analytical chemistry

Article

Predicting the shapes of protein complexes through collision cross section measurements and database searches

Michael Landreh, Cagla Sahin, Joseph Gault, Samira Sadeghi, Chester Lee Drum, Povilas Uzdavinys, David Drew, Timothy M Allison, Matteo T. Degiacomi, and Erik G. Marklund

Anal. Chem., **Just Accepted Manuscript** • DOI: 10.1021/acs.analchem.0c01940 • Publication Date (Web): 13 Jul 2020

is published by the American Chemical Society. 1155 Sixteenth Street N.W., Washington, DC 20036

Predicting the shapes of protein complexes through collision cross section measurements and database searches

Michael Landreh[1,*], Cagla Sahin[1,2], Joseph Gault[3], Samira Sadeghi[4], Chester L. Drum[4], Povilas Uzdavinys[5,$], David Drew[5], Timothy M. Allison[6], Matteo T. Degiacomi[7,*], and Erik G. Marklund[8,*]

[1] Department of Microbiology, Tumor and Cell Biology, Karolinska Institutet, Solnavägen 9. 171 65, Stockholm, Sweden

[2] Department of Biology, University of Copenhagen, Ole Maaløes Vej 5, 2200 Copenhagen N, Denmark

[3] Department of Chemistry, University of Oxford, South Parks Road, Oxford OX1 3QZ, UK

[4] Department of Medicine, Yong Loo Lin School of Medicine, National University of Singapore, 10 Medical Dr, Singapore 119228, Singapore

[5] Department of Biochemistry and Biophysics, Stockholm University, Stockholm, Sweden.

[6] Biomolecular Interaction Centre and School of Physical and Chemical Sciences, University of Canterbury, Private Bag 4800, Christchurch 8140, New Zealand

[7] Department of Physics, Durham University, South Road, DH1 3LE, UK

[8] Department of Chemistry - BMC, Uppsala University, Box 576, 751 23, Uppsala, Sweden

[$] Present address: Department of Structural Biology, Institute of Biophysics and Physical Biochemistry, University of Regensburg, Universitätsstrasse 31, D-93053, Regensburg, Germany

* Correspondence to: michael.landreh@ki.se, matteo.t.degiacomi@durham.ac.uk, or erik.marklund@kemi.uu.se

Abstract

In structural biology, collision cross sections (CCSs) from ion mobility mass spectrometry (IM-MS) measurements are routinely compared to computationally or experimentally derived protein structures. Here, we investigate whether CCS data can inform about the shape of a protein in the absence of specific reference structures. Analysis of the proteins in the CCS database shows that protein complexes with low apparent densities are structurally more diverse than those with a high apparent density. Although assigning protein shapes purely on CCS data is not possible, we find that we can distinguish oblate- and prolate-shaped protein complexesby using the CCS, molecular weight, and oligomeric states to mine the Protein Data Bank (PDB) for potentially similar protein structures. Furthermore, comparing the CCS of a ferritin cage to the solution structures in the PDB reveals significant deviations caused by structural collapse on the gas phase. We then apply the strategy to an integral membrane protein by comparing the shapes of a prokaryotic and a eukaryotic sodium/proton antiporter homologue. We conclude that mining the PDB with IM-MS data is a time-effective way to derive low-resolution structural models.

Key words: Structural proteomics, protein architecture, native mass spectrometry, topology prediction, collision cross sections

Abbreviations: Ion mobility mass spectrometry, IM-MS; collision cross section, CCS; molecular weight, MW

Introduction

The combination of native mass spectrometry (MS) and ion mobility spectrometry (IM), in the form of IM-MS, is a versatile tool for structural biology.[1–3] In this approach, native protein complexes are transferred to the gas phase by nano-electrospray ionization, whilst retaining their non-covalent interactions and a native-like structure.[4] Measuring the mobility of these ions in an electric field inside a gas-filled drift tube allows the determination of collision cross sections (CCSs) from the observed drift times.[5]

In structural biology the CCS measurements of native protein ions are most commonly employed for two reasons: (1) to probe the structure of a protein complex in the gas phase, or (2) to generate structural constraints or restraints to inform computational modeling. The first application can require relatively detailed knowledge about the 3-dimensional organization of the protein of interest. Here, a theoretical native CCS is computed from a high-resolution structure, and then compared to an experimentally determined CCS to assess the integrity of the desolvated complex or monitor conformational changes. Deviations from the theoretical CCS can be used to follow collapse or unfolding of the complex after desolvation. The second common application elucidates the quaternary structures of protein complexes. The CCS is computed from numerous, often coarse-grained models of the assembly and compared to the experimental CCS to identify the most likely structural organization of the protein.[6,7] This strategy has been applied successfully to locate missing subunits in crystal structures,[6] or, when paired with distance restraints obtained by chemical crosslinking, to derive the architectures of complex molecular machineries.[8] The strategy has also been utilized to model complete protein assemblies, for example of polydisperse small heat-shock protein oligomers.[9,10]

Common to the major applications for CCS measurements is that they are interpreted with the help of *a priori* information about the protein of interest, such as protein complex symmetry, high-resolution structures, or the possible connectivities of the subunits in a protein complex.[2] This requirement has prompted us to ask what structural insights can be obtained directly from IM-MS analysis of a native-like protein complex and including only a minimum of other protein-specific structural information. There is a trove of general structural information about proteins, because protein structures are not random: besides the selection that have given rise to specific functions, they have all evolved under common biophysical constraints that dictates what sizes, shapes, and architectures are beneficial.[11–13] It is likely that this has shaped the structural proteome on many levels, creating patterns in how the structural space is populated, which in turn could be modulated by other properties, such as oligomeric state or subcellular location. These patterns may in part be revealed by inspecting collections of known protein structures, and structural databases linking high-resolution structures and CCS information, can consequently be used to indicate the architectures of protein complexes with unknown structures.[14]

Here, by mining the PDB using only molecular weight, CCS, and oligomeric state information, we have determined that it is possible to predict whether a protein complex adopts an oblate or a prolate shape, or a spherical architecture. This simple classification of protein shape can reduce the search space for low-resolution models without a need for reference structures or complex computations.

Experimental

Dataset. MW and CCS for 18 native-like protein ions recorded on a drift tube IM-MS instrument (Waters) in positive ionization mode with helium as drift gas were taken from the Bush CCS database (https://depts.washington.edu/bushlab/ccsdatabase/), accessed 10/2019. The CCS for all PDB entries were computed previously.[15] See Table S1 for all proteins, PDB IDs, MWs, and CCSs used here.

PDB Mining. The PDB mining was carried out as described previously.[14] Briefly, masses and helium CCS for the SAP pentamer and decamer, and the bovine lactoglobulin dimer were taken from the Bush CCS database, or determined experimentally by TWIMS for NHA2 and NapA (see below). CCS and MW were used as input to the Python script find_omega_neighbours.py, which is distributed alongside IMPACT,[15] to find the best matching protein complexes ("neighbors") in terms of CCS and mass (m) among all biological assemblies in the PDB. Default parameters were used, except for the number of neighbours in the output, which was set to 150 instead of the default value of 10. In the script, the CCS is converted to a reduced cross-section (ω), where a ω above (or below) 1.0 signifies a higher (or lower) CCS than expected for a protein of the given mass (see Reference 17), which is then used together with m to compare to the protein complexes in the underlying database. The neighbors are ranked according to their distance to the point in the (ω,m)-plane defined by the reference values given in the input. The distance metric d was defined as an Euclidian norm in the (ω,m)-plane, but since the two coordinate axes represent fundamentally different quantities, weights were introduced to define their respective contributions. Additionally, the mass-component of the distance was taken on a logarithmic scale, $d = \sqrt{\left(w_m log_{10} m/m_r\right)^2 + \left(w_\omega(\omega - \omega_r)\right)^2}$ where "r"-subscripts denote reference values, and the weights were set to $w_m = 1.0$ and $w_\omega = 5.0$. Using the advanced search option "Structure Feature: Number of chains (Biological assembly)" in the PDB, the list of PDB IDs from the Python script was then filtered according to oligomeric state. Resulting entries were checked manually and the ten best matches with the correct homo-oligomeric state according to structure annotation were included in the final list.

Data analysis. CCS and MW were used to calculate apparent densities for spherical proteins as described.[16] The biological assembly structures for all protein entries in the final match-list were downloaded from the PDB. Solvent and salt molecules were deleted from the PDB files, and the inertia tensor for each structure was computed in UCSF Chimera V1.11.2 [17] using the "measure inertia" command. The command returns an ellipsoid that has the same inertia as the structure with all atoms mass-weighted. The calculation also returns the principal axes lengths, moments, and center for the ellipsoid. The vectors v1, v2, and v3 are the principal axes (longest to shortest). The lengths a, b, c are half-diameters along axes v1, v2, and v3, and where used to calculate the principal axes ratio as (a × b) / (b × c), which returns values < 1 for oblates, 0 for perfect spheres, and > 1 for prolates.

Protein preparation. Ferritin from *Archaeoglobus fulgidus* with the F166H mutation was prepared as described.[18] NapA from *Thermus thermophilus* was expressed in *E. coli* and purified as described.[19] NHA2 from *Bison bison* (residues 69-525) was expressed in yeast and purified using the protocol described previously for human NHA2.[20,21] Prior to MS analysis, membrane proteins were exchanged into 100 mM ammonium acetate, pH 7.5, containing 0.01 % C12E9 with a Superdex Increase 200 column on an ÄKTA Purifier FPLC system (GE Healthcare) maintained at 4 °C. Ferritins were exchanged into 100 mM ammonium acetate, pH 7.5, using P-6 Bio-Spin columns (BioRad).

Mass spectrometry. Samples were introduced into the mass spectrometer using gold-coated borosilicate capillaries produced in-house. Mass spectra were recorded on a Synapt G1 T-wave IM mass spectrometer (Waters). Instrument settings were: Capillary voltage 1.5 V, cone voltage 130 V, collision voltages in the trap ranging between 80 and 200 V for NHA2 and 10 V for ferritin, and transfer collision voltage 50 V. The CCS for

NHA2 was measured at a collision voltage of 100 V. The source pressure was 9 mbar. Ion mobility settings were: wave velocity 300 m/s and wave height 13 V in the IMS cell, wave velocity 248 m/s and wave height 13 V in the transfer region. Drift cell gas was N_2 with a pressure of 1.6 Torr. CCS calibrations were performed using alcohol dehydrogenase, concanavalin A and pyruvate kinase for NHA2, and β-galactosidase (all Sigma) for ferritin.[22] The N_2 CCS values reported by Bush *et al.*[23] (alcohol dehydrogenase, concanavalin A and pyruvate kinase) or Benesch *et al.*[24] (β-galactosidase) were used for calibration. MS data were analysed using Mass Lynx 4.1, DriftScope (Waters, Milford, MA) and PULSAR software packages (http://pulsar.chem.ox.ac.uk/).[25] The calibration curves used for CCS determinaton of Ferritin and NHA2 are shown in Figure S1.

Results

Assessing protein shapes through IM-MS and PDB mining

We first asked whether CCS and MW can inform about the shape of a protein. The protein structure universe is highly diverse,[26] and detailed categorizations, such as symmetry classifications, likely fall outside of the resolving power of IM-MS. To investigate whether we can determine protein shape through IM-MS, we chose therefore the simplest possible approach by approximating protein complexes as spheroids. This is not an unreasonable assumption, as previous studies have established that the majority of soluble, ordered proteins adopt a roughly elliptical shape that might facilitate optimal diffusion in the intracellular environment.[11,12]

Multiple studies have revealed that native-like ions of small, single-domain proteins such as ubiquitin have a relatively constant density of approximately 0.8 - 1.1 g/cm^3 in the gas phase,[27] similar to the values determined for proteins in solution.[28] For uniformly shaped protein complexes, at this constant density, CCS should be tightly connected to molecular weight. This correlation thus gives an estimate of the lower CCS boundary for a spherical protein in the gas phase. If the experimentally determined CCS is higher, that is, the density of the protein *appears* to be lower, it is reasonable to assume that the protein ion deviates from the form of a densely packed sphere.[23] We can describe the possible CCS deviations as an elongation (prolate-shaped), flattening (oblate-shaped), or the presence of unoccupied space in the spheroid (hollow sphere), which can be distinguished based on the ratio of their principal axes (Figure 1A). This concept has been successfully applied to synthetic polymers in the gas phase, where deviations in shape could be delineated from a reduction in apparent density at greather chain-lenghts. Recently, these findings were shown to translate to desolvated peptides, illustrating the possibility that candidate shapes can be predicted based on CCS measurements.[29,30]

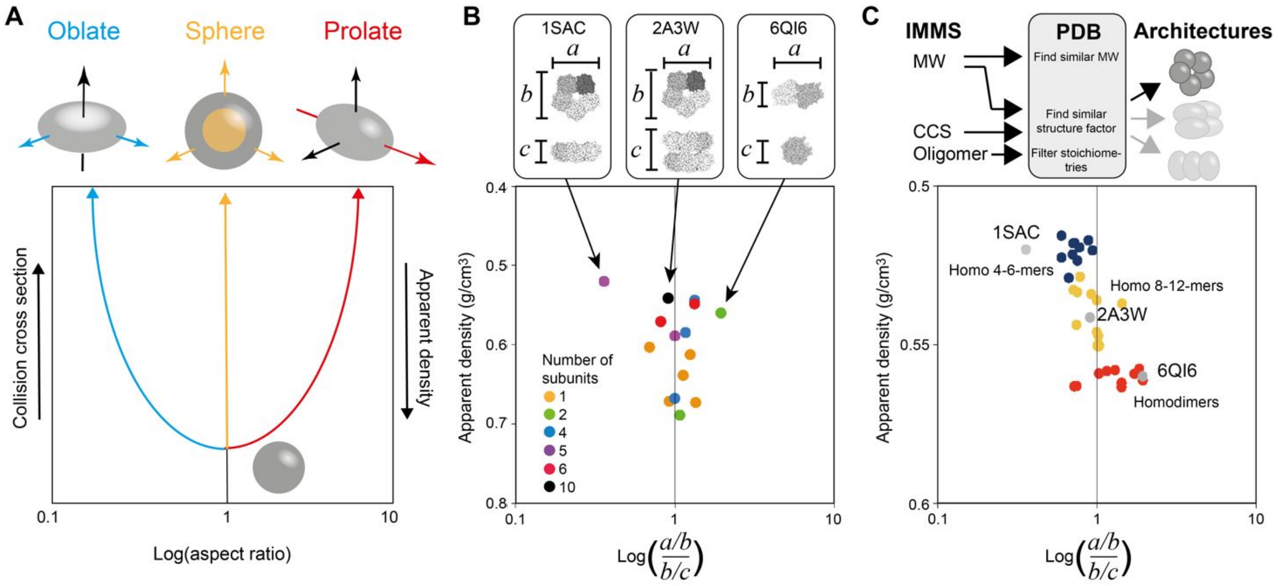

Figure 1. Predicting protein shapes based on IM-MS measurements. (A) In the absence of any structural information, proteins can be approximated as spheroids. An increase in CCS relative to the molecular weight indicates a deviation from the organization into a densely packed sphere. Proteins are coloured according to oligomeric state. (B) When applied to the Bush lab CCS database, we find vastly different densities. If we then model the protein structures in the database as spheroids, we find that the density scales with protein shape: ”dense” proteins tend to be spherical, “less dense” proteins have different principal axes ratios. (C) We selected three representative cases, the oblate serum amyloid protein (SAP) pentamer (blue), the cylindrical SAP decamer (yellow), and the prolate lactoglobulin dimer (red). Using the experimentally determined MWs, CCS values, and oligomeric states, we searched for structural neighbours in the entire PDB. We the computed the principal axes ratios for the ten best matches and found oblates for the SAP pentamer and predominantly prolates for the lactoglobulin dimer ($p = 0.04$). The SAP decamer matches contain a mixture of oblate- and prolate-leaning shapes, in line with its principal axes ratio of ~1.

To test the validity of this framework for IM-MS data of intact protein complexes, we selected 15 monomeric or homo-oligomeric protein complexes from the Bush Lab CCS database (https://depts.washington.edu/bushlab/ccsdatabase/).[23] The dataset fulfils two essential criteria. Firstly, the CCS data were recorded on a drift-tube IM-MS platform under identical conditions, minimizing the error from calibration or parameter variations. Secondly, high-resolution structures are available for all proteins, facilitating comparisons between CCS and protein shape. Hetero-oligomers were excluded in the present study due to their increased possibility for structural complexity. For all proteins in the data set, we calculated the apparent densities based on the measured MW and CCS, assuming a perfectly spherical shape. The resulting values range from 0.7 g/cm^3 for the egg white avidin dimer to 0.52 g/cm^3 for the serum amyloid P component (SAP) pentamer (Figure 1B, Table S1). Next, we calculated the inertia tensor for each associated protein structure to obtain the length of the three principal axes of a spheroid that describes the protein's shape. The ratio between the longest (a) and the intermediate axis (b), and the middle to the shortest axis (c) is <1 for an oblate, 1 for a perfect sphere, and >1 for a prolate. As expected, proteins with low apparent densities displayed a larger variation in principal axes ratios. Notably, the oblate-shaped SAP pentamer (PDB ID 1SAC), the even-sided SAP decamer (PDB ID 2A3W), and the prolate-shaped lactoglobulin dimer (PDB ID 6QI6) all displayed very similar apparent densities (Figure 1B). Similarly, we found no clear correlation between oligomeric state and shape within the dataset (Figure 1B). The observation of different shapes for protein complexes with similar apparent densities or oligomeric states confirm that protein structures are too diverse to assign any specific shape based solely on the oligomeric state, or the CCS and MW.

We then asked whether the combination of all three factors, the oligomeric state, the CCS, and the MW, can inform about the shape of protein complexes. Previously, we computed the CCS for all protein complexes in the PDB (>180 000 structures).[15] By mining the PDB for protein complexes that match an experimentally determined CCS, we were, for example, able to confirm the ring-like shape of a phycobiliprotein complex with an additional subunit.[14] These findings led us to consider the PDB as a large collection of sample structures, with the additional advantage that the structures represent predominantly physiologically relevant assemblies.

Therefore, by mining the PDB for protein complexes with similar stoichiometry, MW, and CCS, it may be possible to identify which shape(s) an unknown protein complex is likely to adopt. To test this idea, we selected the three protein complexes with similar apparent densities but different shapes (the SAP pentamer, the SAP decamer, and the lactoglobulin dimer). We searched the PDB for entries with similar MW and CCS to generate a list of the

150 best matches for apparent density for each protein, which are scored according to their distance from the search values (with a distance of 0 indicating a perfect match, see methods). From this list, we selected the ten best matches that had an oligomeric state similar to the target protein. For the lactoglobulin dimer, the top ten matching dimers were extracted. For the SAP pentamer, the 150 best matches for CCS and MW included only one pentameric complex, so tetramers and hexamers were also included. Similarly, the top 150 matches for CCS and MW of the SAP decamer included only one decamer, so octamers and dodecamers with matching MW and CCS were also considered (Table S2). Plotting the principal axes ratio of each structure against the apparent density (Figure 1C) revealed that the matches for the prolate-shaped lactoglobulin dimer are predominantly prolates, as indicated by their average principal axes ratio of 1.3 ± 0.4. The PDB entries matching the oblate-shaped SAP pentamer are mostly oblates with an average axes ratio of 0.7 ± 0.1. For the SAP decamer with an axes ratio of close to 1, we found matches with axes ratios of 0.8 ± 0.2, which included both oblates and prolates, as well as three complexes with a spherical shape (Figure 1C). Thus the CCS, oligomeric state, and MW of the lactoglobulin dimer match predominantly with prolate-shaped complexes, and those of the SAP pentamer with oblate-shaped complexes. The CCS, oligomeric state, and MW of the SAP decamer, on the other hand, do not match with predominantly prolate- or oblate-shaped proteins, in line with its even axis lengths. In summary, the shape distribution of the protein complexes identified by mining the PDB with CCS, MW, and oligomeric state indicates the likely shape of the target complex.

Comparing gas-phase and solution structures through PDB mining

It is important to note that we effectively compare gas-phase structures to solution structures when mining the PDB using experimental CCS data. For the proteins in the CCS database, the experimental CCSs generally agree with the crystal structures.[23] However, significant deviations between theoretical and experimental CCS have been observed in some proteins, and are commonly caused by the collapse of unsupported or disordered structures.[16,31–33] Interestingly, a recent report outlined how a combination of capillary electrophoresis and native MS can provide comparable insights into protein shapes. The approach is likely similarly sensitive to desolvation-related structural changes.[34] We, therefore, considered how gas-phase changes in protein structures affect the ability to predict the shape of a protein complex. As a test case, we selected the ferritin from *Archaeoglobus fulgidus*, a homo-24-mer that forms a hollow sphere with four large pores and which partially collapses in the gas phase.[35]

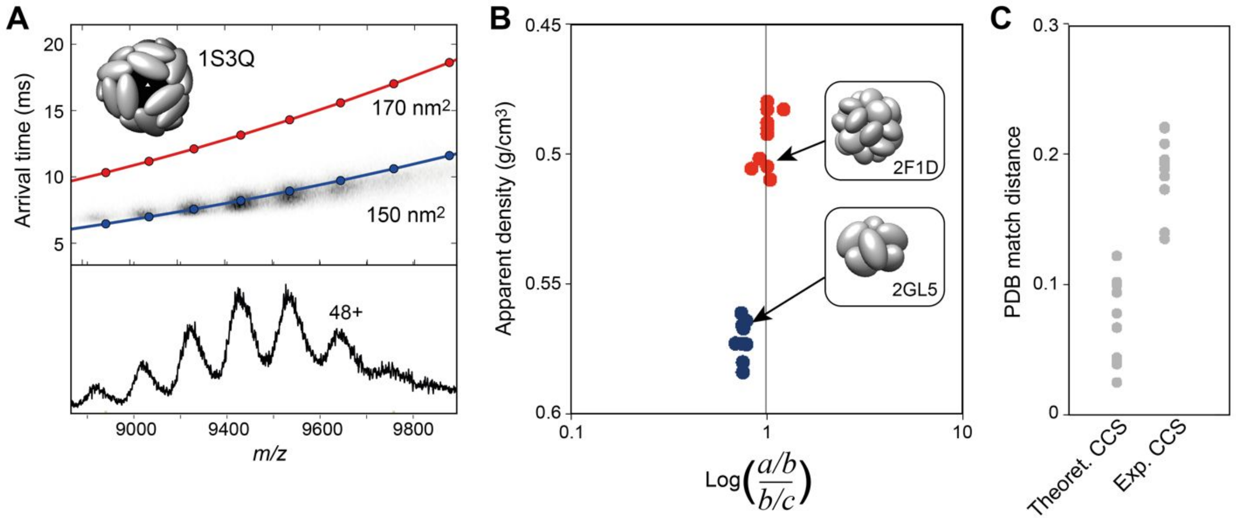

Figure 2. Effect of gas-phase collapse on shape prediction. (A) IM-MS analysis of the intact Archaeoglobus ferritin shows the average experimental CCS of 150 nm^2 (blue line), 15% below the theoretical CCS of 170 nm^2 (red line) computed from the crystal structure (PDB ID 1S3Q, insert). The mobiligram is shown at the top and the mass spectrum at the bottom. (B) Plotting the principal axes ratios and apparent densities of the ten best PDB matches reveals that the theoretical CCS (red) is associated with spherical complexes, while the experimental CCS (blue) returns predominantly oblate-shaped matches. The average densitiy and the standard deviation between the apparent densities for all Ferritin charge states are shown as dashed line shaded area, respectively. (C) The match distance informs about the agreement between the best matches in the PDB and the target CCS and MW. For collapsed ferritin, the distance is significantly larger for the ten best matches identified using the experimental CCS compared to the ten best matches for the theoretical CCS.

We performed IM-MS measurements and found that the intact 24-mer has a CCS of 150 nm^2, approximately 15% below the theoretical value (Figure 2A), suggesting significant compaction of the protein complex in the gas phase compared to the crystal structure. We then mined the PDB using the MW and either the experimental or the theoretical CCS. The oligomeric state was not considered due to the low number of homo-24-mers in the PDB. As expected, the ten best matches for the theoretical CCS show mostly spherical assemblies and include two ferritins. The matches using the experimentally determined

CCS, however, returns mostly oblate-shaped complexes (Figure 2B, Table S3). We conclude that the shift in CCS from native to collapsed structure also shifts the distribution of matching protein shapes. This implies that the use of CCS data to identify similarly shaped proteins in the PDB requires that the protein of interest does not undergo major re-arrangement in the gas phase.

This finding prompted us to ask whether the PDB matches for the collapsed ferritin can provide information about its shape in the gas phase. We analysed the distances between experimental CCS and MW, and the CCS and MW of the best-matching PDB entries. A small distance indicates that a PDB entry closely agrees with the target CCS and MW values, while larger distances indicate poorly representative structures. The ten best matches for the CCS of the collapsed protein complex have an average distance of 0.186. PDB entries matching the same MW and the theoretical CCS have an average distance of 0.071. The different distances reveal that the structures in the PDB are more closely related to the intact ferritin structure than to the collapsed state that has the same MW but a lower CCS. In this context, large distances suggest that the PDB does not contain similarly shaped proteins, and therefore may indicate non-physiological structures such as collapsed or unfolded proteins.

Application to a membrane protein dimer with unknown structure

Having established that IM-MS and PDB mining can provide information about the shapes of proteins, we tested its ability to generate structural constraints for a protein complex with unknown structure. The development of model structures is particularly important for protein systems where structure determination is challenging, such as integral membrane proteins. These proteins are significantly under-represented in the PDB,[36] creating a demand for computational models based on homology information and experimental constraints. We selected sodium-proton antiporters (NHAs), a family of integral membrane proteins responsible for maintaining the intracellular pH, and promising drug targets for hypertension.[37] To date, only structures of prokaryotic Na/H antiporters have been determined, revealing dimers with one core ion-transporting domain per protomer.[38–40] While the transport domains are relatively conserved across all phyla, structures of bacterial homologues show that their dimer interfaces differ significantly through differences in the N-terminus.[19,40–42] In particular, in the Na+/H+ antiporter NhaA from *E. coli* the homodimer is held together by two small N-terminal β-hairpins burying a total surface area of 700 Å^2.[43] The weak dimerization interface has evolved to require the lipid cardiolipin to stabilise the homodimer with functional consequences.[43–45] In contrast, the Na+/H+ antiporters NapA, PaNhaP, MjNhaP lack the β-hairpin extensions and instead dimerise through an additional helix at their N-terminus burying a larger total surface area of > 1700 Å^2.[19,41,42] Indeed, eukaryotic homologues also show longer N-terminal regions than NhaA,[46] suggesting that they may also dimerize in a similar manner.

The lack of reference structures for the N-terminal segment of NHA2 has so far precluded homology-based modelling of its structure. In order to elucidate the overall architecture of the NHA2 dimer, we therefore used IM-MS to compare a mammalian NHA2 to NapA from *Thermus thermophilus* which was previously characterized by X-ray crystallography and IM-MS.[19,20] Native MS analysis of NHA2 in the detergent $C_{12}E_9$ shows that the protein is released from detergent micelles as a stable dimer, requiring high collision voltages to dissociate (Figure 3A). The CCSs of the main charge states 16+, 17+, and 18+ were 585 nm^2, 600 nm^2, and 620 nm^2, respectively. The NHA2 dimer has a molecular weight of 103 kDa, which means that the CCS of NHA2 cannot be directly compared to that of the 82 kDa NapA dimer. Instead, we mined the PDB for homodimers with MW and CCS similar to NapA or NHA2 and computed the principal axes ratios (Figure 3B, Table S4). The best matches for the NapA dimer with 82 kDa and an average CCS of 460 nm^2 were all found to be prolates with an average principal axes ratio of 1.2 ± 0.1. This ratio is lower than the axes ratio of the NapA crystal structure, but in good agreement with the computational model of the protein in the gas-phase (Figure 3B, insert).[20] The ten best matches for the NHA2 dimer with a molecular weight of 103 kDa and an average CCS of 600 nm^2 were also exclusively prolates, with a slightly larger average principal axes ratio of 1.4 ± 0.2. Thus, our analysis indicates that NHA2 has a prolate shape and suggest that NHA2 and NapA have a similar dimer architecture despite further differences in the N-terminal sequences as compared to NhaA and one another (Figure 3C). Furthermore, we note that none of the matching structures we identified are membrane proteins. This is not surprising given the low number of membrane protein structures in the PDB. In the context of IM-MS, it will be interesting to explore whether membrane proteins exhibit similar shape distributions as their globular counterparts.

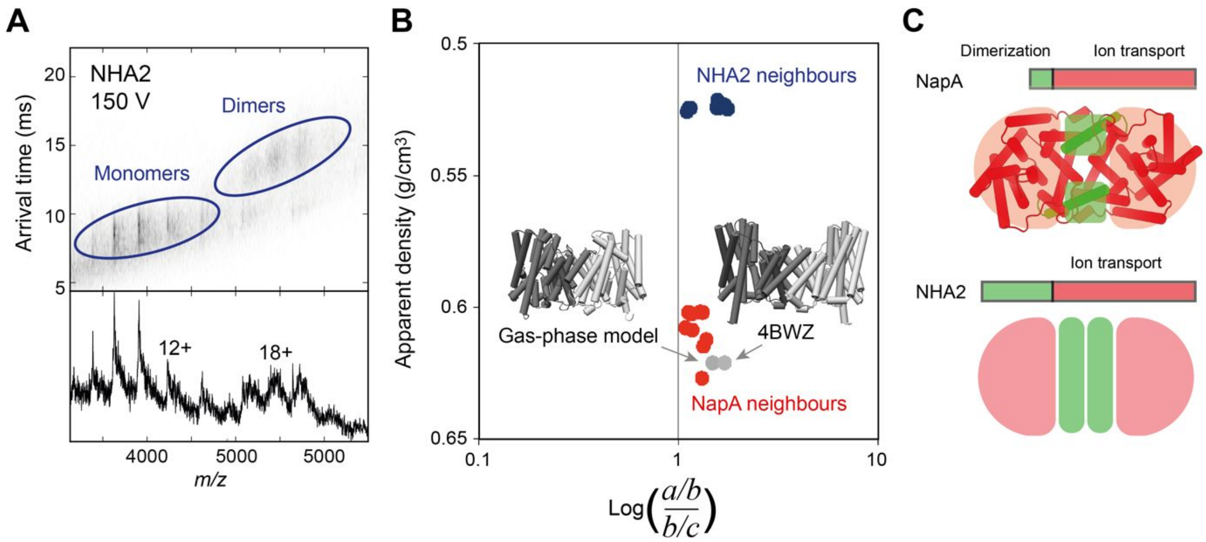

Figure 3. Comparing the shapes of integral membrane proteins NapA and NHA2 by IM-MS and PDB mining. (A) The IM-MS spectrum of NHA2 at a collision voltage of 150V shows that NHA2 is a 103 kDa dimer that can be dissociated by collisional activation. (B) The ten best matches obtained by mining the PDB with the CCS, MW, and oligomeric state of NapA (red) or NHA2 (blue) are prolate-shaped. Insert: The crystal structure of NapA (right) has a higher principal axes ratio than the PDB matches, while that of the gas-phase model structure (left) is in good agreement. The average densitiy and the standard deviation between the apparent densities for all NHA2 charge states are shown as dashed line shaded area, respectively. (C) The dimer interface in NapA (top view) is composed of a single N-terminal helix on each protomer (green) which connects the ion transport domains (red). The sequence of NHA2 has a ~100 residue N-terminal extension. Based on the average principal axes ratio of the PDB matches, we conclude that both proteins share an elongated dimer architecture in which the N-terminal segment of NHA2 likely forms a dimerization interface.

Discussion

Here we demonstrate that it is possible to approximate the shape of a protein complex by mining the structures in the PDB with constraints derived from IM-MS experiments. However, it is important to note that these shapes are not measured but predicted. We find that two main factors have to be taken into consideration:

(1) Although the PDB contains over 180 000 structures, they are not evenly distributed across the MW range (https://www.rcsb.org/stats/distribution_molecular-weight-structure). Therefore, the number of structures that can be mined differs for each protein complex, as

exemplified in the comparison between the 125 kDa SAP pentamer and the 37 kDa lactoglobulin dimer (Figure 1C). There are around 20 000 PDB entries for protein complexes between 30 and 40 kDa, but only 3 000 entries with MW between 120 and 130 kDa. The difference means that there are far more proteins in the lower mass range that potentially match the target values than there are in the higher mass range. This factor likely affects the ability of our strategy to compare proteins with vastly different MW.

(2) Some proteins are significantly over-represented in the PDB and thus cause considerable bias in the mining results.[36] For example, the 150 best matches for the SAP decamer contain 19 highly homologous, dodecameric DNA-binding proteins (Dps) with near-identical CCS and molecular weight. The uneven distribution of unique structures can skew the mining results in favour of the most abundant architectures. More generally, the structural diversity among proteins with similar oligomeric states and CCS likely affects the reliability of the PDB mining results.

We find that the PDB mining strategy presented here yields plausible results even when differences in structure distribution and diversity are not taken into consideration. Going forward, we speculate that both of these factors can be addressed by matching the PDB search space to include a representative set of sample structures. This could be achieved by analyzing the protein shape universe in the context of CCS, which can potentially broaden the applicability of IM-MS as a tool for structural modelling or topology prediction. In addition to providing information about protein complexes where no homologous structures are available, the strategy could help to elucidate the shapes of protein complexes that can only be observed by MS, such as self-assembly intermediates.[47]

Conclusions

We have demonstrated that IM-MS measurements can provide sufficient information to provide low-resolution structural insights into protein complexes without a need for specific reference structures. Using the combination of CCS, MW, and oligomeric state to mine the PDB, we are able to predict the possible shape(s) of intact protein complexes based on IMMS measurements. The strategy can facilitate comparisons of homologous proteins where absolute CCS comparisons are not feasible. For example, it can enable in-depth studies of interactions in individual protein systems, or enable the indentfication of proteins with specific conformations in complex mixtures. However, potential gas phase collapse and PDB bias have to be taken into consideration.

Supporting Information.

Table S1: CCS, MW, and intertia axes of homomeric proteins in the CCS database

Table S2: CCS, MW, and intertia axes of PDB matches

Table S3: CCS, MW, and intertia axes of Ferritin matches)

Table S4: CCS, MW, and intertia axes of NHA2 and NapA matches)

Figure S1: Calibrants with charge states and cross-sections (in nm^2) and PULSAR-generated CCS calibration curves for IMMS analysis of ferritin and NHA2.

Acknowledgements

This work was supported by the Uppsala-Durham Seedcorn Fund to EGM, MTD, and ML. CS is supported by a Novo Nordisk Foundation Postdoctoral Fellowship (NNF19OC0055700). ML gratefully acknowledges technical support from MS Vision, NL. ML is supported by an Ingvar Carlsson Award from the Swedish Foundation for Strategic Research (SSF), a KI faculty-funded Career Position, a KI-StratNeuro starting grant, a Starting Grant from the Swedish Research Council (VR), and a Cancerfonden Project Grant. J Gault is supported by a Junior Research Fellowship at The Queen's College, Oxford. The authors would like thank Prof. Justin Benesch, University of Oxford, for encouragement and helpful discussions, and to Prof. Sir David P. Lane, Karolinska Institutet, for support through Swedish Research Council Grant 2013_08807.

References

(1) Allison, T. M.; Landreh, M. Ion Mobility in Structural Biology. *Compr. Anal. Chem.* **2019**, *83*, 161–195.

(2) Zhong, Y.; Hyung, S. J.; Ruotolo, B. T. Ion Mobility-Mass Spectrometry for Structural Proteomics. *Expert Rev. Proteomics* **2012**, *9* (1), 47–58.

(3) Uetrecht, C.; Rose, R. J.; Van Duijn, E.; Lorenzen, K.; Heck, A. J. R. Ion Mobility Mass Spectrometry of Proteins and Protein Assemblies. *Chem. Soc. Rev.* **2010**, *39* (5), 1633–1655.

(4) Breuker, K.; McLafferty, F. W. Stepwise Evolution of Protein Native Structure with Electrospray into the Gas Phase, 10-12 to 102 S. *Proc. Natl. Acad. Sci.* **2008**, *105* (47), 18145–18152.

(5) Pukala, T. Importance of Collision Cross Section Measurements by Ion Mobility Mass Spectrometry in Structural Biology. *Rapid Commun. Mass Spectrom.* **2019**, *33* (S3), 72–82.

(6) Pukala, T. L.; Ruotolo, B. T.; Zhou, M.; Politis, A.; Stefanescu, R.; Leary, J. A.; Robinson, C. V. Subunit Architecture of Multiprotein Assemblies Determined Using Restraints from Gas-Phase Measurements. *Structure* **2009**, *17* (9), 1235–1243.

(7) Politis, A.; Park, A. Y.; Hyung, S. J.; Barsky, D.; Ruotolo, B. T.; Robinson, C. V. Integrating Ion Mobility Mass Spectrometry with Molecular Modelling to Determine the Architecture of Multiprotein Complexes. *PLoS One* **2010**, *5* (8), e12080.

(8) Politis, A.; Stengel, F.; Hall, Z.; Hernández, H.; Leitner, A.; Walzthoeni, T.; Robinson, C. V.; Aebersold, R. A Mass Spectrometry-Based Hybrid Method for Structural Modeling of Protein Complexes. *Nat. Methods* **2014**, *11* (4), 403–406.

(9) Santhanagopalan, I.; Degiacomi, M. T.; Shepherd, D. A.; Hochberg, G. K. A.; Benesch, J. L. P.; Vierling, E. It Takes a Dimer to Tango: Oligomeric Small Heat Shock Proteins Dissociate to Capture Substrate. *J. Biol. Chem.* **2018**, *293* (51), 19511–19521.

(10) Baldwin, A. J.; Lioe, H.; Hilton, G. R.; Baker, L. A.; Rubinstein, J. L.; Kay, L. E.; Benesch, J. L. P. The Polydispersity of Ab-Crystallin Is Rationalized by an Interconverting Polyhedral Architecture. *Structure* **2011**, *19* (12), 1855–1863.

(11) Shannon, G.; Marples, C. R.; Toofanny, R. D.; Williams, P. M. Evolutionary Drivers of Protein Shape. *Sci. Rep.* **2019**, *9* (1), 11873.

(12) Dima, R. I.; Thirumalai, D. Asymmetry in the Shapes of Folded and Denatured States of Proteins. *J. Phys. Chem. B* **2004**, *108* (21), 6564–6570.

(13) Dill, K. A.; Ghosh, K.; Schmit, J. D. Physical Limits of Cells and Proteomes. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108* (44), 17876–17882.

(14) Kaldmäe, M.; Sahin, C.; Saluri, M.; Marklund, E. G.; Landreh, M. A Strategy for the Identification of Protein Architectures Directly from Ion Mobility Mass Spectrometry Data Reveals Stabilizing Subunit Interactions in Light Harvesting Complexes. *Protein Sci.* **2019**, *28* (6), 1024–1030.

(15) Marklund, E. G.; Degiacomi, M. T.; Robinson, C. V.; Baldwin, A. J.; Benesch, J. L. P. Collision Cross Sections for Structural Proteomics. *Structure* **2015**, *23* (4), 791–799.

(16) Pagel, K.; Natan, E.; Hall, Z.; Fersht, A. R.; Robinson, C. V. Intrinsically Disordered P53 and Its Complexes Populate Compact Conformations in the Gas Phase. *Angew. Chemie - Int. Ed.* **2013**, *52* (1), 361–365.

(17) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera—A Visualization System for Exploratory Research and Analysis. *J Comput Chem* **2004**, *25*, 1605–1612.

(18) Deshpande, S.; Masurkar, N. D.; Girish, V. M.; Desai, M.; Chakraborty, G.; Chan, J. M.; Drum, C. L. Thermostable Exoshells Fold and Stabilize Recombinant Proteins. *Nat. Commun.* **2017**, *8* (1), 1442.

(19) Lee, C.; Kang, H. J.; Von Ballmoos, C.; Newstead, S.; Uzdavinys, P.; Dotson, D. L.; Iwata, S.; Beckstein, O.; Cameron, A. D.; Drew, D. A Two-Domain Elevator Mechanism for Sodium/Proton Antiport. *Nature* **2013**, *501* (7468), 573–577.

(20) Landreh, M.; Marklund, E. G.; Uzdavinys, P.; Degiacomi, M. T.; Coincon, M.; Gault, J.; Gupta, K.; Liko, I.; Benesch, J. L. P.; Drew, D.; et al. Integrating Mass Spectrometry with MD Simulations Reveals the Role of Lipids in Na+/H+ Antiporters. *Nat. Commun.* **2017**, *8*, 13993.

(21) Drew, D.; Newstead, S.; Sonoda, Y.; Kim, H.; von Heijne, G.; Iwata, S. GFP-Based Optimization Scheme for the Overexpression and Purification of Eukaryotic Membrane Proteins in Saccharomyces Cerevisiae. *Nat. Protoc.* **2008**, *3* (5), 784–798.

(22) Allison, T. M.; Landreh, M.; Benesch, J. L. P.; Robinson, C. V. Low Charge and Reduced Mobility of Membrane Protein Complexes Has Implications for Calibration of Collision Cross Section Measurements. *Anal. Chem.* **2016**, *88* (11), 5879–5884.

(23) Bush, M. F.; Hall, Z.; Giles, K.; Hoyes, J.; Robinson, C. V.; Ruotolo, B. T. Collision Cross Sections of Proteins and Their Complexes: A Calibration Framework and Database for Gas-Phase Structural Biology. *Anal. Chem.* **2010**, *82* (22), 9557–9565.

(24) Benesch, J. L. P.; Ruotolo, B. T. Mass Spectrometry: Come of Age for Structural and Dynamical Biology. *Current Opinion in Structural Biology*. 2011, pp 641–649.

(25) Allison, T. M.; Reading, E.; Liko, I.; Baldwin, A. J.; Laganowsky, A.; Robinson, C. V. Quantifying the Stabilizing Effects of Protein–Ligand Interactions in the Gas Phase. *Nat. Commun.* **2015**, *6*, 8551.

(26) Han, X.; Sit, A.; Christoffer, C.; Chen, S.; Kihara, D. A Global Map of the Protein Shape Universe. *PLoS Comput. Biol.* **2019**, *15* (4), :e1006969.

(27) Maißer, A.; Premnath, V.; Ghosh, A.; Nguyen, T. A.; Attoui, M.; Hogan, C. J. Determination of Gas Phase Protein Ion Densities via Ion Mobility Analysis with Charge Reduction. *Phys. Chem. Chem. Phys.* **2011**, *13* (48), 21630–21641.

(28) Ashkarran, A. A.; Suslick, K. S.; Mahmoudi, M. Magnetically Levitated Plasma Proteins. *Anal. Chem.* **2020**, *92* (2), 1663–1668.

(29) Haler, J. R. N.; Massonnet, P.; Far, J.; Upert, G.; Gilles, N.; Mourier, G.; Quinton, L.; De Pauw, E. Can IM-MS Collision Cross Sections of Biomolecules Be Rationalized Using Collision Cross-Section Trends of Polydisperse Synthetic Homopolymers? *J. Am. Soc. Mass Spectrom.* **2020**, *31* (4), 990–995.

(30) Haler, J. R. N.; Morsa, D.; Lecomte, P.; Jérôme, C.; Far, J.; De Pauw, E. Predicting Ion Mobility-Mass Spectrometry Trends of Polymers Using the Concept of Apparent Densities. *Methods* **2018**, *144*, 125–133.

(31) Hall, Z.; Politis, A.; Bush, M. F.; Smith, L. J.; Robinson, C. V. Charge-State Dependent Compaction and Dissociation of Protein Complexes: Insights from Ion Mobility and Molecular Dynamics. *J. Am. Chem. Soc.* **2012**, *134* (7), 3429–3438.

(32) Devine, P. W. A.; Fisher, H. C.; Calabrese, A. N.; Whelan, F.; Higazi, D. R.; Potts, J. R.; Lowe, D. C.; Radford, S. E.; Ashcroft, A. E. Investigating the Structural Compaction of Biomolecules Upon Transition to the Gas-Phase Using ESI-TWIMS-MS. *J. Am. Soc. Mass Spectrom.* **2017**, *28* (9), 1855–1862.

(33) Rolland, A. D.; Prell, J. S. Computational Insights into Compaction of Gas-Phase Protein and Protein Complex Ions in Native Ion Mobility-Mass Spectrometry. *TrAC - Trends Anal. Chem.* **2019**, *116*, 282–291.

(34) Wu, H.; Zhang, R.; Zhanga, W.; Honga, J.; Xiang, Y.; Xu, W. Rapid 3-Dimensional Shape Determination of Globular Proteins by Mobility Capillary Electrophoresis and Native Mass Spectrometry. *Chem. Sci.* **2020**, *DOI: 10.10*.

(35) Kaddis, C. S.; Lomeli, S. H.; Yin, S.; Berhane, B.; Apostol, M. I.; Kickhoefer, V. A.; Rome, L. H.; Loo, J. A. Sizing Large Proteins and Protein Complexes by Electrospray Ionization Mass Spectrometry and Ion Mobility. *J. Am. Soc. Mass Spectrom.* **2007**, *18* (7), 1206–1216.

(36) Peng, K.; Obradovic, Z.; Vucetic, S. Exploring Bias in the Protein Data Bank Using Contrast Classifiers. *Pac. Symp. Biocomput.* **2004**, 435–446.

(37) Padan, E.; Landau, M. Sodium-Proton (Na+/H+) Antiporters: Properties and Roles in Health and Disease. *Met. Ions Life Sci.* **2016**, *16*, 391–458.

(38) Drew, D.; Boudker, O. Shared Molecular Mechanisms of Membrane Transporters. *Annu. Rev. Biochem.* **2016**, *85* (1), 543–572.

(39) Hunte, C.; Screpanti, E.; Venturi, M.; Rimon, A.; Padan, E.; Michel, H. Structure of a Na+/H+ Antiporter and Insights into Mechanism of Action and Regulation by PH. *Nature* **2005**, *435* (7046), 1197–1202.

(40) Lee, C.; Yashiro, S.; Dotson, D. L.; Uzdavinys, P.; Iwata, S.; Sansom, M. S. P.; von Ballmoos, C.; Beckstein, O.; Drew, D.; Cameron, A. D. Crystal Structure of the Sodium-Proton Antiporter NhaA Dimer and New Mechanistic Insights. *J. Gen. Physiol.* **2014**, *144* (6), 529–544.

(41) Wöhlert, D.; Kühlbrandt, W.; Yildiz, O. Structure and Substrate Ion Binding in the Sodium/Proton Antiporter PaNhaP. *Elife* **2014**, *3*, e03579.

(42) Paulino, C.; Wöhlert, D.; Kapotova, E.; Yildiz, Ö.; Kühlbrandt, W. Structure and Transport Mechanism of the Sodium/Proton Antiporter MjNhaP1. *Elife* **2014**, *3*, e03583.

(43) Gupta, K.; Donlan, J. A.; Hopper, J. T.; Uzdavinys, P.; Landreh, M.; Struwe, W. B.; Drew, D.; Baldwin, A. J.; Stansfeld, P. J.; Robinson, C. V. The Role of Interfacial

Lipids in Stabilizing Membrane Protein Oligomers. *Nature* **2017**, *541* (7637), 421–424.

(44) Appel, M.; Hizlan, D.; Vinothkumar, K. R.; Ziegler, C.; Kühlbrandt, W. Conformations of NhaA, the Na/H Exchanger from Escherichia Coli, in the PH-Activated and Ion-Translocating States. *J. Mol. Biol.* **2009**, *386* (2), 351–365.

(45) Rimon, A.; Mondal, R.; Friedler, A.; Padan, E. Cardiolipin Is an Optimal Phospholipid for the Assembly, Stability, and Proper Functionality of the Dimeric Form of NhaA Na+/H+ Antiporter. *Sci. Rep.* **2019**, *9* (1), 17662.

(46) Brett, C. L.; Donowitz, M.; Rao, R. Evolutionary Origins of Eukaryotic Sodium/Proton Exchangers. *Am. J. Physiol. - Cell Physiol.* **2005**, *288*, C223-39.

(47) Landreh, M.; Andersson, M.; Marklund, E. G.; Jia, Q.; Meng, Q.; Johansson, J.; Robinson, C. V.; Rising, A. Mass Spectrometry Captures Structural Intermediates in Protein Fiber Self-Assembly. *Chem. Commun.* **2017**, *53* (23), 3319–3322.

For Table of Contents Only