# INTERPRETING COMPLEX MASS SPECTRA THROUGH DOUBLE DECONVOLUTION ANALYSIS

By

JACK ZENGWEI LIU

_____

A Thesis Submitted to The Honors College

In Partial Fulfillment of the Bachelors degree
With Honors in

Chemistry

THE UNIVERSITY OF ARIZONA

M A Y 2 0 2 1

Approved by:

_____

Dr. Michael T. Marty
Department of Chemistry and Biochemistry

**Abstract**

UniDec provides a rapid and reliable approach to deconvolving native mass spectra into mass and charge components. However, the complexity of interactions in large biomolecules may still render the data impossible to interpret, especially when several variables contribute to the mass. Here, we present a second deconvolution step after conversion to mass space. Using a control spectrum as the point spread function, this mathematically removes a chosen mass-determining factor, enabling the analysis of previously unassignable peaks. We demonstrate the use of this double deconvolution in several applications, including variable metal binding to multiple metalloprotein proteoforms and the incorporation of peptides into lipid bilayer mimetic discs.

## Introduction

Proteins and other large biomolecules play essential roles in cellular processes, often forming dynamic complexes to facilitate their various functions. A greater understanding of these complexes is key for mapping novel interactions and designing related drugs, and consequently, there is a strong interest in techniques that can characterize biochemical structures and interactions. One such technique is mass spectrometry (MS), a widely used analytical tool that measures the mass-to-charge ratio ($m/z$) of ionized samples. Many MS ionization methods exist, but soft ionization techniques are of particular interest in biochemical analysis because they induce little to no fragmentation, thereby allowing one to observe intact macromolecules. One example is electrospray ionization (ESI), which deposits charges on solution-phase analytes as they desolvate.[1] However, the chemical conditions of these methods may not be representative of a protein's native environment, disrupting interactions or creating ones with no biological relevance.

Native MS is a promising technique that uses ESI under non-denaturing conditions, imitating the native biochemical environment through controlling parameters such as pH and ionic strength. These measures maintain the natural folded state of proteins in solution and keep non-covalent interactions intact in the gas phase, enabling MS analysis of structural details such as subunit stoichiometry.[2] However, the use of ESI leads to analytes with a variable number of charges.[1] In tandem with the large size of biological macromolecules and the dynamic non-covalent interactions measured through native MS, this greatly complicates analysis.

UniDec software provides a rapid and reliable way to separate ESI mass spectra into their corresponding mass and charge components. The algorithm approaches MS data analysis through a Bayesian deconvolution approach, framing the *m*/*z* spectrum as the convolution between a peak shape and a set of weighted delta functions describing the contributions of specific ions. UniDec then iteratively deconvolves the spectrum to separate the mass and charge dimensions.[3] While this is sufficient for many native MS experiments, analysis remains difficult for samples involving multiple mass-determining factors. Because each factor introduces a distribution of mass peaks that are in turn split into distributions based on other factors, the complexity of the mass spectrum scales multiplicatively, potentially rendering the data impossible to interpret even after conversion to mass space.

Here, we present a novel double deconvolution approach to native MS data analysis. We apply a second deconvolution step after conversion to mass space, framing the data as the convolution between two mass spectra—the underlying image and the point-spread function—describing the contributions of individual mass-determining factors. The algorithm is an application of the standard Richardson-Lucy deconvolution.[4,5] By selecting control mass spectra containing variation from only one factor as the point-spread function, we mathematically separate this factor from the data and retrieve the underlying image. We demonstrate the use of double deconvolution in the analysis of variable zinc binding to rhodopsin and the incorporation of LL-37 peptides into nanodiscs. Rhodopsin is a G-protein-coupled receptor that binds $Zn^{2+}$ ions, the loss of which is known to result in age-related vision loss,[6] while LL-37 is a human antimicrobial peptide whose functionality is based around lipid membrane interactions.[7] Nanodiscs are membrane

mimetics constructed from a lipid bilayer surrounded by membrane scaffold proteins (MSPs).[8] In these examples, we explore the use of double deconvolution in observing zinc binding across various rhodopsin proteoforms and enhancing the natural monodispersity of LL-37-bearing nanodiscs, respectively. Overall, we show the applicability of double deconvolution in resolving highly complex biochemical systems.

## Methods

Software Development and Algorithm

Our double deconvolution software (DoubleDec) is implemented as user options in UniDec. The data analysis is written in C, while the graphical user interface (GUI) is written in Python.

Given a mass spectrum containing multiple mass-determining factors, we define the problem as iteratively retrieving the underlying main signal, which contains mass variation due to only the desired factors, from its convolution with a point-spread function (Equation 1):

$$M_{i+1} = M_i \left( \left( \frac{R}{M_i * P} \right) * P^* \right) \qquad \text{(Equation 1)}$$

where $M_{i+1}$ is the current main signal, $M_i$ is the main signal from the previous iteration, $R$ is the recorded signal, $P$ is the point-spread function, and $P^*$ is the flipped point-spread function. We initially define the main signal as equal to the recorded signal. At each iteration, we calculate the convolution of the main signal $M_i$ with the point-spread function $P$, then use it as the divisor in a point-wise division with the recorded signal $R$. The quotient function is then convolved with $P^*$, and the result is point-wise multiplied with $M_i$ to yield $M_{i+1}$. After some number of iterations $n$, this process calculates the deconvolution of the recorded signal with the point-spread function in $M_n$, converging on the solution of maximum likelihood.

DoubleDec uses mass spectra as both the recorded signal and point-spread function. Users are asked to select the files corresponding to both spectra, referred to as the data file and kernel file, respectively, allowing the program to account for any sort of mass-determining factor as long as the user can create the appropriate control

samples for native MS. All parts of the data analysis are performed in linear time except for the two convolutions per iteration, which are done by applying the Fourier convolution theorem with fast Fourier transforms. Iterations are calculated until the sum-of-squares difference from the previous main signal is negligible or until the maximum number of iterations is reached. Because these fast Fourier transforms run in $O(n \log n)$ time, the asymptotic runtime of the overall algorithm is also $O(n \log n)$.

Specific Applications

We describe applications of DoubleDec to analyze data for rhodopsin-zinc binding and LL-37 incorporation into nanodiscs. Native mass spectra were collected for rhodopsin samples at varying zinc concentrations as previously described.[9] Mass spectra were also collected for LL-37 incorporated into 2-dipalmitoyl-*sn*-glycero-3-phospho-(1′-rac-glycerol) (DPPG) nanodiscs as previously described.[10]

**Results and Discussion**

To assess the correctness of the DoubleDec algorithm, we first describe its use on a set of empty 1-palmitoyl-2-oleoyl-*sn*-glycero-3-phosphocholine (POPC) nanodiscs, using the resulting native mass spectrum as both the data and kernel file (**Figure 1**). The kernel file contains the data for the point-spread function, and therefore represents the control sample containing the mass-determining factors to be mathematically removed. By selecting identical data for the recorded signal and point-spread function, DoubleDec is expected to remove all mass variation from the spectrum.

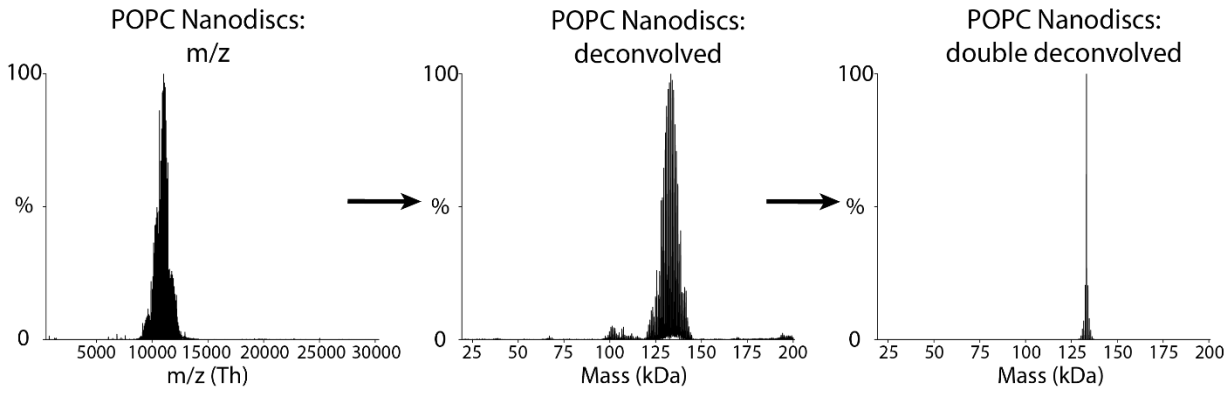


**Figure 1.** Demonstration of double deconvolution algorithm on a POPC nanodisc mass spectrum. The first deconvolution from *m*/*z* to mass is done with standard UniDec, and the second deconvolution is done using the resultant spectrum as a kernel on itself.

With a kernel equivalent to the data, a single distribution containing all the mass variation of the kernel is found: the entire spectrum itself. All mass variation from the kernel is then collapsed into one peak, resulting in the singular peak shown above (**Figure 1**). DoubleDec is therefore capable of removing mass-determining factors as expected.

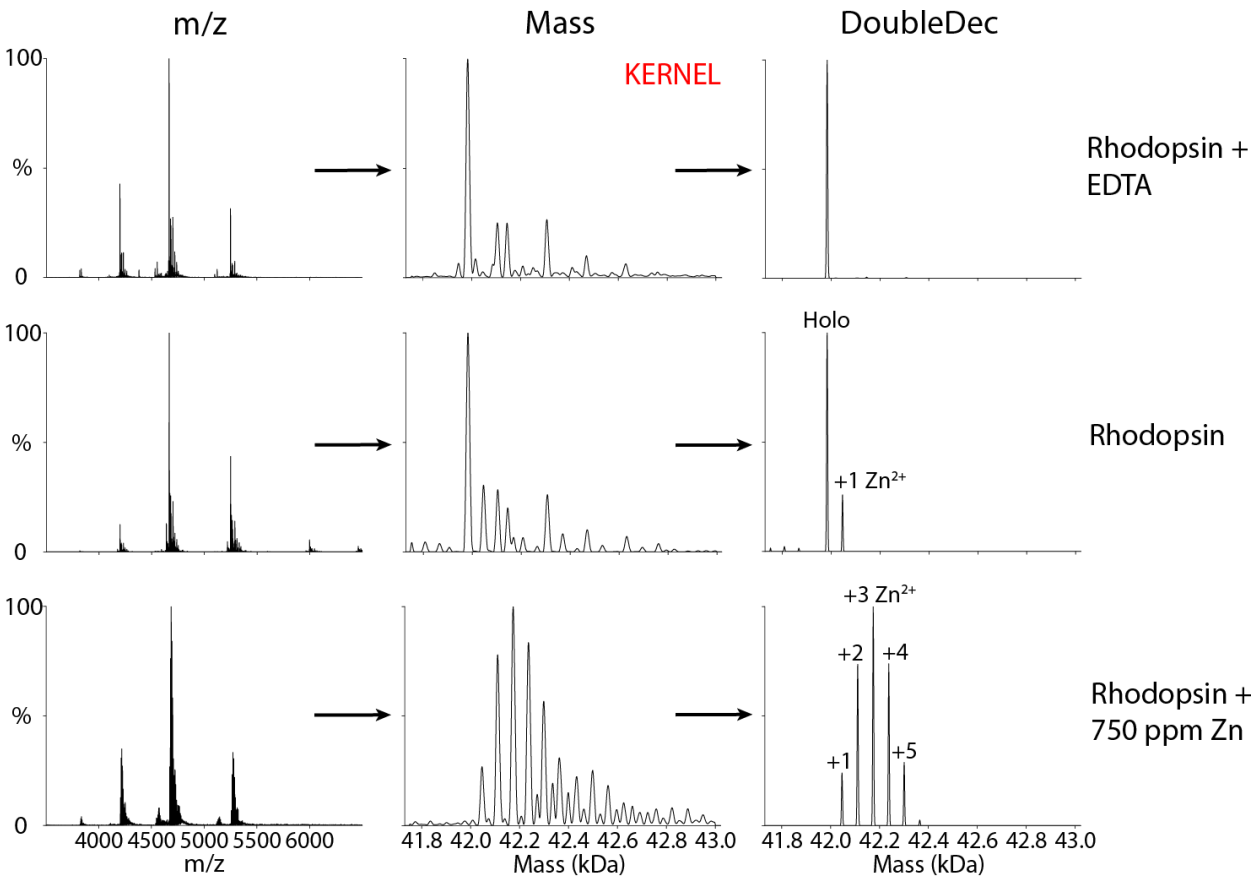


**Figure 2.** Mass spectra of rhodopsin samples with EDTA, no additives, and additional 750 ppm Zn. Each row shows the raw mass spectra to the left, the deconvolved mass spectra in the center, and double-deconvolved mass spectra to the right. The rows correspond to rhodopsin with EDTA at the top, the unaltered rhodopsin sample in the middle, and rhodopsin with additional 750 ppm Zn at the bottom.

Next, we show the use of DoubleDec in observing the variable binding of zinc to rhodopsin. Native MS was performed on rhodopsin samples containing ethylenediaminetetraacetic acid (EDTA), no additional additives, or additional zinc.[9] Because EDTA is a strong chelating agent known to remove bound metals, the

rhodopsin-EDTA spectrum is used as a control representing mass variation solely due to multiple rhodopsin proteoforms. The deconvolved rhodopsin-EDTA spectrum was used as the kernel for applying double deconvolution to all samples, thus combining the contributions from all proteoforms into one peak. This mathematically removes proteoform variation as a mass-determining factor, leaving only variation from bound zinc and significantly improving its quantification (**Figure 2**). These results demonstrate the viability of applying double deconvolution in any mass spectra complicated by multiple mass-determining factors, allowing one to consider factors in isolation and greatly simplifying analysis.

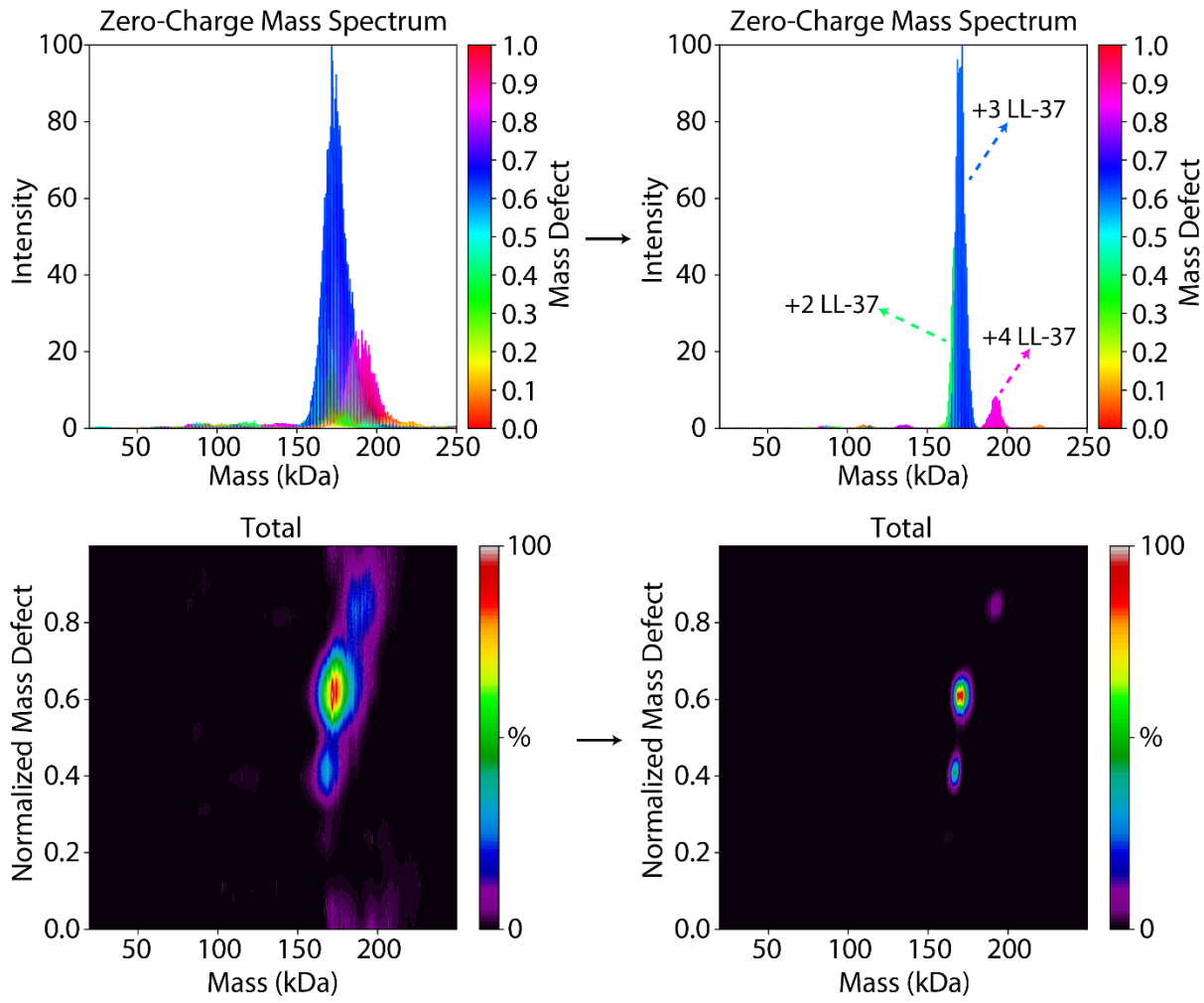


**Figure 3.** Mass spectra of LL-37 DPPG nanodiscs made at a 3:1 LL-37:nanodisc ratio. The top row shows the zero-charge mass spectra, while the bottom row shows the corresponding mass defect plots. Single-deconvolved data are on the left, while double-deconvolved data are on the right.

Lastly, we demonstrate the use of DoubleDec to enhance the resolution of mass peak distributions in LL-37 DPPG nanodiscs. Shown above is the mass spectrum for DPPG nanodiscs containing LL-37 peptide added at a 3:1 peptide:nanodisc ratio (**Figure 3**). Mass defect analysis was applied to quantify the number of LL-37 peptides

incorporated into the nanodisc. The mass defect is calculated as the remainder of the measured nanodisc mass divided by the mass of the DPPG lipid. Because the DPPG lipid mass is the divisor, the mass defect value is unaffected by the number of lipids incorporated. Any changes to the mass defect are therefore due to other compounds in the nanodisc complex; here, we observe LL-37 incorporation in increments of 0.214 in the mass defect.

The data were double-deconvolved using an empty DPPG nanodisc spectrum as the kernel. This mathematically separates the natural polydispersity of DPPG nanodiscs from variation due to LL-37 incorporation, greatly improving the resolution of peak distributions with distinct mass defects. The peaks distributed at mass defects of 0.4, 0.6, and 0.85 correspond to nanodiscs with two, three, and four LL-37 peptides incorporated, respectively.

Granted, the polydispersity of the nanodiscs is not a significant obstacle to analysis in this case, with mass defect peaks clearly appearing at the same values even without double deconvolution. Indeed, the relatively monodisperse nature of nanodiscs is a commonly cited motivation for their use.[8] However, the demonstrated ability to mathematically reduce polydispersity remains relevant for systems where it can be a problem. Possible future applications include the reconstitution of membrane proteins in saposin lipid nanoparticles (SapNPs), the modular nature of which naturally leads to increased mass variation.[11]

## Conclusions

Double deconvolution provides a novel approach to resolving complex mass spectra by applying a second deconvolution step, allowing users to mathematically separate mass-determining factors. Because any mass spectra could be used as the data or kernel, DoubleDec is widely applicable. We have demonstrated its use in isolating zinc binding in rhodopsin and improving the resolution of peptide incorporation peaks in lipoprotein nanodiscs.

A typical DoubleDec workflow is described as follows (**Figure 4**).

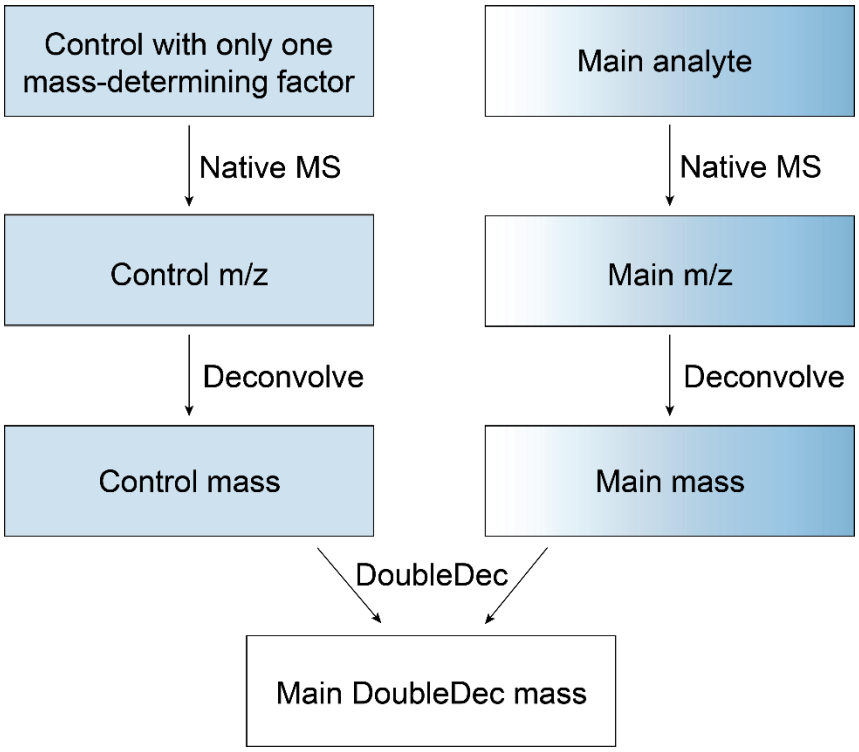


**Figure 4.** Flowchart depicting the application of DoubleDec to an unspecified native MS experiment.

To separate mass-determining factors, users first consider the possible factors that may exist in their sample. DoubleDec has limited use in samples where only one factor is present, but the complexity of biological systems makes such samples relatively rare. Users then create a control sample that varies in mass only due to the factor they wish to remove. They analyze both the control and data samples through native MS, and the deconvolved mass spectrum of the control sample is used as the kernel for double deconvolution of the data.

The flexibility of double deconvolution further expands on the universal nature of UniDec's Bayesian approach to MS data analysis. DoubleDec shows great promise in a wide range of MS experiments, and we expect its applicability to only grow over time as native MS systems become increasingly complex.

## References


1. Konermann, L.; Ahadi, E.; Rodriguez, A. D.; Vahidi, S. Unraveling the Mechanism of Electrospray Ionization. *Anal. Chem.* **2013**, *85* (1), 2–9.
2. Leney, A. C.; Heck, A. J. R. Native Mass Spectrometry: What is in the Name? *J. Am. Soc. Mass Spectrom.* **2017**, *28* (1), 5–13.
3. Marty, M. T.; Baldwin A. J.; Marklund, E. G.; Hochberg, G. K. A.; Benesch, J. L. P.; Robinson C. V. Bayesian Deconvolution of Mass and Ion Mobility Spectra: From Binary Interactions to Polydisperse Ensembles. *Anal. Chem.* **2015**, *87*, 4370-4376.
4. Lucy, L. B. An iterative technique for the rectification of observed distributions. *Astron. J.* **1974**, *79* (6).
5. Richardson, W. H. Bayesian-based iterative method of image restoration. *J. Opt. Soc. Am.* **1972**, *62* (1), 55.
6. Handa, J. T.; Cano, M.; Wang, L.; Datta, S.; Liu, T. Lipids, Oxidized Lipids, Oxidation-Specific Epitopes, and Age-Related Macular Degeneration. *Biochim. Biophys. Acta Mol. Cell Biol. Lipids* **2017**, *1862* (4), 430–440.
7. Xhindoli, D.; Pacor, S.; Benincasa, M.; Scocchi, M.; Gennaro, R.; Tossi, A. The Human Cathelicidin LL-37 - A Pore-Forming Antibacterial Peptide and Host-Cell Modulator. *Biochim. Biophys. Acta* **2016**, *1858* (3), 546–566.
8. Sligar, S. G.; Denisov, I. G. Nanodiscs: A Toolkit for Membrane Protein Science. *Protein Sci.* **2021**, *30* (2), 297–315.
9. Norris, C.; Keener, J. E.; Perera, S. M. D. C.; 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*. **2020**. DOI: 10.1016/j.ijms.2020.116477.
10. Kostelic, M. M.; Zak, C. K.; Jayasekera, H. S.; Marty, M. T. Assembly of Model Membrane Nanodiscs for Native Mass Spectrometry. *Anal. Chem.* **2021**, *93* (14), 5972–5979.
11. Flayhan, A.; Mertens, H. D. T.; Ural-Blimke, Y.; Molledo, M. M.; Svergun, D. I.; Löw, C. Saposin Lipid Nanoparticles: A Highly Versatile and Modular Tool for Membrane Protein Research. *Structure* **2018**, *26* (2), 345–355.