Received: 17 March 2022 | Revised: 7 April 2022 | Accepted: 8 April 2022

DOI: 10.1002/pro.4333

FOR THE RECORD

THE PROTEIN SOCIETY WILEY

Complementing machine learning-based structure predictions with native mass spectrometry

Timothy M. Allison[1] | **Matteo T. Degiacomi**[2] | **Erik G. Marklund**[3] | **Luca Jovine**[4] | **Arne Elofsson**[5] | **Justin L. P. Benesch**[6] | **Michael Landreh**[7]

[1]Biomolecular Interaction Centre, School of Physical and Chemical Sciences, University of Canterbury, Christchurch, New Zealand

[2]Department of Physics, Durham University, Durham, UK

[3]Department of Chemistry – BMC, Uppsala University, Uppsala, Sweden

[4]Department of Biosciences and Nutrition, Karolinska Institutet, Huddinge, Sweden

[5]Science for Life Laboratory and Department of Biochemistry and Biophysics, Stockholm University, Solna, Sweden

[6]Department of Chemistry, University of Oxford, Oxford, UK

[7]Department of Microbiology, Tumor and Cell Biology, Karolinska Institutet – Biomedicum, Stockholm, Sweden

Correspondence
Michael Landreh, Department of Microbiology, Tumor and Cell Biology, Karolinska Institutet – Biomedicum, Tomtebodavägen 9, 171 65 Stockholm, Sweden.
Email: michael.landreh@ki.se

Funding information
Cancerfonden, Grant/Award Number: 19 0480; Vetenskapsrådet, Grant/Award Number: 2019-01961

Review Editor: Nir Ben-Tal

Abstract

The advent of machine learning-based structure prediction algorithms such as AlphaFold2 (AF2) and RoseTTa Fold have moved the generation of accurate structural models for the entire cellular protein machinery into the reach of the scientific community. However, structure predictions of protein complexes are based on user-provided input and may require experimental validation. Mass spectrometry (MS) is a versatile, time-effective tool that provides information on post-translational modifications, ligand interactions, conformational changes, and higher-order oligomerization. Using three protein systems, we show that native MS experiments can uncover structural features of ligand interactions, homology models, and point mutations that are undetectable by AF2 alone. We conclude that machine learning can be complemented with MS to yield more accurate structural models on a small and large scale.

KEYWORDS

integrative modeling, machine learning, protein structure prediction, structural proteomics

1 | INTRODUCTION

Machine learning (ML)-based algorithms have been hailed as the solution to the protein structure prediction problem and are already being used to predict structures across entire proteomes.[1,2] For example, using protein sequence data as the only user input, AF2[3] can generate models of ordered, monomeric proteins that rival in quality experimentally derived structures,[4] which can be assembled into complexes using AF2 Multimer.[5]

However, it is important to remember that the models are generated according to user-provided input. For example, AF2 Multimer does not suggest an oligomeric state; instead, the stoichiometry for the model must be specified along with the sequences of the components. Moreover, AF2 may propose seemingly plausible models for a protein interaction even if this is not biologically relevant, for example, because the proteins are in different cellular compartments. Furthermore, using AF2 to predict interactions involving dynamic regions,[6] ligand binding sites, or point mutations,[7] all of which are major focal points of structural biology, remains challenging.[8] In these cases, additional structural data may be required to assess the validity of the computed structures, for example, from X-ray crystallography and cryo-EM. However, obtaining such data is challenging, resulting in a need for alternative strategies.

Mass spectrometry (MS), with its rapidly expanding structural biology toolbox,[9] can provide structural data that are directly complementary to ML (Figure 1a). Despite not being a stand-alone structure determination technique, MS offers a wealth of information for hybrid structural biology approaches.[9] It has a well-developed capacity to provide proteoform primary structure information, such as post-translational modifications, via MS-sequencing. In combination with in-solution labeling methods such as hydrogen-deuterium exchange (HDX), MS can inform about local structural dynamics. Native MS, where the non-covalent interfaces in macromolecules are preserved in the experiment, is widely used to determine oligomeric states, which is of particular importance when building models of protein complexes. Crosslinking and ion mobility (IM) measurements reveal the spatial arrangements of components in a protein complex. Unlike other biophysical methods, MS offers the crucial advantage of being able to provide structural data on the proteome scale. For example, proteome-wide crosslinking studies can help to filter biologically irrelevant interactions.[14] Collision-cross sections (CCSs, effectively 2D-projections of the structures) can be calculated for entire model proteomes and used to filter complex architectures by IM-MS.[15] Last, hybrid MS methods, such as NativeOmics, can reveal direct connections between primary and quaternary structure variations, as well as help to identify ligands or cofactors that may be structurally and functionally important.[16]

We, therefore, asked whether native MS, which is widely employed to study protein interactions, can be readily used to assess the plausibility of structural models generated by AF2. For this purpose, we selected three protein complexes whose interactions involve disordered regions, ligands, and point mutations. In all three cases, the native MS data show specific effects that are not detectable by AF2 alone, illustrating the complementarity of the two approaches.

2 | RESULTS AND DISCUSSION

As a first example, we tested the ability of AF2 to predict the structure of dihydroorotate dehydrogenase (DHODH), a mitochondrial enzyme involved in uracil synthesis. Inhibition of DHODH selectively kills cancer cells, making it a prime target for the development of novel therapeutics.[17] When using AF2 to predict the structure of the soluble domain of DHODH, the result is nearly indistinguishable from the available X-ray structures,[18] with a Cα root-mean-square deviation (RMSD) of 0.5 Å^2 (Figure 1b), with the exception that the predicted structure contains a central cavity which in the experimental structures is occupied by the cofactor flavin mononucleotide (FMN). In fact, overlaying the ligand-binding sites of the AF2 prediction and the X-ray structure reveals a nearly identical arrangement of the residues that coordinate FMN (Figure S1a). We have previously used native MS to assess the relationship between ligand binding and folding of DHODH[10] and found that the protein exists mostly in the holo-form. We also detected a small apo population with higher charge states, indicating unfolding in solution. Indeed, IM-MS revealed that FMN-bound protein adopts a compact conformation, whereas the FMN-free protein is largely unfolded, as evident from the CCS distributions of the 13+ charge state of both populations (Figure 1b).[10] When we computed the CCSs of the experimental and the predicted structures, we found them to be virtually identical (Figure 1b). Taken together, we find that AF2 predicts the fold of the holo-form of DHODH even without the co-factor. The recently solved crystal structure of the FMN-free form of the homologous DHODH from *Trypanosoma brucei* reveals backbone re-arrangements in the FMN pocket which result in increased local flexibility.[19] Native MS shows that the human protein cannot maintain the correct conformation in the absence of FMN in MS, which strongly supports that FMN is required to adopt a stable conformation. This discrepancy could arise from co-factor-bound proteins being part of the AF2 training set, yet the co-factors themselves are not considered in the prediction. Although alternative computational tools may be used to incorporate ligands in AF2 models, the connection between binding and folding is not considered in the predictions. As shown for DHODH, native MS can inform about the role of the co-factor in promoting the correct fold of DHODH, a role that is not evident from the ML-based prediction alone.

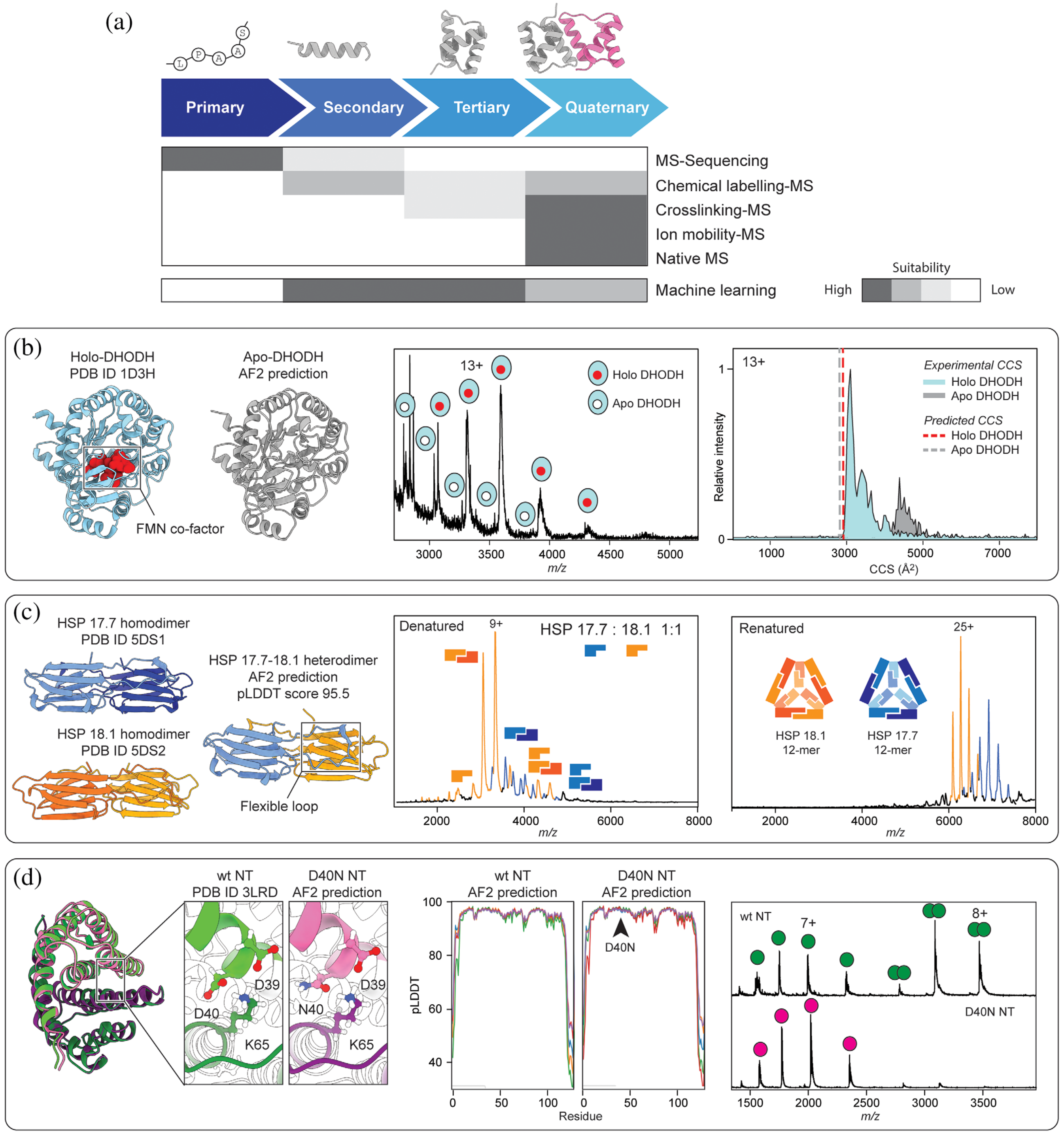

FIGURE 1 (a) The structural mass spectrometry (MS) toolbox offers information that is directly complementary to machine learning-based structure prediction. MS can inform about proteoforms (MS sequencing), structural dynamics (HDX-MS), the spatial arrangements of proteins in a complex (ion mobility and crosslinking MS), and oligomeric states (native MS). (b) Left: Experimental and predicted structures for holo- (left) and apo-DHODH show near-identical three-dimensional folds. Middle: Native MS reveals the presence of a small population of apo protein.[10] Right: IM-MS of the 13+ charge states of apo- and holo-DHODH shows that the protein with co-factor has a native-like CCS, whereas the protein without co-factor is unfolded. (c) Left: Crystal structures for the HSP 17.7 and 18.1 homodimers are virtually indistinguishable from the AF2-predicted heterodimer. Native MS of a mixture of HSP 17.7 and 18.1 under denaturing conditions (middle) and after refolding (right) reveals that no heterodimer formation takes place.[11] (d) Left: AF2 predicts that the D40N mutant of MaSp1 NT forms a homodimer that closely resembles the dimeric structure of wt MaSp1 NT, despite showing partial loss of the D39/D40/K65 salt bridge. Middle: pLDDT plots indicate that the D40N mutation does not affect the prediction confidence for the subunits in the NT dimer. Right: Native MS analysis of both NT variants at pH 6.0 shows that the D40N mutation abolishes NT dimerization.[12] All AF predictions were carried out using ColabFold V1.5, using AF2 Multimer 2.2.[13] Predictions were run with the AMBER refinement step but without templates. The MS data for all three proteins were taken from each respective reference publications.

Next, we asked whether native MS and AF2 could capture the effect of a flexible segment on the formation of a protein complex. For this purpose, we turned to the paralogous small heat shock proteins 17.7 and 18.1 from *Pisum sativum*. Both form highly similar homodimeric protomers via a conserved dimerization interface and swapping of a flexible loop, which then assemble into tetrahedral dodecamers.[11] Using AF2, we could correctly predict both homodimers (Figure 1c), and also the hypothetical HSP 17.7–18.1 heterodimer with a per-residue confidence score (pLDDT, which corresponds to the model's predicted score on the Local Distance Difference Test and measures distances between atom pairs[20]) equal to those of the homodimers, and a Cα RMSD of 0.73 and 0.66 Å^2 for the 17.7 and 18.1 heterodimer, respectively. Similarly, the predicted alignment error plots show no discernable difference (Figure S1). We also used the pTM score in Alphafold Multimer 2.2[21] to assess the quality of the interface predictions and found that the heterodimer scored essentially the same (0.891 ± 0.005) as the homodimers (0.879 ± 0.007 and 0.882 ± 0.007). However, the proteins do not coassemble in vivo, despite being colocalized and coexpressed to high concentrations during heat stress.[12] Upon refolding a mixture of denatured HSP 17.7 and 18.1, native MS revealed homodimer formation and assembly into dodecamers, while at the same time suggesting that, despite no direct steric hindrance and seemingly compatible dimer interfaces, heterodimeri zation is practically impossible (Figure 1c). This preference arises from an inability of the different monomers to bind each other's flexible loops due to differences in non-interfacial residues, which provides a penalty for hetero-oligomerization.[11] Such a preference of homo- over hetero-oligomerization is likely a wide-spread phenomenon.[11] However, as it is mediated by a flexible region outside of the well-defined dimerization surface, it has no significant impact on the confidence of the AF2 model, but can be readily detected by MS.

Last, we investigated the ability of MS and AF2 to capture the impact of point mutations on protein complex formation. Mutations that do not introduce significant steric hindrance yield near-identical AF2 structures[7] that nonetheless show measurable differences in stability.[8] However, it is unclear to what extent AF2 can inform about the effect of mutations on protein–protein interactions. We chose the N-terminal domain (NT) of the spider silk protein Major ampullate Spidroin 1 (MaSp1) from *Euprosthenops australis*, which is monomeric above, and dimeric below, pH 6.5.[12,22] This pH sensitivity is in part due to a conserved salt bridge between D39/D40 and K65 on the opposing subunit.[23,24] We used AF2 to predict the structure of the dimeric wild-type protein, as well as a point mutant with a weakened salt bridge, D40N (Figure 1c). Importantly, AF2 does not explicitly address the protonation state of ionizable residues, but may indirectly reflect the interactions observed under the solution conditions used to solve the structures included in the training set. Comparison of the pLDDT scores of the top five models for each variant showed no discernable differences (Figure 1d) with a Cα RMSD of 0.2 Å^2, indicating highly similar structures. Native MS analysis of both proteins at pH 6.0, on the other hand, showed that the D40N mutation abolished dimerization nearly completely (Figure 1d).[12] In summary, mutating aspartate 40 to asparagine does not introduce structural changes or steric clashes and does not appear to have notable consequences for the F2 model of the dimer. The impact of losing this salt bridge on dimer formation, therefore, requires experimental validation, such as through native MS analysis.

3 | CONCLUSIONS

Here, we examined the ability of MS to provide complementary information to ML-based structure predictions of protein complexes. While AF2 predictions are generally highly accurate, they do not specifically address the influence of bound ligands, flexible regions, and point mutations on protein interactions. Native MS, on the other hand, does not provide structural details but can capture a wide range of protein interactions with a single measurement. Of particular importance for structure prediction is the ability of MS to provide accurate information on protein oligomeric states. While MS is unrivaled in the detail of the mass measurements, reliable mass measurement of multimeric stoichiometries can be obtained from various alternative techniques, opening even more ways to complement ML predictions. Going forward, MS should be combined with ML either by defining the modeling question a priori using MS data (MS/AI) or by using MS data to identify a likely model a posteriori (AI/MS). We anticipate that whole-proteome structural MS data, and even mass measurements in physiological solutions, such as analytical ultracentrifugation and small-angle X-ray scattering, but also new methods like mass photometry,[25] could be incorporated into large-scale ML predictions, for example in the form of constraints, to generate accurate structural maps of the entire cellular environment.

AUTHOR CONTRIBUTIONS

Timothy M. Allison: Conceptualization (equal); writing – review and editing (equal). **Matteo T. Degiacomi:** Conceptualization (equal); writing – review and editing (equal). **Erik G. Marklund:** Conceptualization (equal);

writing – review and editing (equal). **Luca Jovine:** Conceptualization (supporting); writing – review and editing (equal). **Arne Elofsson:** Conceptualization (supporting); writing – review and editing (equal). **Justin L. P. Benesch:** Conceptualization (supporting); writing – review and editing (equal). **Michael Landreh:** Conceptualization (equal); project administration (lead); writing – original draft (lead); writing – review and editing (equal).

ORCID

Michael Landreh https://orcid.org/0000-0002-7958-4074

REFERENCES

1. Baek M, DiMaio F, Anishchenko I, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science. 2021;373:6557.
2. Tunyasuvunakool K, Adler J, Wu Z, et al. Highly accurate protein structure prediction for the human proteome. Nature. 2021;596:590–596. https://doi.org/10.1038/s41586-021-03828-1.
3. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596:583–589. https://doi.org/10.1038/s41586-021-03819-2.
4. Zweckstetter M. NMR hawk-eyed view of AlphaFold2 structures. Protein Sci. 2021;30:2333–2337. https://doi.org/10.1002/pro.4175.
5. Bryant P, Pozzati G, Elofsson A. Improved prediction of protein-protein interactions using AlphaFold2. Nat Commun. 2022;13:1265.
6. Ruff KM, Pappu RV. AlphaFold and implications for intrinsically disordered proteins. J Mol Biol. 2021;433:167208. https://doi.org/10.1016/j.jmb.2021.167208.
7. Buel GR, Walters KJ. Can AlphaFold2 predict the impact of missense mutations on structure? Nat Struct Mol Biol. 2022;29:1–2. https://doi.org/10.1038/s41594-021-00714-2.
8. Akdel M, Pires DEV, Pardo EP, et al. A structural biology community assessment of AlphaFold 2 applications. BioRxiv. 2021. https://doi.org/10.1101/2021.09.26.461876.
9. Lössl P, van de Waterbeemd M, Heck AJ. The diverse and expanding role of mass spectrometry in structural and molecular biology. EMBO J. 2016;35:2634–2657. https://doi.org/10.15252/embj.201694818.
10. Costeira-Paulo J, Gault J, Popova G, et al. Lipids shape the electron acceptor-binding site of the peripheral membrane protein dihydroorotate dehydrogenase. Cell Chem Biol. 2018;25:309–317.e4. https://doi.org/10.1016/j.chembiol.2017.12.012.
11. Hochberg GKA, Shepherd DA, Marklund EG, et al. Structural principles that enable oligomeric small heat-shock protein paralogs to evolve distinct functions. Science. 2018;359:930–935.
12. Landreh M, Askarieh G, Nordling K, et al. A pH-dependent dimer lock in spider silk protein. J Mol Biol. 2010;404:328–336. https://doi.org/10.1016/j.jmb.2010.09.054.
13. Mirdita M, Ovchinnikov S, Steinegger M. ColabFold - making protein folding accessible to all. BioRxiv. 2021. https://doi.org/10.1101/2021.08.15.456425.
14. Liu F, Rijkers DTS, Post H, Heck AJR. Proteome-wide profiling of protein assemblies by cross-linking mass spectrometry. Nat Methods. 2015;12:1179–1184. https://doi.org/10.1038/nmeth.3603.
15. Marklund EG, Degiacomi MT, Robinson CV, Baldwin AJ, Benesch JLP. Collision cross sections for structural proteomics. Structure. 2015;23:791–799. https://doi.org/10.1016/j.str.2015.02.010.
16. Gault J, Liko I, Landreh M, et al. Combining native and 'omics' mass spectrometry to identify endogenous ligands bound to membrane proteins. Nat Methods. 2020;17:505–508. https://doi.org/10.1038/s41592-020-0821-0.
17. Ladds MJGW, van Leeuwen IMM, Drummond CJ, et al. A DHODH inhibitor increases p53 synthesis and enhances tumor cell killing by p53 degradation blockage. Nat Commun. 2018;9:1107.
18. Reis RAG, Calil FA, Feliciano PR, Pinheiro MP, Nonato MC. The dihydroorotate dehydrogenases: Past and present. Arch Biochem Biophys. 2017;632:175–191. https://doi.org/10.1016/j.abb.2017.06.019.
19. Kubota T, Tani O, Yamaguchi T, et al. Crystal structures of FMN-bound and FMN-free forms of dihydroorotate dehydrogenase from Trypanosoma brucei. FEBS Open Bio. 2018;8:680–691. https://doi.org/10.1002/2211-5463.12403.
20. Mariani V, Biasini M, Barbato A, Schwede T. lDDT: A local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics. 2013;29:2722–2728. https://doi.org/10.1093/bioinformatics/btt473.
21. Richard Evans R, O'Neill M, Pritzel A. Protein complex prediction with AlphaFold-Multimer. BioRxiv. 2021. https://doi.org/10.1101/2021.10.04.463034.
22. Gaines WA, Sehorn MG, Marcotte WR. Spidroin N-terminal domain promotes a ph-dependent association of silk proteins during self-assembly. J Biol Chem. 2010;285:40745–40753. https://doi.org/10.1074/jbc.M110.163121.
23. Askarieh G, Hedhammar M, Nordling K, et al. Self-assembly of spider silk proteins is controlled by a pH-sensitive relay. Nature. 2010;465:236–238. https://doi.org/10.1038/nature08962.
24. Kronqvist N, Otikovs M, Chmyrov V, et al. Sequential pH-driven dimerization and stabilization of the N-terminal domain enables rapid spider silk formation. Nat Commun. 2014;5:3254.
25. Young G, Hundt N, Cole D, et al. Quantitative mass imaging of single biological macromolecules. Science. 2018;360:423–427. https://doi.org/10.1126/science.aar5839.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Allison TM, Degiacomi MT, Marklund EG, Jovine L, Elofsson A, Benesch JLP, et al. Complementing machine learning-based structure predictions with native mass spectrometry. Protein Science. 2022; 31(6):e4333. https://doi.org/10.1002/pro.4333