results. The dashed line shows perfect agreement. Residuals shown below indicate the difference between published and UPP.

After loading this spreadsheet file into UPP, the full “Run All” process took 99 seconds to convert/average the data, deconvolve the results, and assign the peaks for all 115 files. Thus, a standard laptop was able to process the data set with less than a second per file. After data has been converted and averaged from raw file into a text file, data conversion can be removed for subsequent reanalysis, which shortens the deconvolution and analysis steps to 90 seconds. Removing the deconvolution process shortened the time needed for peak picking and data analysis to 60 seconds, but shortcuts in the code could shorten that further by removing optional file imports. Importantly, these results demonstrate that UPP can process data faster than it can be collected, even with the highest throughput systems.¹⁷

As part of the workflow, UPP calculated the percentage of correctly paired bispecific antibody from the relative amounts of BsAb, LC1 Mispair, and LC2 Mispair.⁴ Example data in Figure 4A illustrates a relatively high correct pairing, with low amounts of incorrectly paired byproducts. Figure 4B illustrates an example with relatively low correct pairing with higher amounts of incorrectly paired species. The results excellently matched prior analysis,⁴ with a root mean squared deviation of 1.1% (Figure 4C). The maximum absolute difference was 6.7%, and only two files had absolute differences greater than 3%. Together, these data demonstrate that UPP can rapidly and accurately process native MS and intact protein ESI data from large screening studies and provide valuable quantitative outputs.

Application to DAR Calculations

To test the DAR calculation workflow, we applied UPP to a set of 10 data files collected on a Thermo Scientific Exactive EMR with online SEC with native MS, using a previously described LC/MS method.⁷ This data set contained two antibodies with duplicates of either biotin or drug conjugation. An unmodified control was included for each antibody. The mass of each antibody was supplied along with the mass of the conjugate. The minimum number of conjugates was set to 0 and the maximum was set to 15.

Analysis of these 10 files took around 6 seconds, less than 1 second per file. After the data had been converted and deconvolved, reanalysis took only 4 seconds for the data set. A screenshot of the output is shown in Figure 3, and example deconvolutions are shown in Figure 5. An example results file and an example report are provided in the supporting information. All conjugates had DAR values around 4 that matched manual calculations. Unmodified controls both had DAR values of 0, as expected.

Interestingly, the deglycosylation was partially incomplete (Figure 5A), which led to a series of unmatched peaks (shown in yellow in Figures 5A, C, and E). The DAR workflow does not currently support variable modifications in the same way as the BsAb workflow, so to correct for incomplete deglycosylation, we used the DoubleDec feature in UniDec. DoubleDec loads a template mass distribution that is used to deconvolve the output of the primary UniDec deconvolution.³⁰ Essentially, it specifies a complex peak shape pattern (Figure 5A) and then collapses that fixed pattern into a single peak in a second round of deconvolution (Figure 5B). Importantly, it assumes that the pattern of post-translational modification is constant for all drug conju-

gated states. DoubleDec has previously been used to measure zinc and lipid binding to rhodopsin, which has a complex set of post-translational modifications (PTMs),³⁰ and to measure tryptophan binding to TRAP, also combining a set of PTMs into a single peak.³¹

To use DoubleDec in UPP, we first manually deconvolved the unmodified antibodies to obtain a kernel file (Figure 5A). The deconvolved mass distributions from each antibody were saved separately, and the paths to those files were included in the spreadsheet as the “DoubleDec Kernel File”. After deconvolving with these kernel files in the automated UPP deconvolution, the second series of peaks was largely removed (Figure 5B, D, and E).

DoubleDec systematically lowered the calculated DAR values, as seen in Figure 5. All conjugates had lower DARs with DoubleDec. The DARs for the 4 biotin conjugates decreased by an average of 3.5%, and DARs for the 4 drug conjugates decreased by an average of 2%. The biotin conjugate was more affected because it has a smaller mass difference (339 Da) than the drug conjugate (471 Da). Thus, the drug conjugate has more space between the peaks to accommodate the incompletely deglycosylated peaks. In contrast, the second incompletely deglycosylated peak (+331 Da) overlapped with the biotin conjugation (+339 Da), and this overlap caused slightly higher signal for larger conjugates and thus a systematically high DAR. DoubleDec corrects this subtle error and enables accurate DAR calculation.

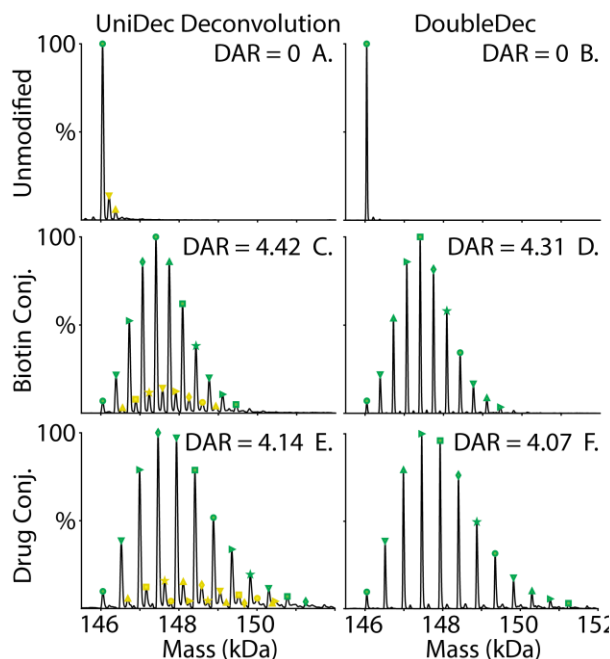


Figure 5: UPP results for DAR calculations of an antibody with UniDec (A, C, E) and DoubleDec (B, D, E) with unmodified forms (A, B), biotin conjugates (C, D), and drug conjugates (E, F).

Overall, these results demonstrate the power of UPP for quickly calculating DAR values for a set of data. Although we generated the DoubleDec kernel files manually, it would be possible to semi-automate this process by having a separate spreadsheet of kernel files that are deconvolved first and

validated. Files from this first spreadsheet could then be entered as kernel files in subsequent spreadsheets. Reviewing the HTML reports can help to alert the user when DoubleDec is needed, and a cutoff for the percentage of unknown peaks could be set up with custom code to automatically trigger DoubleDec. Moreover, DAR values could be calculated for multiple species within the same spectrum, such as dissociated antibody chains or different proteoforms, by using additional rows for each file with different protein masses. With flexible scripting and spreadsheet frameworks, teams can customize their workflow and automate these complex analyses.

CONCLUSIONS

Here, we described a new module to the UniDec software package, the UniDec Processing Pipeline. UPP offers several advantages for high-throughput data processing. Because it is open source, labs and companies can develop custom workflows. The example workflows shown here demonstrate its potential for biopharmaceutical applications, but the same framework could be readily applied to drug discovery¹⁵ or protein design³² by simply adjusting the spreadsheet columns. For example, UPP could be applied to high-throughput analysis of non-covalent MHC complexes to screen for neoantigen candidates by quantifying the amount of successful peptide exchange.³³ Accessible inputs and outputs make the software easy to interface with other tools. Finally, because it is free, cross platform, and containerized, it can be run in individual workstations, local servers, or cloud providers without licensing restrictions or requirements to transfer data offsite or to a 3rd party ecosystem. It can be run in either GUI or command line modes, and the results can be viewed with standard desktop tools: a web browser and a spreadsheet application.

Alongside these advantages, several limitations remain. First, only a subset of deconvolution settings can be controlled from the spreadsheet currently. However, because each parameter takes only a few extra lines of code, we will add additional settings as needed and requested. If desired, users can also control all deconvolution settings by specifying an external config file in each line of the table.

Second, as discussed above, UPP is fast but not perfectly efficient. Much of the computational time is spent reading and writing from the hard drive, which could be streamlined with future code developments to pass data in the memory between the Python scripts and the core UniDec binaries, ideally by developing a shared library and Python wrapper.

Finally, for simple systems and abundant species, deconvolution with standard parameters is very robust. However, for complex data or low abundance species, automated processing with default parameters may not be reliable. If deconvolution settings need to be adjusted for each file, UniDec can be opened for manual deconvolution on each, but that defeats the purpose of batch processing. In any case, we recommend that users carefully validate the tool and regularly check the reports to ensure that the deconvolution results are correct.

Overall, UPP provides a flexible template to build complex workflows on, presenting a streamlined interface to batch process, deconvolve, and analyze data. We welcome

users to build custom in-house pipelines, which they can either keep private or contribute back to the free and open-source code base. In future iterations, it would also be possible to link other UniDec engines for CD-MS analysis³⁴ and more sophisticated LC/MS analysis with chromatographic peak picking. Pairing a flexible spreadsheet input with these deconvolution engines will significantly advance high throughput biotherapeutic analysis by mass spectrometry.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website. Supporting files are provided for: the help file HTML; an individual example BsAb HTML report; combined example DAR HTML reports with and with DoubleDec; an example input spreadsheet for BsAb; an example results spreadsheet for BsAb; and an example results spreadsheet for DAR.

AUTHOR INFORMATION

Corresponding Author

* Michael T. Marty

Department of Chemistry and Biochemistry, Bio5 Institute, University of Arizona Tucson, AZ 85721, United States email: mtmarty@arizona.edu

ACKNOWLEDGMENT

This work was funded by the National Science Foundation (CHE-1845230 to M.T.M.). The authors thank the community of UniDec users for valuable suggestions and motivating the work. We also thank the Antibody Engineering and BioAnalytical Sciences departments at Genentech for the bispecific antibody and antibody-drug conjugate materials, respectively.

REFERENCES

- Karch, K. R.; Snyder, D. T.; Harvey, S. R.; Wysocki, V. H., Native Mass Spectrometry: Recent Progress and Remaining Challenges. *Annu. Rev. Biophys.* **2022**, *51* (1), 157-179.
- Campuzano, I. D. G.; Sandoval, W., Denaturing and Native Mass Spectrometric Analytics for Biotherapeutic Drug Discovery Research: Historical, Current, and Future Personal Perspectives. *J Am Soc Mass Spectr* **2021**, *32* (8), 1861-1885.
- Kellie, J. F.; Tran, J. C.; Jian, W.; Jones, B.; Mehl, J. T.; Ge, Y.; Henion, J.; Bateman, K. P., Intact Protein Mass Spectrometry for Therapeutic Protein Quantitation, Pharmacokinetics, and Biotransformation in Preclinical and Clinical Studies: An Industry Perspective. *J Am Soc Mass Spectr* **2021**, *32* (8), 1886-1900.
- Yin, Y.; Han, G.; Zhou, J.; Dillon, M.; McCarty, L.; Gavino, L.; Ellerman, D.; Spiess, C.; Sandoval, W.; Carter, P. J., Precise quantification of mixtures of bispecific IgG produced in single host cells by liquid chromatography-Orbitrap high-resolution mass spectrometry. *MAbs* **2016**, *8* (8), 1467-1476.
- Yan, Y.; Xing, T.; Wang, S.; Daly, T. J.; Li, N., Coupling Mixed-Mode Size Exclusion Chromatography with Native Mass Spectrometry for Sensitive Detection and Quantitation of Homodimer Impurities in Bispecific IgG. *Anal. Chem.* **2019**, *91* (17), 11417-11424.
- Grunert, I.; Heinrich, K.; Hingar, M.; Ernst, J.; Winter, M.; Bomans, K.; Wagner, K.; Fevre, A.; Reusch, D.; Wuhler, M.; Bulau, P., Comprehensive Multidimensional Liquid Chromatography-Mass Spectrometry for the Characterization of Charge Variants of a Bispecific Antibody. *J Am Soc Mass Spectr* **2022**, *33* (12), 2319-2327.
- Joshi, K. K.; Phung, W.; Han, G.; Yin, Y.; Kim, I.; Sandoval, W.; Carter, P. J., Elucidating heavy/light chain pairing preferences to facilitate the assembly of bispecific IgG in single cells. *mAbs* **2019**, *11* (7), 1254-1265.
- Campuzano, I. D. G.; Nshanian, M.; Spahr, C.; Lantz, C.; Netirojjanakul, C.; Li, H.; Wongkongkathap, P.; Wolff, J. J.; Loo, J. A., High Mass Analysis with a Fourier Transform Ion Cyclotron Resonance Mass Spectrometer: From Inorganic Salt Clusters to Antibody Conjugates and Beyond. *J. Am. Soc. Mass Spectrom.* **2020**, *31* (5), 1155-1162.
- Dyachenko, A.; Wang, G.; Belov, M.; Makarov, A.; de Jong, R. N.; van den Bremer, E. T. J.; Parren, P. W. H. I.; Heck, A. J. R., Tandem Native Mass-Spectrometry on Antibody-Drug Conjugates and Submillion Da Antibody-Antigen Protein Assemblies on an Orbitrap EMR Equipped with a High-Mass Quadrupole Mass Selector. *Anal. Chem.* **2015**, *87* (12), 6095-6102.
- Ehkirch, A.; D'Atri, V.; Rouviere, F.; Hernandez-Alba, O.; Goyon, A.; Colas, O.; Sarrut, M.; Beck, A.; Guilleme, D.; Heinisch, S.; Cianferani, S., An Online Four-Dimensional HICxSEC-IMxMS Methodology for Proof-of-Concept Characterization of Antibody Drug Conjugates. *Anal. Chem.* **2018**, *90* (3), 1578-1586.
- Nagornov, K. O.; Gasilova, N.; Kozhinov, A. N.; Virta, P.; Holm, P.; Menin, L.; Nesatyy, V. J.; Tsybin, Y. O., Drug-to-Antibody Ratio Estimation via Proteoform Peak Integration in the Analysis of Antibody-Oligonucleotide Conjugates with Orbitrap Fourier Transform Mass Spectrometry. *Anal. Chem.* **2021**, *93* (38), 12930-12937.
- Pacholarz, K. J.; Barran, P. E., Use of a charge reducing agent to enable intact mass analysis of cysteine-linked antibody-drug-conjugates by native mass spectrometry. *EuPA Open Proteomics* **2016**, *11*, 23-27.
- Baah, S.; Laws, M.; Rahman, K. M., Antibody-Drug Conjugates-A Tutorial Review. *Molecules* **2021**, *26* (10).
- VanAernum, Z. L.; Busch, F.; Jones, B. J.; Jia, M.; Chen, Z.; Boyken, S. E.; Sahasrabudde, A.; Baker, D.; Wysocki, V. H., Rapid online buffer exchange for screening of proteins, protein complexes and cell lysates by native mass spectrometry. *Nat Protoc* **2020**, *15* (3), 1132-1157.
- Ren, C.; Bailey, A. O.; VanderPorten, E.; Oh, A.; Phung, W.; Mulvihill, M. M.; Harris, S. F.; Liu, Y.; Han, G.; Sandoval, W., Quantitative Determination of Protein-Ligand Affinity by Size Exclusion Chromatography Directly Coupled to High-Resolution Native Mass Spectrometry. *Anal. Chem.* **2019**, *91* (1), 903-911.
- Campuzano, I. D. G.; Pelegri-O'Day, E. M.; Srinivasan, N.; Lippens, J. L.; Egea, P.; Umeda, A.; Aral, J.; Zhang, T.; Laganowsky, A.; Netirojjanakul, C., High-Throughput Mass Spectrometry for Biopharma: A Universal Modality and Target Independent Analytical Method for Accurate Biomolecule Characterization. *J. Am. Soc. Mass Spectrom.* **2022**, *33* (11), 2191-2198.
- Zacharias, A. O.; Liu, C.; VanAernum, Z. L.; Covey, T. R.; Bateman, K. P.; Wen, X.; McLaren, D. G., Ultrahigh-Throughput Intact Protein Analysis with Acoustic Ejection Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **2023**, *34* (1), 4-9.
- Rolland, A. D.; Prell, J. S., Approaches to Heterogeneity in Native Mass Spectrometry. *Chem. Rev.* **2022**, *122* (8), 7909-7951.
- Marty, M. T.; Baldwin, A. J.; Marklund, E. G.; Hochberg, G. K.; Benesch, J. L.; Robinson, C. V., Bayesian deconvolution of mass and ion mobility spectra: from binary interactions to polydisperse ensembles. *Anal. Chem.* **2015**, *87* (8), 4370-6.
- Marty, M. T., A Universal Score for Deconvolution of Intact Protein and Native Electrospray Mass Spectra. *Anal. Chem.* **2020**, *92* (6), 4395-4401.
- Marty, M. T., Eliminating Artifacts in Electrospray Deconvolution with a SoftMax Function. *J. Am. Soc. Mass Spectrom.* **2019**, *30* (10), 2174-2177.
- Reid, D. J.; Diesing, J. M.; Miller, M. A.; Perry, S. M.; Wales, J. A.; Montfort, W. R.; Marty, M. T., MetaUniDec: High-Throughput Deconvolution of Native Mass Spectra. *J. Am. Soc. Mass Spectrom.* **2019**, *30* (1), 118-127.

23. McKinney, W. In *Data structures for statistical computing in python*, 2010; Austin, TX: 2010; pp 51-56.
24. Bald, T.; Barth, J.; Niehues, A.; Specht, M.; Hippler, M.; Fufezan, C., pymzML—Python module for high-throughput bioinformatics on mass spectrometry data. *Bioinformatics* **2012**, *28* (7), 1052-1053.
25. Goloborodko, A. A.; Levitsky, L. I.; Ivanov, M. V.; Gorshkov, M. V., Pyteomics—a Python Framework for Exploratory Data Analysis and Rapid Software Prototyping in Proteomics. *J Am Soc Mass Spectr* **2013**, *24* (2), 301-304.
26. Levitsky, L. I.; Klein, J. A.; Ivanov, M. V.; Gorshkov, M. V., Pyteomics 4.0: Five Years of Development of a Python Proteomics Framework. *J Proteome Res* **2019**, *18* (2), 709-714.
27. Alexander, W. M.; Ficarro, S. B.; Adelmant, G.; Marto, J. A., multiplierz v2.0: A Python-based ecosystem for shared access and analysis of native mass spectrometry data. *Proteomics* **2017**, *17* (15-16).
28. Kostelic, M. M.; Marty, M. T., Deconvolving Native and Intact Protein Mass Spectra with UniDec. In *Proteoform Identification: Methods and Protocols*, Sun, L.; Liu, X., Eds. Springer US: New York, NY, 2022; pp 159-180.
29. Jones, J.; Pack, L.; Hunter, J. H.; Valliere-Douglass, J. F., Native size-exclusion chromatography-mass spectrometry: suitability for antibody-drug conjugate drug-to-antibody ratio quantitation across a range of chemotypes and drug-loading levels. *MAbs* **2020**, *12* (1), 1682895.
30. Norris, C. E.; Keener, J. E.; Perera, S.; Weerasinghe, N.; Fried, S. D. E.; Resager, W. C.; Rohrbough, J. G.; Brown, M. F.; Marty, M. T., Native Mass Spectrometry Reveals the Simultaneous Binding of Lipids and Zinc to Rhodopsin. *Int. J. Mass Spectrom.* **2021**, *460*, 116477.
31. Li, W.; Norris, A. S.; Lichtenthal, K.; Kelly, S.; Ihms, E. C.; Gollnick, P.; Wysocki, V. H.; Foster, M. P., Thermodynamic coupling between neighboring binding sites in homo-oligomeric ligand sensing proteins from mass resolved ligand-dependent population distributions. *Protein Sci.* **2022**, *31* (10), e4424.
32. Vorobieva, A. A.; White, P.; Liang, B.; Horne, J. E.; Bera, A. K.; Chow, C. M.; Gerben, S.; Marx, S.; Kang, A.; Stiving, A. Q.; Harvey, S. R.; Marx, D. C.; Khan, G. N.; Fleming, K. G.; Wysocki, V. H.; Brockwell, D. J.; Tamm, L. K.; Radford, S. E.; Baker, D., De novo design of transmembrane β barrels. *Science* **2021**, *371* (6531), eabc8182.
33. Schachner, L. F.; Phung, W.; Han, G.; Darwish, M.; Bell, A.; Mellors, J. S.; Szrentic, K.; Huguet, R.; Blanchette, C.; Sandoval, W., High-Throughput, Quantitative Analysis of Peptide-Exchanged MHCI Complexes by Native Mass Spectrometry. *Anal. Chem.* **2022**, *94* (42), 14593-14602.
34. Kostelic, M. M.; Zak, C. K.; Liu, Y.; Chen, V. S.; Wu, Z.; Sivinski, J.; Chapman, E.; Marty, M. T., UniDecCD: Deconvolution of Charge Detection-Mass Spectrometry Data. *Anal. Chem.* **2021**, *93* (44), 14722-14729.

For Table of Contents Only:

