Erschienen in: Nature Methods ; 11 (2014), 4. - S. 403-406
https://dx.doi.org/10.1038/nmeth.2841

A mass spectrometry–based hybrid method for structural modeling of protein complexes

Argyris Politis[1,4,5], Florian Stengel[2,5], Zoe Hall[1], Helena Hernández[1], Alexander Leitner[2], Thomas Walzthoeni[2], Carol V Robinson[1] & Ruedi Aebersold[2,3]

We describe a method that integrates data derived from different mass spectrometry (MS)-based techniques with a modeling strategy for structural characterization of protein assemblies. We encoded structural data derived from native MS, bottom-up proteomics, ion mobility–MS and chemical cross-linking MS into modeling restraints to compute the most likely structure of a protein assembly. We used the method to generate near-native models for three known structures and characterized an assembly intermediate of the proteasomal base.

Cells contain macromolecular assemblies, which are composed of physically interacting proteins. Elucidating the structure and dynamics of these assemblies are primary goals of structural biology.

Recently, analysis of protein complexes using hybrid methods has garnered great interest[1–3], enabling insights for systems that remain refractory to structure determination by a single method[4]. Among the methods that contribute to structural analyses, structural MS is generally applicable and requires only small sample amounts. Different types of MS measurements can provide multiple and orthogonal data sets for a specific protein complex. Label-free, quantitative bottom-up analyses by liquid chromatography–tandem MS (LC-MS/MS) define the composition and relative abundance of the complex subunits. Native MS of intact protein complexes and their subcomplexes provides information on the overall stoichiometry and protein-protein interactions. MS coupled with ion mobility (IM), IM-MS, elucidates protein architectures and dynamics by measuring their collisional cross-sections (CCSs)[5,6]. Chemical cross-linking coupled with MS (CX-MS) technology identifies protein subunit interfaces[7]. Although the utility of the individual techniques has been documented, combining information from all four MS-based approaches with modeling has not been reported to our knowledge.

Here we describe a generic hybrid structural biology method that integrates orthogonal data sets for the same protein complex generated by native MS, label-free quantification (LFQ) by LC-MS/MS, IM-MS and CX-MS. This hybrid method differs from other approaches because of its ability to generate orthogonal data sets and to computationally integrate diverse MS data sets with different levels of resolution and information content from the same sample. Overall, the method enables accurate prediction of multiprotein and heterogeneous complexes when high-resolution information of the individual subunits is used, and it consists of experimental techniques that require only low microgram sample amounts and that exhibit high measuring speed and tolerance for heterogeneous sample environments[8].

The method involves four steps: (i) protein purification and data collection by the respective MS technique (aliquots of the purified protein complex are first analyzed by LFQ and CX-MS experiments and then, after buffer exchange, IM-MS and native MS (Online Methods)); (ii) encoding MS data into restraints; (iii) structure prediction by iterative sampling and scoring of models; and (iv) ensemble analysis to generate most likely structures (**Fig. 1a** and Online Methods).

We developed and benchmarked the method using three well-characterized complexes exhibiting distinct topologies: methane monooxygenase hydroxylase (MMOH) from *Methylococcus capsulatus*, toluene/*o*-xylene monooxygenase hydroxylase (ToMOH) from *Pseudomonas stutzeri* and urease from *Klebsiella aerogenes* (Online Methods, **Supplementary Note 1** and **Supplementary Fig. 1**). Native MS allowed us to determine the stoichiometry of the complexes and their subunit connectivities[5] (**Supplementary Fig. 2**). IM-MS added orientationally averaged CCSs[9], and CX-MS allowed us to identify high-confidence inter- and intra-protein interactions[10–12]. Using these MS-based restraints allowed sampling of complex models. Next we refined the models using an optimization step and ranked the models with a weighted scoring function. We selected representative structures from the pool of highly ranked models upon pairwise clustering of their α-carbon r.m.s. deviations (Cα RMSDs). A refinement step ensured physical interactions between subunits (Online Methods). For all complexes we found good agreement (RMSDs < 12 Å) of the best-scored models with their native structures (**Fig. 1b**,**c** and **Supplementary Figs. 3**–**7**).

To evaluate contributions of each restraint for predicting near-native structures, we carried out statistical tests using receiver operating characteristics (ROCs) (**Supplementary Note 2**).

[1]Department of Chemistry, University of Oxford, Oxford, UK. [2]Department of Biology, Institute of Molecular Systems Biology, Eidgenossische Technische Hochschule (ETH) Zurich, Zurich, Switzerland. [3]Faculty of Science, University of Zurich, Zurich, Switzerland. [4]Current address: Department of Life and Health Sciences, School of Biomedical Sciences, University of Ulster, Londonderry, UK. [5]These authors contributed equally to this work. Correspondence should be addressed to R.A. (aebersold@imsb.biol.ethz.ch) or C.V.R. (carol.robinson@chem.ox.ac.uk).

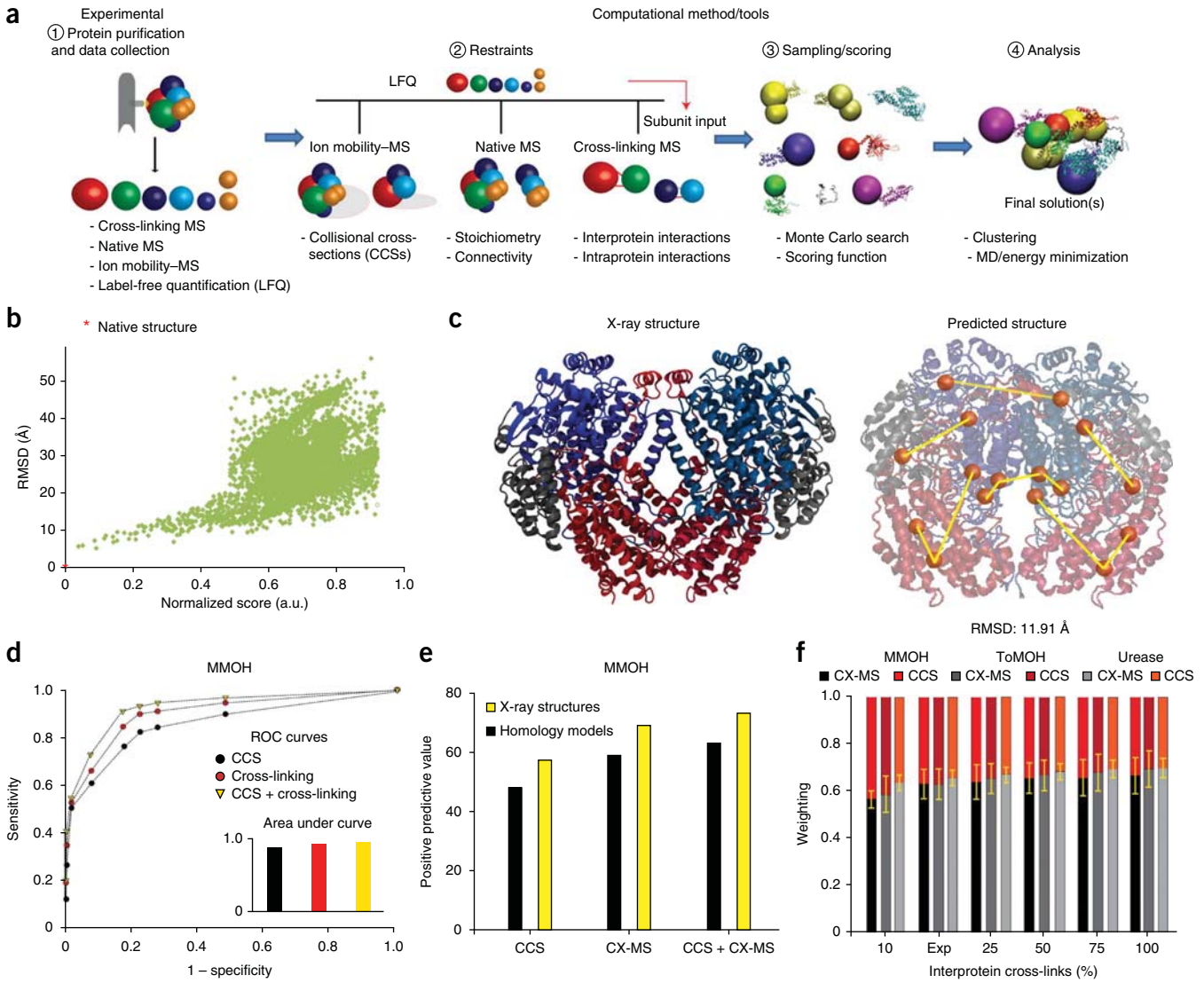

Figure 1 | Workflow and benchmark of a hybrid method for structure determination of protein assemblies using complementary MS data. (**a**) The workflow is composed of four steps. (1) The complex of interest is purified, either by a recombinant expression system or by affinity purification, and analyzed by four complementary MS-based approaches: bottom-up proteomics (LFQ), native MS, IM-MS and CX-MS. (2) The acquired data are translated into restraints, which provide information about the overall shape of subunits and subcomplexes (IM-MS), their stoichiometry and connectivity (native MS, LFQ) and interprotein proximities (CX-MS). (3) Models are generated by sampling the conformational space using a Monte Carlo search (>10,000 models), which is followed by a refinement step and evaluation. (4) Clustering of the best-scoring models determines the final solution(s). (**b**) The structural similarity of the models to the native structure is evaluated using their pairwise r.m.s. deviation (RMSD). (**c**) A representative structure of the best-scored ensemble of structures for MMOH oligomer (6-mer) reveals good agreement with an X-ray structure. (**d**) ROC curves were used to assess the accuracy and confidence levels of all restraints, individually and combined. Sensitivity is TP/(TP + FN), and specificity is TN/(TN + FP), where TP is true positive, FP is false positive, FN is false negative and TN is true negative. (**e**) Positive predictive values (TP/(TP + FP)) were calculated for all restraints, individually and combined, for the benchmarked complexes. (**f**) Weighting of the scoring function that accounts for both IM-MS and CX-MS restraints. The probability of identifying TPs is plotted for each restraint against the percentage of interprotein cross-links available. Errors bars, s.d. Exp, experimental data.

The plotted ROC curves and their predictive values show that combining restraints from IM-MS and CX-MS increased predictability (by ~10%; **Fig. 1d** and **Supplementary Figs. 8–12**). Next we assessed the impact on predictability when partial or no high-resolution structures were available. The results showed a decrease in predictability (by ~10%) when only homology models were used (**Supplementary Table 1**). If no high-resolution subunit information is available or can be computed, predictability will be substantially reduced. However, combining restraints still increased the predictive power of the method (**Fig. 1e** and **Supplementary Table 2**).

We further assessed the individual contribution of CX-MS and IM-MS restraints to the scoring function by weighting their impact in a training set of complexes. To optimize weighting, we calculated true positives for varied degrees of input data (Online Methods). We defined a true positive as a model with RMSD < 12 Å from the native structure. We calculated optimal weightings of 0.64 and 0.36 (± 0.05 s.d.) for CX-MS and IM-MS restraints, respectively (**Fig. 1f**). We henceforth used these values for complexes with unknown structures.

Next we applied our method to a biologically important assembly, the proteasome. Our structural knowledge of the intact

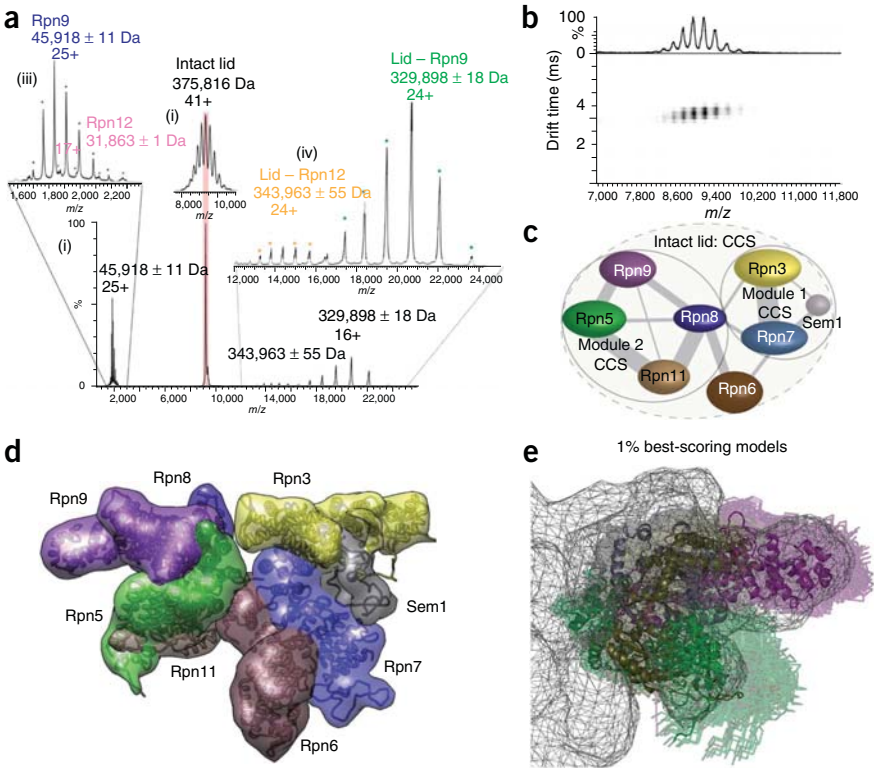

Figure 2 | Structural models of the intact proteasomal lid and two distinct submodules. (**a**) Mass spectra of the intact proteasomal lid and two of its subcomplexes as observed by native MS. Insets, assigned spectra of peripheral subunits Rpn9 and Rpn12 and of the remaining 'stripped' subcomplexes. (**b**) IM data plotted as drift time versus *m/z*. (**c**) Connectivity map of the proteasome lid generated by integrating subcomplex information from native MS with pairwise subunit contacts identified by CX-MS. (**d**) Three-dimensional model of the lid predicted by integrating all MS-derived restraints. The individual subunits are depicted as simulated density maps, generated by the UCSF Chimera package. (**e**) We overlaid the 1% best-scoring ensemble of structures (~100 conformations) of the Rpn5-Rpn8-Rpn9-Rpn11 module and subsequently docked them into a high-resolution EM density map. All models exhibited a marked similarity (RMSDs < 10 Å) to each other. The representative, best-scored model is shown as a cartoon.

complex is derived from two electron microscopy (EM) maps containing all but the smallest 'lid' subunit (Sem1)[4,13]. By isolating the proteasomal lid using pull-downs of tagged lid subunits and subjecting aliquots to the various MS methods, we confirmed successful enrichment of the lid subunits with LFQ (**Supplementary Fig. 13**). Exemplary mass spectra of the intact lid and its subcomplexes are shown (**Fig. 2a,b**, **Supplementary Figs. 14** and **15** and **Supplementary Table 3**) together with corresponding CCSs derived from IM-MS (**Fig. 2b** and **Supplementary Table 4**). We identified a total of 170 interlinks (28 nonredundant) between nonidentical subunits within the lid (**Supplementary Tables 5–9**).

Native and CX-MS data defined two distinct modules in the lid (Rpn5-Rpn8-Rpn9-Rpn11 and Rpn3-Rpn11-Sem1) (**Fig. 2c** and **Supplementary Figs. 15–17**). Using our hybrid method, we predicted models of the lid that were in good agreement with the corresponding EM maps[4,13] (**Fig. 2d**, **Supplementary Fig. 18** and

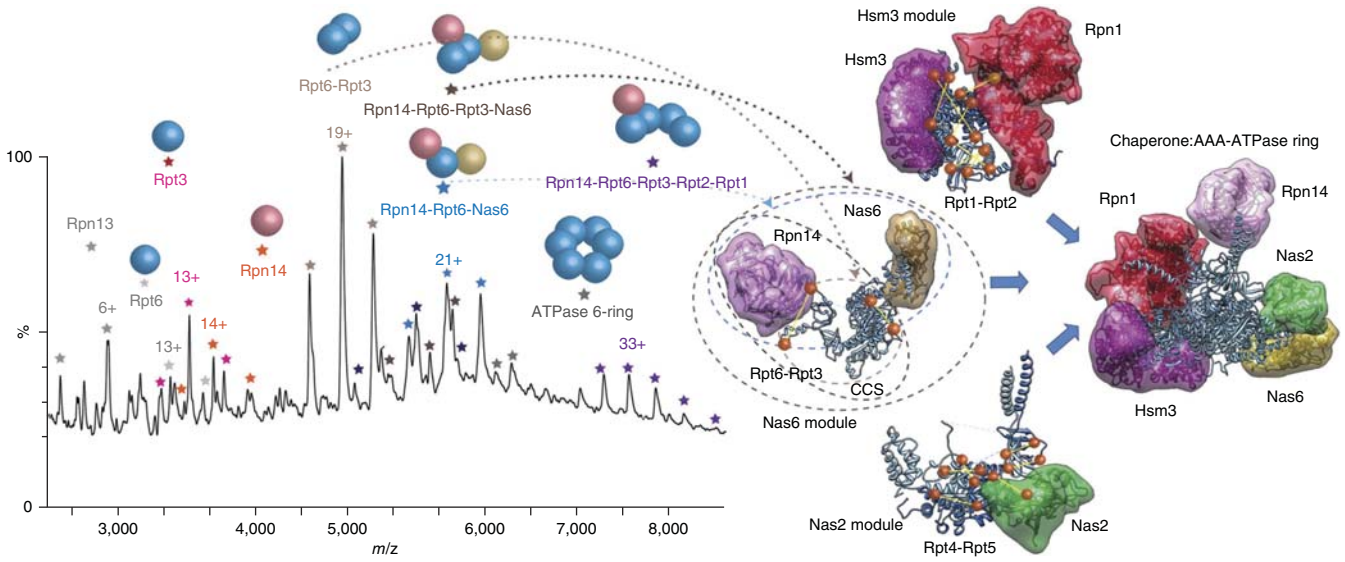

Figure 3 | Structural models of chaperone-base assembly intermediates involved in the formation of the proteasomal base complex. We generated homology models and collected X-ray crystal structures of all individual subunits (base subcomplex and associated PIP chaperones) for downstream analysis using the MS-restrained modeling strategy. A native MS spectrum from an Rpn14 pulldown (left) shows the intact Rpn14-Rpt6-Rpt3-Nas6 and subcomplexes thereof (stars indicate measured charge state series). We built a structural model for the Rpn14-Rpt6-Rpt3-Nas6 module (the best-scoring model of an ensemble of structures) combining native MS, IM-MS and CX-MS. We proposed a structural model of the assembly pathway of the proteasomal base consistent with the MS-derived data sets. Experimentally identified cross-links, subcomplexes and CCS measurements are indicated. Base-dedicated chaperones with their simulated density-map envelopes are shown.

Supplementary Table 10). We showed a marked similarity for the best-scoring ensemble of models of the Rpn5-Rpn8-Rpn9-Rpn11 module using hierarchical clustering (**Supplementary Fig. 19**) and by overlaying them onto the corresponding density map[13] (**Fig. 2e** and **Supplementary Figs. 19** and **20**). Interestingly, in our model we placed Sem1 in the density cleft formed between subunits Rpn3 and Rpn7 (**Fig. 2d**), which is consistent with data from recent studies using EM, MS, and deletion strains of Sem1 and Rpn15 (refs. 14,15).

Next we attempted to characterize assembly intermediates, which are notoriously challenging targets for classical structural biology methods. Molecular and biochemical studies have shown that the proteasomal base is assembled via a multistep process wherein precursors are transiently associated with proteasome-dedicated chaperones or proteasome-interacting proteins (PIPs). Despite some successes on smaller complexes[16,17], efforts to uncover high-resolution structures of intact assembly intermediates have failed, presumably owing to the heterogeneous and transient nature of these complexes[18].

The combined LFQ data from lid affinity pulldowns (**Supplementary Fig. 21**) indicated that in addition to all known 19S subunits, we detected the PIPs Hsm3, Rpn14, Nas2, Ubp6 and Nas6 (PSD10) that assist assembly of the base[18,19]. To probe these PIP-containing complexes, we used pulldowns from Rpn14- and Nas6-tagged cells. LFQ confirmed that the base subunits are the main interacting partners of these PIPs (Online Methods and **Supplementary Fig. 22**). Native MS revealed the intact Nas6-Rpt3-Rpt6-Rpn14 precursor as well as multiple stable subcomplexes thereof (**Fig. 3** and **Supplementary Figs. 23** and **24**). IM yielded the CCS of the Rpt3-Rpt6-Nas6 trimer, and CX-MS confirmed four unique high-confidence PIP-based inter-cross-links (**Supplementary Table 11**). These data together with crystallographic information on the Nas6-Rpt3 interface[16] allowed us to confidently predict a structural ensemble of the intact Nas6-Rpt3-Rpt6-Rpn14 precursor (**Fig. 3** and **Supplementary Table 12**).

We also detected multiple high-quality interlinks for the base ATPase hexamer (Rpt1–Rpt6), all in agreement with the proposed order of subunits[20] (Rpt1-Rpt2-Rpt6-Rpt3-Rpt4-Rpt5; **Supplementary Results** and **Supplementary Figs. 25** and **26**). Together with the known composition and stoichiometry of the precursors[18], this allowed us to propose a structural model for early steps in base assembly (**Fig. 3**). We further proposed, on the basis of LFQ and CX-MS data, the structural organization of other known intermediate precursors (Nas2-Rpt4-Rpt5, Hsm3-Rpt1-Rpt2-Rpn1 and Rpn2-Rpn13 modules) that act as building blocks for the formation of the base[18,19] (**Fig. 3** and **Supplementary Figs. 27**–**30**).

Overall, we developed, validated and applied a generic method consisting of complementary MS-based approaches and computational data integration for structural analysis of protein complexes. The computational data integration is available as **Supplementary Software**, and its Python package documentation is described in **Supplementary Note 3**. Because this hybrid method can be coupled to any purification protocol, provided expression levels are micromolar, we anticipate it will be very useful for probing heterogeneous assemblies, especially in the 50- to 300-kDa range that is challenging for current EM approaches.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

MMOH and ToMOH were a gift of S.J. Lippard (Massachusetts Institute of Technology). Urease from *K. aerogenes* was a gift from R.P. Hausinger (Michigan State University). This work was supported by funding from PROSPECTS (Proteomics Specification in Space and Time Grant HEALTH-F4-2008-201648) within the European Union 7th Framework Program (A.P., C.V.R. and R.A.) and from European Research Council advanced grants "Proteomics v3.0" (233226) and "IMPRESS" (268851) to R.A. and C.V.R. H.H. is funded by Medical Research Council programme grant (G1000819). F.S. is a Sir Henry Wellcome Fellow funded by the Wellcome Trust (grant 095951), and C.V.R. is funded by the Royal Society.

AUTHOR CONTRIBUTIONS

F.S. and A.P. conceived the study; F.S., A.P., C.V.R. and R.A. designed the research; A.P. performed all modeling and developed the software; F.S. carried out the experiments; Z.H. and H.H. performed part of the IM-MS and native MS experiments. A.L. and T.W. supported CX-MS experiments and analysis; F.S. and A.P. analyzed the data; A.P., F.S., C.V.R. and R.A. wrote the paper; all authors commented on and edited the final version of the paper.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Robinson, C.V., Sali, A. & Baumeister, W. *Nature* **450**, 973–982 (2007).
2. Alber, F. *et al. Nature* **450**, 683–694 (2007).
3. Stengel, F., Aebersold, R. & Robinson, C.V. *Mol. Cell. Proteomics* **11**, R111.014027 (2012).
4. Lasker, K. *et al. Proc. Natl. Acad. Sci. USA* **109**, 1380–1387 (2012).
5. Hall, Z., Politis, A. & Robinson, C.V. *Structure* **20**, 1596–1609 (2012).
6. Politis, A. *et al. PLoS ONE* **5**, e12080 (2010).
7. Leitner, A. *et al. Mol. Cell. Proteomics* **9**, 1634–1649 (2010).
8. Walzthoeni, T., Leitner, A., Stengel, F. & Aebersold, R. *Curr. Opin. Struct. Biol.* **23**, 252–260 (2013).
9. Ruotolo, B.T. *et al. Nat. Protoc.* **3**, 1139–1152 (2008).
10. Leitner, A., Walzthoeni, T. & Aebersold, R. *Nat. Protoc.* **9**, 120–137 (2014).
11. Walzthoeni, T. *et al. Nat. Methods* **9**, 901–903 (2012).
12. Rinner, O. *et al. Nat. Methods* **5**, 315–318 (2008).
13. Lander, G.C. *et al. Nature* **482**, 186–191 (2012).
14. Kao, A. *et al. Mol. Cell. Proteomics* **11**, 1566–1577 (2012).
15. Bohn, S. *et al. Biochem. Biophys. Res. Commun.* **435**, 250–254 (2013).
16. Nakamura, Y. *et al. Biochem. Biophys. Res. Commun.* **359**, 503–509 (2007).
17. Barrault, M.B. *et al. Proc. Natl. Acad. Sci. USA* **109**, E1001–E1010 (2012).
18. Saeki, Y. *et al. Cell* **137**, 900–913 (2009).
19. Roelofs, J. *et al. Nature* **459**, 861–865 (2009).
20. Tomko, R.J. Jr. *et al. Mol. Cell* **38**, 393–403 (2010).

ONLINE METHODS

Overall workflow. First, the protein complex of interest is purified, either by a recombinant expression system or by affinity purification and, if needed, subsequently enriched by centrifugal concentration. Then the sample is split and used for LFQ and CX-MS and, after buffer exchange, for IM-MS and native MS experiments. LFQ generates a list of subunits and their relative abundance present in the sample. Native MS of the intact complexes yields the composition and stoichiometry of protein complexes while further information is attained from gas-phase dissociation techniques such as collision-induced dissociation (CID), which reveals subunit interaction networks[5]. IM coupled with MS provides topological information in the form of an orientationally averaged CCS[9]. Furthermore, the CCSs of stable subcomplexes can be used to reveal the structures of the building blocks of a complex. We identified multiple high-confidence inter- and intraprotein interactions by applying isotopically labeled cross-linkers and searching and validating the identified cross-linked peptides against a database generated from the LFQ experiment using the xQuest and xProphet pipeline[11,12]. We used the identified cross-links as upper-bound distance restraints (35 Å) for structural modeling.

With the data encoded into spatial restraints in hand, we applied our computational strategy for structure determination of protein complexes. We first selected an appropriate representation scheme that best reflects the resolution of the available data. In order to be able to generate pseudoatomic models, we used high-resolution information of the individual subunits. These can be X-ray crystals, NMR structures or high-confidence homology models given available templates. We used the subunit list from LFQ to generate the structural input for the various subunits of the proteasomal assembly. For full exploitation of the cross-linking information (residue level), high-resolution structures should be available for the individual subunits within the complexes. We therefore generated homology models for all subunits for which no high-resolution structures are available. Sequence Id for the test case proteins was between 20% and 100% (**Supplementary Table 1**) and between 19% and 56% for the lid proteins (**Supplementary Table 10**), respectively.

Next we set out to build a large number of structural models of protein complexes from their building blocks. A critical part of sampling is to accurately determine the stoichiometry and copy number of subunits and subcomplexes within the intact assembly. We acquired this information by combining LFQ data with the native MS data of the intact complexes and additional subcomplexes identified by CID that allowed us to build structural models consistent with the experiments. We generated model structures that satisfy the input data using a Monte Carlo search algorithm and subsequently optimized through a conjugate gradient optimization. Then we scored the candidate models using a weighted scoring function, which encodes the three types of restraints. We selected the representative structures from the pool of highly ranked models upon pairwise clustering (described below). Finally, a flexibility step using energy minimization/molecular dynamics (MD) simulations allowed us to search for energetically favorable structures and eliminate potential steric clashes.

Protein purification. We used a training set of three well-characterized complexes exhibiting distinct topologies to develop and optimize our method. The complexes are (i) toluene/*o*-xylene monooxygenase hydroxylase from *P. stutzeri*[21] (ToMOH, PDB ID: 2INC; 212 kDa), an $\alpha_2\beta_2\gamma_2$ globular heterohexamer; (ii) methane monooxygenase hydroxylase from *M. capsulatus*[21] (MMOH, PDB ID: 1MTY; 251 kDa), a rectangular-shaped $\alpha_2\beta_2\chi_2$ complex; and (iii) urease from *K. aereogenes*[22] (PDB ID: 1KRA; 249 kDa for the apo enzyme), an $\alpha_3\beta_3\chi_3$ triangular-shaped assembly (**Supplementary Fig. 1**).

We purified the proteasome lid and its subcomplexes from RPNX-3xFlag strains (*MAT**a*** rpnX::RPNX-3xFlag-His3) essentially as described before[23]. Additionally, we performed control pulldowns for the proteasome-interacting proteins (PIPs) using the commercially available Tap-Tagged library[24].

Briefly, RPNX-3xFlag cells were cultured, lysed and pulled down with anti-Flag M2 agarose beads. We then subjected affinity-purified proteasomes to anion-exchange chromatography after treatment with high salt to promote dissociation of the 26S proteasome and before elution with Flag peptide. For enrichment of each subcomplex, we subjected the eluted samples to a 15–40% sucrose gradient, which was followed by fractionation and SDS-PAGE. Prior to MS analysis, we pooled and concentrated lower fractions using Vivaspin centrifugal concentrators (10K MWCO, Sartorius) followed by cross-linking or buffer exchange using Micro Bio-spin 6 columns (Bio-Rad) into ammonium acetate, pH 7.5, for the MS of intact assemblies and ion-mobility analysis.

We lysed and pulled down Tap-Tag strains with IgG beads (Sigma I5006) coupled to Dynabeads (M-270 Epoxy, 143.01, Invitrogen). We then washed the proteins bound to beads after IP three times with 50 mM HEPES, pH 7.1, 100 mM NaCl, 10 mM MgCl plus protease inhibitors (Roche), which was followed by a concentration step and MS analysis as described.

Cross-linking coupled to mass spectrometry (CX-MS). For cross-linking experiments, equimolar amounts of light and heavy isotopically labeled cross-linkers disuccinimidyl suberate (DSS)-d0/DSS-d12 (Creative Molecules) dissolved in dimethylformamide (DMF, Thermo Scientific) at a stock concentration of 25 mM were used. We added cross-linkers to the proteins at a final concentration of 1 mM and incubated the sample for 30 min at 37 °C with slight shaking before the cross-linking reaction was quenched with ammonium bicarbonate at a final concentration of 50 mM for 10 min at 37 °C. We then reduced (alkylated) and digested the proteins with trypsin using standard protocols followed by a SEC enrichment step before LC-MS/MS measurement on a Thermo LTQ Orbitrap XL or Thermo Orbitrap Elite mass spectrometer (LIT-Orbitrap, linear ion trap–Orbitrap) equipped with a standard nanoelectrospray source. We loaded the peptides onto a 75-µm-ID analytical column, packed in-house with Michrom Magic C18 material (3-µm particle size, 200-Å pore size). We separated the peptides at a flow rate of 300 nL min^{-1} ramping a gradient from 5% to 35% mobile phase B (water/acetonitrile/formic acid; 3:97:0.1). We set the ion source and transmission parameters of the mass spectrometer to a spray voltage of 2 kV, capillary temperature at 200 °C, capillary voltage at 60 V and tube lens voltage at 135 V. We operated the mass spectrometer in data-dependent mode, selecting up to five precursors from a MS1 scan (resolution = 60,000) in the range of m/z 350–1,600 for CID. We rejected singly and doubly charged

precursor ions and precursors of unknown charge states. CID was performed for 30 ms using 35% normalized collision energy and an activation q of 0.25. We activated the dynamic exclusion with a repeat count of 1, exclusion duration of 30 s, list size of 300 and a mass window of ±50 p.p.m. Ion target values were 1,000,000 (or maximum 500-ms fill time) for full scans and 10,000 (or maximum 200-ms fill time) for MS/MS scans, respectively.

We analyzed cross-linked peptides using the xQuest[12] and xProphet[11] software platforms, unless otherwise indicated. We considered only cross-links that scored a FDR of <0.05 after xProphet analysis. For some of the reciprocal PIP pulldowns and some of the recombinant 'test-case' protein samples, a valid FDR could not be calculated, as not enough decoy matches could be generated. In those cases, we considered as cutoff the absolute Id threshold of Id 25 (PIPs) or Id 18 (recombinant test cases) and a deltaScore of <0.95. We further analyzed all spectra by visual inspection in order to ensure good matches of ion series on both cross-linked peptide chains for the most abundant peaks.

Label-free quantification (LFQ). We performed LFQ using Progenesis 4.0 (Nonlinear Dynamics) by automatic alignment of total ion chromatograms of raw files, using imported pep.xml files from X!Tandem searches against the yeast UniProtKB/Swiss-Prot protein database. We then calculated protein abundances by taking the sum of MS1 raw abundances over all biological replicates and samples and corrected for the number of amino acids of each protein. We used the resulting identifications to generate the library for subsequent cross-linking searches and identification of subcomplexes in native MS experiments.

Nanoelectrospray mass spectrometry of intact complexes. We obtained mass spectra for MS and tandem MS of intact assemblies on a Q-ToF 2 (Waters/Micromass UK) modified for high-mass operation[25], using a previously described protocol to preserve noncovalent interactions[26], with the following instrumental parameters: nanoelectrospray capillary, 1,600 V; sample cone, 40 V; extractor cone, 0 V; ion transfer stage pressure, 9.5×10^{-3} bar and up to 35 µbar of argon in the collision cell. Voltage in the collision cell was at 25 V for MS and up to 200 V for tandem MS experiments. We externally calibrated spectra using a 33 mg mL^{-1} aqueous solution of cesium iodide (Sigma). We processed the acquired data with MassLynx software (Waters). The data are shown with minimal smoothing.

NanoES ion-mobility analysis (absolute measurements). We collected mass spectra and drift time (DT) profiles for absolute CCS measurements on a quadrupole–IM–time-of-flight (ToF) mass spectrometer in positive ion mode (Synapt G1 HDMS, Waters) with a custom-made 18-cm ion-mobility cell that has a radial RF ion confinement (radio frequency of 2.7 MHz and peak-to-peak amplitude of 200 V) and a linear voltage gradient to direct ions along the axis of transmission to the time-of-flight mass analyzer[27]. We acquired the measurements at 20 °C and at 0.994 torr using helium in the mobility cell and monitored the pressures with a calibrated absolute pressure transducer (MKS Baratron model 626A) connected directly to the ion-mobility cell. We kept the cone voltage at 60 V (or 15 V for a second series of experiments), extraction cone at 1 V, trap at 10 V (5 V) and bias at 20 V. Source pressure was ~5.7 mbar, trap and IMS at 4.9×10^{-2} mbar and 1.4 mbar, respectively, and ToF analyzer pressure at 2.3×10^{-6} mbar. We determined the Ω values directly from the slopes of DT versus reciprocal drift voltage plots[28,29], using drift voltages ranging from 50 to 200 V, where the difference in potentials between the entrance and exit electrodes denotes the drift voltage.

Spatial restraints. We converted the experimental data from the different MS approaches into restraints for subsequent modeling analysis. We used the LFQ data to define all potential members of the proteasomal assembly and the various native MS measurements were used to define overall stoichiometries of the intact protein complex and its various subcomplexes. From all MS data, we built an experimental tree of the proteasomal assembly (**Supplementary Fig. 16**). We subsequently used this tree to sample and score the generated models. In addition, we constructed an interaction map of all subunits within the complex by integrating native MS with identified binary interactions from CX-MS (**Supplementary Figs. 15** and **16b**). We also used the CCSs derived from IM as restraints, implemented as a harmonic function, to measure the closeness of fit between experiments and calculated CCSs for models. Finally, we used the confirmed high-quality cross-links as upper-bound distance restraints between the residues in proteins. We further segregated the cross-links into interprotein cross-links that specify distance restraints between the cross-linked residues in interacting subunits and intraprotein cross-links that were not used in this study to compute the models but that can be used to examine the consistency of atomic coordinates (crystal structures or homology models) with the identified cross-links.

Sampling and optimization. Generating an adequate number of models is a critical step of our approach. Here we built models of the subcomplexes observed in our experiments in a stepwise manner starting from the smallest subcomplex identified in our MS-based experiments (usually a dimer) and building up to the oligomeric state of the intact complex (for example, 6-mer for MMOH and ToMOH and 9-mer for urease). In order to adequately sample the conformational space of proteins, we utilized a Monte Carlo sampling approach guided by the connectivity restraints derived from MS-based experiments. We incorporated the MS connectivity restraint for use during sampling (http://salilab.org/imp/nightly/doc/). This restraint ensured that all subunits remained connected and also enabled evaluation of the ensemble of generated structures by their deviation to the experimental tree derived from MS and CX-MS data. Furthermore, the sampling explored only positions consistent with the overall stoichiometry (number of subunit copies and intersubunit connectivities) of the respective complex under investigation. This step was followed by a conjugate gradient optimization step as implemented in Integrative Modeling Platform (IMP; http://salilab.org/imp/)[30]. Overall, at each step we generated 10,000–20,000 model structures at the atomic level, depending on the size, shape and composition of the complex. Next we subjected these models to further analysis by measuring their closeness to the experimental data.

Scoring function. The scoring function captured the encoded information from the raw data and was used to score the candidate model structures. Along with the imposed optimization process, the restraints ensure consistency of the models generated with the experimentally available data. In the cases studied here, we first filtered our structures using the interaction maps constructed from native MS and LFQ data. Next we evaluated the structures consistent with the input data by penalizing the violation of restraints provided by the various types of structural information, namely CX-MS and CCS. We gave a penalty of a unit score to model structures for each violation of an identified residue-specific intersubunit cross-link. We implemented the CCS restraint as a harmonic function, where perfect agreement between the model and experimental CCS would take a value of 0 and violations of restraint would result in higher values[5]. Therefore, we used the CCS restraint as shown in the equation (1)

$$S_{\mathrm{CCS}} = \left(\frac{\mathrm{CCS}' - \mathrm{CCS}}{\sigma'} \right)^2 \tag{1}$$

where the S_{CCS} score is computed by the closeness of fit between the experimental (CCS′) and calculated (CCS) values. σ' denotes the experimental error in the data. In our experiments, the CCS accuracy, measured using a linear drift tube, is estimated to be <3%. Here, in order to ensure realistic errors, we used σ' of ±6%.

We expressed the scoring function as a probability density function of the Cartesian coordinates of the assembly proteins (C) given information (I) on a restrained feature, p_f (ref. 2).

$$p(C/I) = \prod_f p_f(C/I_f) \tag{2}$$

We can then write the overall scoring function as the logarithm of the probability density function

$$S(C) = -\ln \prod_f p_f(C/I_f) = \sum_f r_f(C) \tag{3}$$

Practically, we calculated the scoring function, $S(C)$, by summing individual restraints r with weights w.

$$S(C) = \sum_f w_f r_f \tag{4}$$

We used the weighting scoring scheme, which integrates information from CX-MS and IM-MS, to evaluate all structural models that satisfy the input restraints derived from LFQ and native MS. Adequate sampling is critical in order to exhaustively search the conformational space of structures fitting the data. For example, IMP makes use of Monte Carlo sampling algorithms to generate tens of thousands of random configurations. We then optimized these structures by simultaneously minimizing violations of input restraints. We achieved this using conjugate gradients, and simulated annealing molecular dynamics, which refine the position of particles[4,31,32]. Ideally, the global optimum corresponds to the native assembly structure.

Weighting. As discussed in the main text, we optimized the scoring function using the training set of complexes. Bringing together data from varied sources into a single scoring function introduces heterogeneities and inconsistencies, which can be tackled by weighting the impact of the different data sets. Moreover, each of these data sets has different error features associated with both the experimental methods and the computational approaches. Here we calculated the impact of each individual source of data as in equation (5), where $P_{(\mathrm{TP}/y)}$ denotes the probability of identifying true positives from a certain type of data and the sum of probabilities of all types is described by $\Sigma_f P_{(TP/f)}$.

$$W_y = \frac{P_{(\mathrm{TP}/y)}}{\sum_f P_{(\mathrm{TP}/f)}} \tag{5}$$

The probability of identifying true positives from a certain type of data is given by equation (6), where TP/y denotes the true positives of a certain type and $\Sigma_f TP/f$ is the sum of true positives of all types.

$$P_{(\mathrm{TP}/y)} = \frac{\mathrm{TP}/y}{\sum_f \mathrm{TP}/f} \tag{6}$$

Such an approach allowed us to estimate the weights for the complete data sets from both types as well as for various levels of incomplete data for CX-MS. Therefore, using the values derived for the theoretical cross-links, we weighted the impact of our data from CX-MS experiments in the training set of complexes.

To estimate the impact of each individual experiment when incomplete data sets are available, we calculated the individual weights using various percentages of data available from each type. We estimated the weights for complete IM-MS and CX-MS data sets using equations (3) and (4), yielding the values of $W_{\mathrm{IM\text{-}MS}} = 0.361$ and $W_{\mathrm{CX\text{-}MS}} = 0.639$ (± 0.05 s.d. in both cases) for MMOH, ToMOH and urease. Thus, as protein complexes with very different shapes and stoichiometries assigned with very similar weighting scores, we are able to use this as a generic setting for our subsequent predictions of complexes with unknown high-resolution structures.

Clustering analysis. We judged the uniqueness of the candidate models by performing clustering analysis. As such, we clustered the best-scoring models into distinct subsets on the basis of their structural similarities, using a hierarchical tree approach[33]. Here we hierarchically clustered the 1% of best-scoring models according to their pairwise RMSDs and represented each identified cluster by the model with the best score.

Flexibility. In a final step, to account for flexibility we subjected the best-scoring models to dynamical analysis using NAMD[34]. Thus, we refined the atomic positions of the subunits within the subcomplexes by performing energy minimization. We performed such an analysis at all intermediate steps needed to build the assembly. This allowed us not only to eliminate any steric clashes in the final models but also to search for the most energetically favorable conformation(s).

Rigid docking on the density map. To confirm the validity of our models, we fitted the model structures assembled for all complexes and subcomplexes of the proteasomal lid into the corresponding density map[13] using the UCSF Chimera package (version 16.2)[35]. Briefly, we first manually placed the model structure into the map and then rigidly docked using the automated docking tool as implemented in UCSF Chimera. We quantitatively assessed the quality of fit of the best-scoring structures of the intact lid complex and subcomplexes to the density map using the cross-correlation coefficient.

Homology modeling. We performed homology modeling for MMOH, ToMOH and urease benchmark cases (**Supplementary Table 1**), the proteasomal lid (**Supplementary Table 10**) and the base subcomplexes (**Supplementary Table 12**) using Modeller (version 9.11). We selected the final structures upon satisfaction of spatial restraints and the discrete optimized protein energy (DOPE) assessment scores[36] as implemented in Modeller[37]. Finally, we verified the predicted structures using the Prochek validation program[38].

Software. Software documentation for the method is described in **Supplementary Note 3**, and the software is available as **Supplementary Software** and can be found at https://github.com/integrativemodeling/hybrid_ms_method/.

21. McCormick, M.S., Sazinsky, M.H., Condon, K.L. & Lippard, S.J. *J. Am. Chem. Soc.* **128**, 15108–15110 (2006).
22. Jabri, E. & Karplus, A. *Biochemistry* **35**, 10616–10626 (1996).
23. Sakata, E. *et al. Mol. Cell* **42**, 637–649 (2011).
24. Ghaemmaghami, S. *et al. Nature* **425**, 737–741 (2003).
25. Sobott, F., Hernández, H., McCammon, M.G., Tito, M.A. & Robinson, C.V. *Anal. Chem.* **74**, 1402–1407 (2002).
26. Hernández, H. & Robinson, C.V. *Nat. Protoc.* **2**, 715–726 (2007).
27. Pringle, S.D. *et al. Int. J. Mass Spectrom.* **261**, 1–12 (2007).
28. Kemper, P.R., Dupuis, N.F. & Bowers, M.T. *Int. J. Mass Spectrom.* **287**, 46–57 (2009).
29. Bush, M.F. *et al. Anal. Chem.* **82**, 9557–9565 (2010).
30. Russel, D. *et al. PLoS Biol.* **10**, e1001244 (2012).
31. Alber, F., Kim, M.F. & Sali, A. *Structure* **13**, 435–445 (2005).
32. Alber, F., Forster, F., Korkin, D., Topf, M. & Sali, A. *Annu. Rev. Biochem.* **77**, 443–477 (2008).
33. Johnson, S.C. *Psychometrika* **32**, 241–254 (1967).
34. Phillips, J.C. *et al. J. Comput. Chem.* **26**, 1781–1802 (2005).
35. Pettersen, E.F. *et al. J. Comput. Chem.* **25**, 1605–1612 (2004).
36. Shen, M.-y. & Sali, A. *Protein Sci.* **15**, 2507–2524 (2006).
37. Šali, A. & Blundell, T.L. *J. Mol. Biol.* **234**, 779–815 (1993).
38. Laskowski, R.A., MacArthur, M.W., Moss, D.S. & Thornton, J.M. *J. Appl. Crystallogr.* **26**, 283–291 (1993).