

# Collision cross-sections for structural proteomics

Erik G. Marklund, Matteo T. Degiacomi,

Carol V. Robinson, Andrew J. Baldwin\*, Justin L.P. Benesch\*

*Department of Chemistry, Physical & Theoretical Chemistry Laboratory, University of  
Oxford, South Parks Road, Oxford, Oxfordshire, OX1 3QZ, U.K.*

\*Correspondence to: [justin.benesch@chem.ox.ac.uk](mailto:justin.benesch@chem.ox.ac.uk), [andrew.baldwin@chem.ox.ac.uk](mailto:andrew.baldwin@chem.ox.ac.uk)

**Keywords:** Mass spectrometry, ion mobility spectrometry, collision cross-sections, structural proteomics, structural biology.

**Running title:** CCSs for structural proteomics

## **SUMMARY**

Ion mobility mass spectrometry (IM-MS) allows the structural interrogation of biomolecules by reporting their collision cross-sections (CCSs). The major bottleneck for exploiting IM-MS for structural proteomics lies in the lack of speed at which structures and models can be related to experimental data. Here we present IMPACT (Ion Mobility Projection Approximation Calculation Tool), which overcomes these twin challenges, providing accurate CCSs up to  $10^6$  times faster than alternative methods. This allows us to assess the CCS space presented by the entire structural proteome, interrogate ensembles of protein conformers, and monitor molecular dynamics trajectories. Our data demonstrate that the CCS is a highly informative parameter, and that IM-MS is of considerable practical value to structural biologists.

## INTRODUCTION

The function of proteins, and the multi-component complexes they assemble into, is directly related to the structures they adopt and the motions that facilitate their inter-conversion (Robinson et al., 2007; Russel et al., 2009). The twin fields of structural biology and structural genomics have met with considerable success over the last two decades, however many significant structures remain elusive and the conformational heterogeneity important for function remains challenging to access experimentally (Ward et al., 2013). Ion-mobility mass-spectrometry (IM-MS) provides a novel means to complement and integrate with existing methodologies, providing information on the stoichiometry and physical size of protein assemblies, and distribution of conformations they adopt (Konijnenberg et al., 2013; Sharon, 2013; Stengel et al., 2012; Thalassinos et al., 2013; Zhou and Robinson, 2014).

IM-MS reports the collision cross-section (CCS) of ions through measuring the time taken for them to traverse a region of inert gas under the influence of a weak electric field (Bohrer et al., 2008; Ruotolo et al., 2008). For globular protein assemblies at least, there is an excellent correlation, with an error <3%, between the CCSs of protein assemblies measured experimentally using travelling-wave or drift-tube IM-MS instruments, and those estimated from their high-resolution atomic coordinates (Benesch and Ruotolo, 2011). While this observation motivates the use of IM-MS for interrogating the structure of protein assemblies, a significant bottleneck in the analysis process is imposed by the challenges involved in calculating reliable CCS values from atomic coordinates. For IM-MS to have a significant impact on integrative structural biology calculations, it is necessary to be able to routinely determine CCS values of  $>10^6$  models, with molecular masses typically in the >100 kDa range (Baldwin et al., 2011;

Laganowsky et al., 2014). Unfortunately, current computational approaches are not adequate for this task, drastically limiting the scope of IM-MS.

A number of CCS calculation algorithms are available, each developed for particular applications and molecular size-ranges (Jurneczko and Barran, 2011; Uetrecht et al., 2010). In the main these algorithms are Monte Carlo integrations in which “probes”, representing the IM gas, are “fired” upon the randomly oriented “target”, the structure under investigation. The various algorithms differ in the assumptions and approximations made in modelling the collisions between probe and target. The most complex are the trajectory method (TJM) (Mesleh et al., 1996) and the diffuse trajectory method (DTM) (Larriba and Hogan Jr, 2013), which take into account long-range interactions through, for example, Lennard-Jones potentials, to approximate the momentum transferred from each incident probe to the target. Though some deviation from experimental values is observed (Hewitt et al., 2014; Jurneczko and Barran, 2011), these methods are typically considered to provide the best CCS estimates (Bleiholder et al., 2011). However, both the TJM and DTM require the integration of forces in order to calculate the probes’ trajectories, making them very computationally expensive. The introduction of various additional approximations has led to the exact hard-sphere scattering (EHSS) (Shvartsburg and Jarrold, 1996), and projected superposition approximation (PSA) (Bleiholder et al., 2011) methods, both of which are significantly faster than the TJM and DTM.

The projection approximation (PA) is the simplest approach of all, equating the CCS to the average projected area of the target (Mack, 1925), taking into account the size of the IM gas (Fig. 1A)(Vonhelden et al., 1993). Because the PA ignores scattering and long-range interactions, the calculation is fast but also leads to a systematic underestimation

of the CCS (Bleiholder et al., 2011; Jurneczko and Barran, 2011; Larriba and Hogan Jr, 2013). However, for macromolecules, comparison between the PA and TJM reveal an excellent correlation, with the fitted values matching the TJM data to within <2% (Bleiholder et al., 2011). These observations demonstrate that the approximations made by the PA approach are not a significant drawback when examining larger targets such as proteins and their assemblies, allowing us to exploit its simplicity for application to structural proteomics.

Here we present IMPACT (Ion Mobility Projection Approximation Calculation Tool), an algorithm that has been written specifically with structural proteomics applications in mind. To accommodate the varying experiments that contribute to modern “hybrid” structural biology investigations, IMPACT is able to include models derived from X-ray crystallography, nuclear magnetic resonance spectroscopy (NMR), small angle X-ray scattering (SAXS), and electron microscopy (EM). We show that IMPACT is able to calculate CCSs orders of magnitude more rapidly than previously possible, and that this is achieved without significant drop in accuracy. We demonstrate this dramatic enhancement in performance by calculating the CCS of all the proteins in the Protein Data Bank in Europe (PDBe) in just a few hours on a single processor. The results enable us to define a shape factor that reports on how similar a given structure is to others of similar mass, a measurement that will be useful in the characterisation of unknown protein assemblies, and as a restraint in hybrid structure modelling. Furthermore, we calculate the CCS distribution of a conformational ensemble of ubiquitin, and show that its distribution of values is in close agreement with that measured experimentally. Finally, we show IMPACT is sufficiently fast for on-the-fly calculation of CCSs during molecular dynamics (MD) simulations, providing information complementary to the

radius of gyration, revealing IMPACT to be a useful addition to the molecular-modelling toolbox. In sum, our work will enable the application of IM-MS across the breadth of structural biology and structural proteomics.

## RESULTS

We formulated four principal requirements for a CCS calculator tailored for structural biology: 1) be able to calculate CCS accurately and to a well-defined precision; 2) have the capacity to calculate the CCS of very large biomolecular assemblies; 3) allow the processing of a variety of structural data and models; and 4) do all of this rapidly, enabling routine analysis of large sets of structures. Our solution, IMPACT, is a CCS calculator written from scratch in the C programming language, which can function both as a command-line tool and as a library to facilitate integration with other computational structural biology tools.

### **High accuracy collision cross-section of protein assemblies**

As CCS calculations are typically based on a Monte-Carlo integration using random numbers, run-to-run variability is expected, the magnitude of which we define as ‘precision’. To determine the statistical certainty of the result, it is crucial to monitor the convergence of the calculations. We accomplish this in IMPACT by interleaving multiple instances of the calculation (Williams et al., 2009), with a robust convergence criterion based on the standard error of their mean (Supplemental Information). We benchmarked IMPACT’s accuracy by comparing it to the TJM, the method typically considered the “gold-standard” for CCS calculation, for a reference database of 442 native protein structures spanning a range from 348 to 33,000 Å<sup>2</sup> (Fig. 1B and Supplemental Information). We find that the CCSs calculated using the two methods are highly correlated across the entire range, and show that using the power law  $\Omega_{TJM} = 0.843 \cdot \Omega_{IMPACT}^{1.05}$  to calibrate IMPACT results in a root mean square relative

error of ~1% between the two sets of values (Fig. 1B, inset; Supplemental Table S1). Correcting for the finite precision in the calculations (Supplemental Experimental Procedures), we find that the remaining error, which defines the accuracy, is 0.95%. We have therefore selected 1% as the default convergence level in IMPACT. Importantly, since the calculation error is considerably smaller than the 3% precision by which CCS values can be compared to solution structures (Benesch and Ruotolo, 2011), IMPACT is equivalently accurate for structural biology applications as the TJM.

### **IMPACT calculates collision cross-sections with unprecedented speed**

To achieve high computational performance we implemented two strategies. First, we separated the calculation of rotation matrices from their application to the atomic coordinates (Supplemental Information) (Williams et al., 2009). In addition, we implemented a strategy that pre-arranges atoms into cuboidal sub-sections, each of which can in turn be further subdivided, yielding a multi-level “octree” (Fig. 2A, Supplemental Figure S2) (Meagher, 1982), which is a type of data structure for accelerating geometric modelling. For an incoming probe, IMPACT first assesses whether or not it has collided with the bounding box of the outer level of the octree. If so, the smaller boxes contained inside the outer box are interrogated for collision with the probe, and so on until reaching the deepest level where the atoms are stored. The probe will in most cases only hit a minority of boxes, enabling a faster calculation as many atoms can be omitted from the search for collisions. The process of subdivision adds overhead to the calculation, however for assemblies  $\gg 100$  kDa (Supplementary Table S2) such as the ribosome (protein and nucleic acid, 2.4 MDa), satellite tobacco necrosis virus (STNV; capsid, genome, salt, and water; 1.8 MDa), the vault (3.5 MDa),



and the adenovirus capsid (89 MDa) we obtained an additional acceleration up to 20-fold (Fig. 2B, Supplemental Information). Importantly, the introduction of octrees does not result in a loss of accuracy, returning the same CCS values as without. We find that there is a strong correlation between the number of atoms and optimal number of subdivisions (maximum octree depth) (Supplemental Figure S2D), which is exploited by IMPACT at runtime for maximum performance.

To test the impact of these strategies we performed a benchmark between IMPACT and other available CCS calculators (Supplemental Information), using the 170 kDa asymmetric unit from the Norwalk virus capsid protein (PDB code 1IHM) to compare with other studies (Bleiholder et al., 2011; Paizs, 2014). We find that to approach a precision of 1%, the TJM requires 17 h, and the EHSS 28 min. The time taken for the PA varies between different implementations, with the fastest existing form converging in 0.43 s. By comparison, IMPACT requires only 70 ms to reach completion. The speed improvement of IMPACT is therefore substantial, ranging approximately between one and six orders of magnitude when compared to alternative CCS calculators (Figs. 2C). Notably, disk access constitutes approximately 20% of IMPACT's wall time for these calculations, indicating that when data is supplied from a coordinate file, rather than being available and properly formatted in rapid-access memory, the performance of IMPACT is close to the unavoidable limitations posed by the hardware.

### **The structural proteome displays significant variation in collision cross-section**

The accuracy, precision and speed of IMPACT allow us to interrogate large structural datasets. We therefore set out to determine the CCS for all the biological assemblies in

the PDBe (Gutmanas et al., 2014), comprising >300,000 structure models, a calculation that took ~5 hrs on a single processor (Supplemental Experimental Procedures). In order to obtain an unbiased depiction, we reduced the repetition in this database by selecting only a single structure when several were available for the same sequence, based on which was most likely according to the PISA algorithm (Krissinel and Henrick, 2007). This resulted in a curated subset of 266,516 structures, representing the entire structural proteome for which atomic co-ordinates exist. We also calculated the mass of these structures, based on summing the mass of all residues present in the coordinate file, in order to mitigate for unresolved regions of sequence. We find that, across the PDBe, CCS is approximately proportional to mass to the two-thirds ( $m^{2/3}$ ) (Fig. 3A). This is consistent with experimental observations made for smaller datasets (Bush et al., 2010; Kaddis et al., 2007; Ruotolo et al., 2008), and is a scaling predicted for structures of any shape with a constant density.

Many proteins in the PDBe are redundant in the sense that they are represented by several entries, where there is range of mutants or otherwise closely related structures. As this may lead to bias in the trend of CCS versus mass, we examined additionally the PiQSi database, a manually curated and non-redundant database of 1,755 biological assembly structures (Levy, 2007). This analysis of the PiQSi database also fits well to the CCS scaling with  $m^{2/3}$  (Supplemental Figure), and allows us to confidently fit a relation between CCS and mass for a typical protein, of  $\Omega_{\text{eff}}=0.457m^{2/3}$

By approximating globular proteins as spheres, entailing  $\Omega_{\text{sphere}} = \pi(3m/4\pi\rho)^{2/3}$ , effective gas-phase densities,  $\rho_{\text{eff}}$  of 0.37 Da/Å<sup>3</sup> (Bush et al., 2010; Kaddis et al., 2007; Ruotolo et al., 2008) and 0.48 Da/Å<sup>3</sup> (Bush et al., 2010; Kaddis et al., 2007; Ruotolo et al., 2008) have been inferred from experimental CCS. At face value these appear

inconsistent with the density of  $0.87 \text{ Da}/\text{\AA}^3$  reported from X-ray structures (Fischer et al., 2004). With the CCSs of all proteins in the PDBe and PiQSi at our disposal we are in a position to assess whether this difference arises from the assumption of spherical proteins (Supplemental Information), or if the lower density reflects non-native conformations in the experiments. We find that CCSs for native protein structures are consistently larger than what is expected for the approximation of perfect spheres, and determine a  $\rho_{eff}$  of  $0.33 \text{ Da}/\text{\AA}^3$  and  $0.31 \text{ Da}/\text{\AA}^3$  for the structures in PDBe and PiQSi, respectively (Supplemental Figure S3B-C), consistent with earlier findings. Since  $\rho_{eff}$  was obtained using native structures, our results suggest that it is the simplistic assumption of spherical proteins that is responsible for the apparent deviation from the solution-phase density of proteins.

In addition to these general trends, the CCSs are distributed around the fitted line for both PDBe and PiQSi. Examination of the CCSs for proteins of similar mass in both the PDBe and PiQSi by taking “slices” of data at different mass windows (e.g. 80-90 kDa, Fig. 3B) revealed that the deviation is larger than the 1% precision used in the CCS calculations. This suggests that, providing the resolving power of the instrumentation is sufficient, proteins of similar mass can often be separated with IM.

### **A shape factor enables facile assessment of unknown structures and the shape variations in the proteome**

In order to interrogate the variation in CCS more succinctly we decoupled the shape variations from the inherent scaling of CCS with  $m^{2/3}$ . Analogous to the Perrin friction factor (Perrin, 1936), we can quantify how the CCS of a given structure,  $\Omega$ , deviates from

the line fitted through the CCS distribution (Fig. 3A, Supplemental Figure S3A) invoking a dimensionless shape factor  $\omega \equiv \Omega/\Omega_{fit}$ . This can be derived from any experimentally determined CCS given the mass, which is itself an outcome of the IM-MS experiment.  $\omega$  will be equal to unity when a protein has structural characteristics of an “average” protein at that given mass,  $<1$  when a protein is more compact, and  $>1$  when it is more extended than typical (e.g. high aspect ratio, or containing a cavity). Similar to other shape factors in the physical sciences, by enabling comparison to the average,  $\omega$  can therefore be used to provide gross structural information on an unknown (Hewitt et al., 2014).

Examining  $\omega$  as a function of mass provides an overview of the distribution of protein shapes in the PDBe (Fig. 3C). In the mass range  $<200$  kDa, the distribution of  $\omega$  is centred close to unity, increasing slightly with mass. In this same range, the relative standard deviation,  $\sigma_r$ , is approximately 10%, indicating that there are considerable structural variations between proteins of similar mass. Above 200 kDa, where PDBe data are sparser,  $\omega$  and  $\sigma_r$  both increase markedly. Notably, the variation of 10% or more is significantly in excess of the typical resolution ( $<2.5\%$ ) of modern IM-MS instrumentation used in the study of protein assemblies (Zhong et al., 2011) and discrepancy ( $<3\%$ ) between CCS measurement and estimation (Benesch and Ruotolo, 2011). This indicates the discriminatory ability of not just CCS as a structural quantity, but also the utility of current IM-MS experiments. These results suggest that the shape factor is a suitable means to distinguish between types of structure in unknown proteins, and potentially useful for relating proteins of unknown structure to structural homologues in the PDBe.

## Using collision cross-sections in integrative structural biology

Even though the CCS provides useful information about the structure of a given protein, it is not sufficient for complete atomistic structure determination. Hence, IM-MS is most useful for structural biology when combined with other structural data. For IMPACT to be of wide practical utility, it is therefore important that it be able to accept inputs from a variety of experimental sources. Consequently we designed it such that the CCSs of coarse-grained models, including those obtained from SAXS or EM (Fig 4A), can be determined by allowing for custom atomic radii to be read from a parameter file. To exemplify the utility of this feature we created bead models from an EM density map of GroEL (EMD code 1457) by placing beads at all voxels that had an electron density above a specified threshold. By repeating this procedure at different thresholds, we obtained 500 bead models varying in volume, and calculated their CCSs (Fig. 4A). We find that the CCS increases with increasing volume and, when imposing a protein density of  $0.87 \text{ Da}/\text{\AA}^3$  (Fischer et al., 2004), the model with a volume corresponding to the mass of GroEL (801 kDa,  $943,000 \text{ \AA}^3$ ) gave a CCS that closely matched the crystal structure (PDB code 1OEL)(Fig. 4A). At larger volumes, the CCS increased sharply, suggesting that CCS is sensitive to artefacts resulting from over-filling of the EM map, and might therefore be an effective alternative means for thresholding EM data.

*Ab-initio* bead models can also be constructed from SAXS data, in a process that typically involves the pooling of several separate simulations to form an average structure, and filtered such that they match a specified target volume. Using the ATSAS package (Svergun et al., 1995; Volkov and Svergun, 2003) we generated 100 bead models from simulated SAXS data of GroEL, using a range of different target volumes (Fig. 4A). We found that the CCSs calculated for the resulting models increased approximately linearly

with increasing volume. Notably, the model that best matched the mass of GroEL had a CCS 17% lower than the X-ray structure, whereas the model that best matched the CCS of the X-ray structure gave a volume 75% larger than expected. For the latter case, the large volume can in part be attributed to cavities in GroEL being filled up by beads (Fig. 4A, inset), which has little impact on the CCS. The results therefore indicate that CCS hold structural information that is complementary to the SAXS data, and might therefore be used to support *ab-initio* modelling.

### **Enabling collision cross-sections for assessment of structural dynamics**

IM-MS measurements record a snapshot of the structural ensemble present in solution, thereby intrinsically reflecting not just the predominantly populated state of the protein, but also minor conformers (Wytttenbach et al., 2014). The speed of IMPACT allows us to analyse structural ensembles, collections of structures that aim to encompass all the conformers dynamically populated at equilibrium. IM-MS measurements of protein and protein complex ions reveal CCS distributions considerably wider than the instrumental resolution (Koeniger et al., 2006; Zhong et al., 2011). In addition, the widths of these distributions are sensitive to experimental conditions that affect the conformational flexibility in solution (Zhong et al., 2014; Zhou et al., 2014). These observations indicate how IM-MS is sensitive to the structural dynamics of proteins (Wytttenbach et al., 2014). To investigate this quantitatively, we examined two protein ensembles generated from nuclear magnetic resonance spectroscopy (NMR) data, for the small protein ubiquitin (PDB identifiers: 2K39 (Lange et al., 2008) and 2K0X (Bryn Fenwick et al., 2011)). Calculating the CCS distribution of these ensembles reveals only a minor difference in

mean CCS between them ( $\langle\Omega\rangle_{2K39} = 1052 \text{ \AA}^2$ ,  $\langle\Omega\rangle_{2K0X} = 1031 \text{ \AA}^2$ ), and a close correspondence with that obtained from IM-MS ( $\langle\Omega\rangle_{\text{Exp}} \approx 1000 \text{ \AA}^2$ ) (Wytttenbach and Bowers, 2011). There is however significant variation in the width of the CCS distributions, with 2K39 notably wider than the more recently refined 2K0X. Comparison with experimental CCSs reveals much closer correspondence with the latter. This demonstrates that our approach can be used to enable the experimental quantification of conformational heterogeneity, and cross-validation of structural ensembles.

The performance of IMPACT suggests that CCS values could be calculated on-the-fly during MD simulations without unduly retarding their progress. To investigate this possibility we calculated the CCS for a short MD trajectory of lysozyme in the gas phase (Marklund et al., 2009) every 10 ps (Fig 4C) to a precision of 0.5%. This single-core calculation for 1500 structures took 1 min. By comparison, the simulations had been run on several processors for several weeks, confirming IMPACT's ability to restrain MD simulations with IM-MS data. Because CCS calculations have been a limiting factor, attempts to integrate IM data into MD simulations have previously been done by approximating the CCS to the squared radius of gyration,  $R_g^2$ , which is more readily computed (Chirot et al., 2012). A plot of  $R_g^2$  versus CCS for the trajectory analysed here reveals only a weak correlation,  $R^2=0.73$  ( $R^2=0.73$  also for CCS vs  $R_g$ ) (Fig. 4D). This indicates that the two parameters are not generally interchangeable as restraints. Together, these observations demonstrate the potential for enabling the use of IM data to the restrain MD simulations of macromolecules by CCS directly, in order to focus the sampling to conformations that comply with experimental observations.

## DISCUSSION

Recent methodological advances have enabled IM-MS measurements to be made across a wide variety of proteins, including some of the most challenging of biological assemblies (Benesch and Ruotolo, 2011; Konijnenberg et al., 2013; Sharon, 2013; Stengel et al., 2012; Thalassinos et al., 2013; Zhou and Robinson, 2014). It is well known that the conformations of folded proteins and protein assemblies are well reflected in the gas-phase ensemble obtained on the timescale of typical IM-MS experiments (Breuker and McLafferty, 2008; Ruotolo and Robinson, 2006). At first glance this may seem to conflict with the seemingly low effective densities inferred from gas-phase experiments. However, by analysing all biological assemblies in the PDBe and PiQSi we were able to reproduce the experimentally inferred gas-phase density of proteins, which shows that the low apparent density can be explained as a direct consequence of the simplistic way densities have been inferred from CCS. As such our analysis resolves this apparent inconsistency with effective gas-phase protein density being lower than the density of crystal structures, and shows that the reported results are consistent with proteins remaining native-like in vacuum on the timescale of the IM experiment.

From our analyses we found that the spread in CCS at any given mass of the PDBe is significantly in excess of the resolution of modern IM-MS instrumentation. We have introduced the shape factor  $\omega$  in order to assess the averageness of a measured CCS, relative to the structural proteome. The variation in shape factor increases with mass such that CCS is more discriminatory for larger molecules, which may explain why other studies have found the CCS to be relatively insensitive for small protein assemblies (Karaca and Bonvin, 2013). Moreover, if the spread of the shape factor were explained by overall shape anisotropy alone, for which IM is relatively insensitive (Hewitt et al.,



2014), then the typical protein would need to have one dimension several times shorter or longer than the others. This suggests that the CCSs also reflect finer levels of protein structure. While these results indicate that the CCS represents a useful structural restraint for molecular modelling, we have also demonstrated by comparison with NMR data that experimentally derived CCS distributions can hold information about the conformational ensemble that can be inferred from IM spectra. This is consistent with the observation of apparent resolving powers observed for a protein complex being lower than for small molecules (Zhong et al., 2011), the presence of conformations that do not inter-convert on the IM timescale (Koeniger et al., 2006), and the visibility of solution-phase effects in the IM peak widths (Zhong et al., 2014; Zhou et al., 2014).

We have shown that the CCS provides information distinct to that given by the  $R_g$ , rendering it potentially useful for restraining MD simulations. Furthermore, the performance of IMPACT is such that calculating CCS during the MD trajectory does not add significant overhead to the simulation. However, two additional factors will need to be taken into consideration when calculating the CCS of some proteins: the possibility of “collapse” in the gas phase of intrinsically disordered (Pagel et al., 2013) or hinging (Hogan et al., 2011) regions; and different scattering characteristics associated with extended or unusually concave surfaces (Bleiholder et al., 2011). The former consideration affects the accuracy of all methods for CCS calculations, and the latter also applies to experiments in which the protein is intentionally unfolded in the gas phase, prior to IM measurement (Hopper and Oldham, 2009; Ruotolo et al., 2007; Laganowsky et al., 2014). Future investigations will allow refinement of the CCS calculation procedure, to accommodate the possibility of such effects.

In summary, we have demonstrated that our CCS calculator, IMPACT, is well suited to the varied demands of modern structural biology and structural proteomics. It is capable of estimating the CCS of proteins and protein assemblies from both atomic coordinates and low-resolution structural information. IMPACT furthermore displays a dramatic advantage in terms of speed (up to  $10^6$ -fold) relative to alternative methods, without compromising on accuracy of the final measurement. By employing IMPACT to examine the complete structural proteome, as well as specific structural biology data, we have demonstrated that IM has the capability to be a very useful approach for the biosciences: in its own right, as a validation tool, and also in contributing to hybrid approaches that combine information from multiple sources.

## METHODS

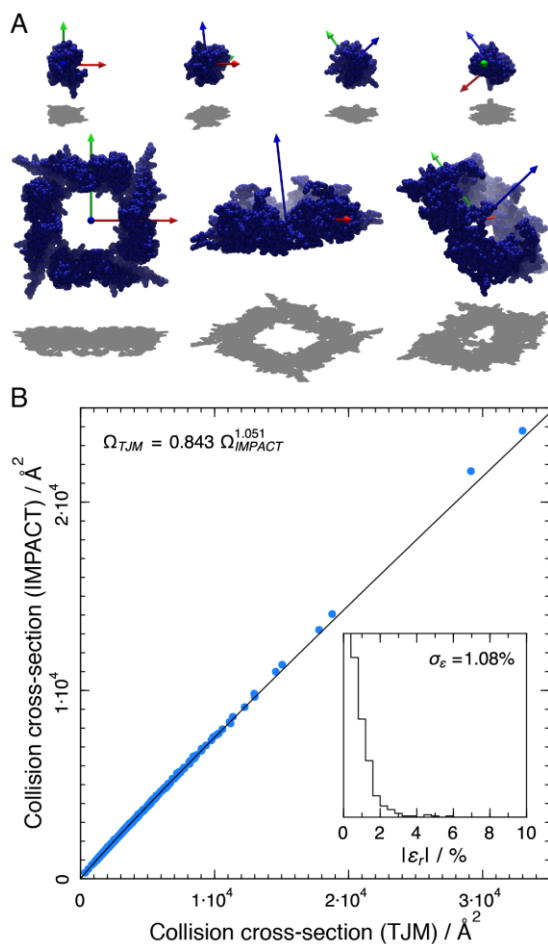
IMPACT is written in the C programming language and builds into both a library for linking with other software, and a standalone command-line tool for Windows, Linux/UNIX and Mac OS X, all available for download at [impact.chem.ox.ac.uk](http://impact.chem.ox.ac.uk), together with documentation for usage and installation. Also available are a list of CCSs for all proteins in the PDB, as well as a tool to search within for proteins with similar mass, CCS, and/or  $\omega$ . IMPACT is currently capable of reading xyz and pdb files with single or multiple models, and atomic radii can optionally be read from a separate file, which effectively allows for the processing of both all-atom and coarse-grained models as described elsewhere in this manuscript, or combinations thereof. A more detailed description and further calculation details are described in Supplemental Experimental Methods.

## **AUTHOR CONTRIBUTIONS**

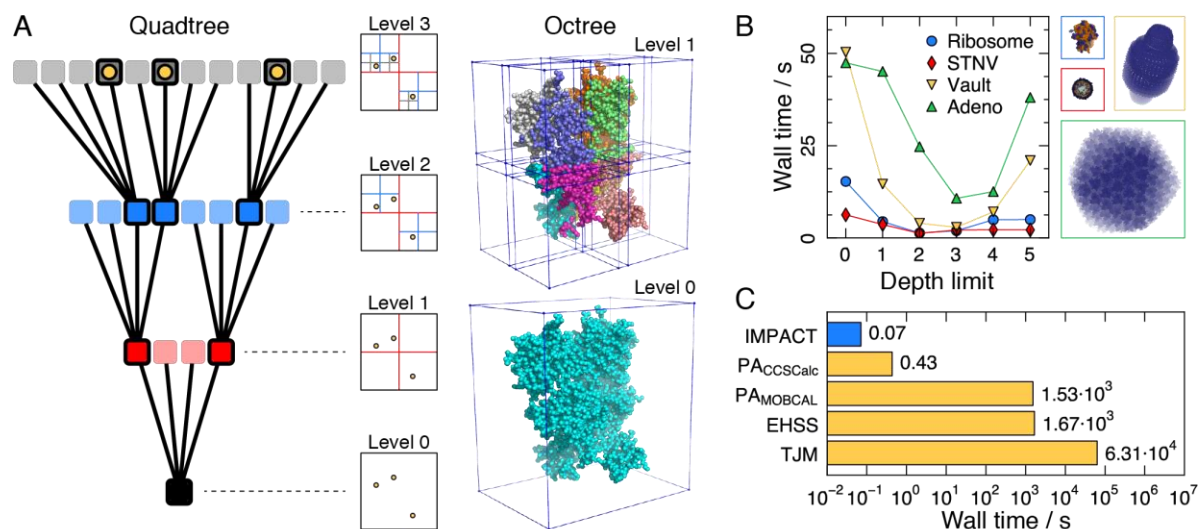
EGM wrote the IMPACT algorithm, performed the calculations, and analysed the data. MTD developed the algorithm for conversion of EM data. AJB performed initial work into accelerating CCS calculations. EGM, AJB, and JLPB designed the experiment and derived the relevant theory. All authors contributed to writing the manuscript.

## **ACKNOWLEDGEMENTS**

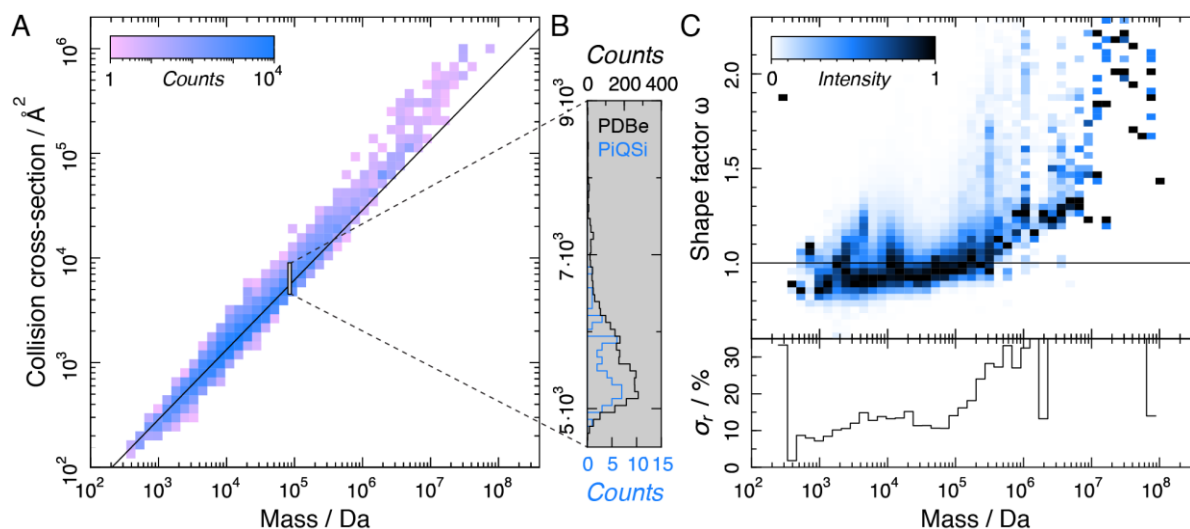
We thank Christian Bleiholder (Florida State University) and Michael T Bowers (University of California, Santa Barbara) for providing datasets for benchmarking, and insightful discussions. We thank Keith Richardson (Waters Corp.) for insight into his CCS calculation approaches and helpful suggestions. We thank Gerard Kleywegt and Sameer Velankar (EMBL-EBI) for providing the assembled contents of PDBe. EGM is supported by The Carl Trygger Foundation for Scientific Research, the Swedish Royal Academy of Sciences, and Helge Ax:son Johnson Foundation; MTD is supported by the Swiss National Science Foundation, CVR is a Royal Society Professor, AJB holds a David Phillips Fellowship from the BBSRC, and JLPB holds a Royal Society University Research Fellowship. CVR and JLPB thank the Medical Research Council for funding through the Confidence in Concept scheme, and JLPB the Biotechnology and Biological Sciences Research Council for grant BB/K004247/1.



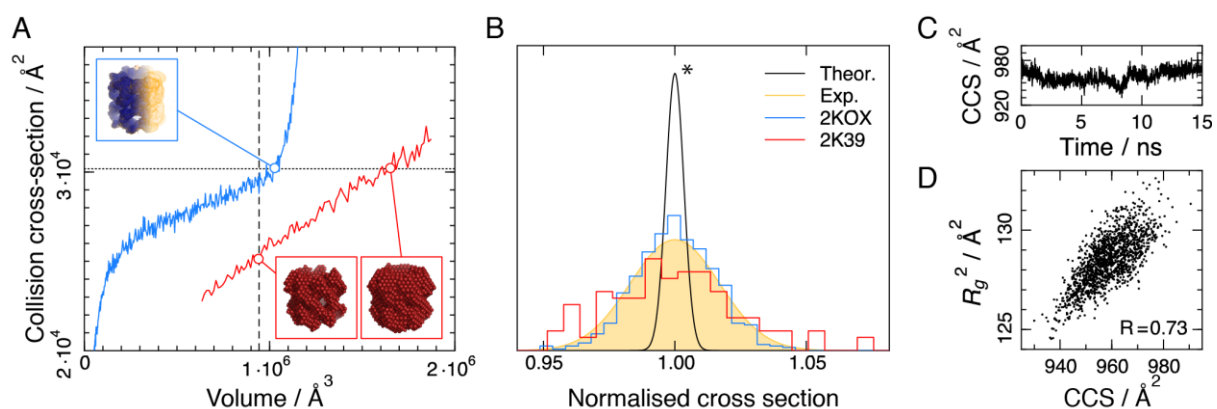
**Figure 1: An accurate CCS calculator.** **A:** The PA approximates the CCS as the rotationally averaged projected area of the target molecule, adjusted for the finite radii of the IM gas probes. The molecule is rotated randomly many times during the calculations to sample rotational space, and the average projected area is determined through Monte Carlo integration. **B:** Comparing the CCS reported from IMPACT to that obtained by the TJM reveals an excellent correlation. The relative error for the 442 structures in the benchmarking dataset is  $\sim 1\%$  (inset), and show no correlation with mass. Overall, the error is lower than the 3% inherent experimental uncertainty when interpreting these values in structural biology applications. See also Supplemental Figure 1A.



**Figure 2: IMPACT provides a dramatic increase in speed of CCS calculations. A:** Schematic showing in 2D/3D how quadtrees/octrees are constructed for a target through recursive subdivision. A quadtree representation of a structure with three atoms (yellow dots) is shown at depths 0 to 3, and the two first levels of an octree for the *lac*-repressor (PDB code 1EFA). The bounding boxes enclosing the subdivisions at each level lets us omit large parts of the target from the search for collisions with the probe, saving time in the process (see also Supplemental Figure S2A-C). **B:** Computational wall time plotted against maximum octree depth  $D$  for a series of large macromolecular complexes (Supplemental Table S2). Octrees provide the biggest boost to speed for large targets, being almost a factor of 20 for the Vault. IMPACT automatically determines the optimum octree depth in a calculation (Supplemental Figure S2D). **C:** Performance benchmarks, where the CCS of the asymmetric unit from a crystal structure of the Norwalk virus capsid (PDB code 1IHM) was calculated to 1% precision, reveal that IMPACT outperforms other PA implementations, and is approximate  $\sim 10^6$  times faster than TJM without significant loss in accuracy (see Fig 1B).



**Figure 3: The structural proteome displays large variations in CCS.** **A:** Histogram of CCS values of the 266,516 proteins in the PDBe calculated using IMPACT. A clear trend can be seen that follows  $\Omega_{fit} = 0.457m^{2/3} \text{ \AA}^2$  (black, determined from the curated PiQSi dataset, Supplemental Figure S3), which follows the expected scaling law for CCSs. **B:** A slice through this histogram at 80-90 kDa shows the variation in CCS without most of the inherent mass dependence. The large variation reveals the distinguishing power of IM-MS. **C:** The PDBe data expressed using shape factor  $\omega$  (upper panel), and the relative standard deviation of the  $\omega$  as a function of mass. (lower panel). The variations that are observed across the whole mass range are considerably greater than both the experimental error and instrument resolution, which are approximately 3% and 2% respectively, and reveals that the discriminatory power of the IM-MS approach increases with increasing molecular mass. See also Supplemental Figure S3.



**Figure 4: Applying IMPACT to diverse structural biology approaches. A:** IMPACT can analyse coarse-grained representations, such as those generated from EM density maps or SAXS data, to give a CCS value for comparison to other data. Calculating the CCS for the EM map of GroEL (EMD code 1457) as a function of represented volume (blue), reveals that the model matching the volume expected from the mass of GroEL (dashed line) has a CCS very close to that of the X-ray structure (dotted line). Inset is such a bead representation of GroEL, superimposed on the EM density map. *Ab-initio* reconstructions of GroEL from SAXS data simulated from the X-ray structure, were filtered to match a range of volumes (black). The best match with the X-ray structure was found at a volume considerably in excess of that expected for GroEL, indicating that the CCS holds information valuable for the processing of SAXS bead models. **B:** The peak from an experimentally derived IM-MS measurement of 7+ charge state of ubiquitin (yellow), which corresponds to native solution conformations (Wytttenbach and Bowers, 2011), is broader than that expected for a single conformation (Koeniger et al., 2006) (black, scaled down to fit the y-scale of the graph). Traces calculated using IMPACT from two NMR-derived ensembles, 2KOX (blue,  $\langle \Omega \rangle = 1031 \text{\AA}^2$ ) (Bryn Fenwick et al., 2011) and 2K39 (red,  $\langle \Omega \rangle = 1052 \text{\AA}^2$ ) (Lange et al., 2008) reveal that the former is in good agreement with the IM-MS measurement in terms of width, although both ensembles



match the experimental value of  $\sim 1000 \text{ \AA}^2$ . **C:** The CCS was calculated for a 15 ns MD trajectory of lysozyme in the gas phase every 10 ps using IMPACT, taking 1 min, demonstrating the possibility of using IMPACT for restraining MD simulations (upper). The radius of gyration,  $R_g^2$ , has previously been taken as a proxy for CCS (Chirot et al., 2012). For this trajectory, the two quantities were weakly correlated revealing that a CCS and  $R_g$  are sensitive to different molecular properties and are thus complementary quantities (lower).

## REFERENCES

- Baldwin, A.J., Lioe, H., Hilton, G.R., Baker, L.A., Rubinstein, J.L., Kay, L.K., and Benesch, J.L.P. (2011). The polydispersity of  $\alpha$ B-crystallin Is rationalized by an interconverting polyhedral architecture. *Structure* 19, 1855-1863.
- Benesch, J.L.P., and Ruotolo, B.T. (2011). Mass spectrometry: come of age for structural and dynamical biology. *Curr Opin Struct Biol* 21, 641-649.
- Bleiholder, C., Wyttenbach, T., and Bowers, M.T. (2011). A novel projection approximation algorithm for the fast and accurate computation of molecular collision cross sections (I). *Method. Int J Mass Spectrom* 308, 1-10.
- Bohrer, B.C., Mererbloom, S.I., Koeniger, S.L., Hilderbrand, A.E., and Clemmer, D.E. (2008). Biomolecule analysis by ion mobility spectrometry. *Annu Rev Anal Chem* 1, 293-327.
- Breuker, K., and McLafferty, F.W. (2008). Stepwise evolution of protein native structure with electrospray into the gas phase,  $10^{-12}$  to  $10^2$  s. *Proc Natl Acad Sci USA* 105, 18145-18152.
- Bryn Fenwick, R., Esteban-Martin, S., Richter, B., Lee, D., Walter, K.F.A., Milovanovic, D., Becker, S., Lakomek, N.A., Griesinger, C., and Salvatella, X. (2011). Weak long-range

correlated motions in a surface patch of ubiquitin involved in molecular recognition. *J Am Chem Soc* *133*, 10336-10339.

Bush, M.F., Hall, Z., Giles, K., Hoyes, J., Robinson, C.V., and Ruotolo, B.T. (2010). Collision cross sections of proteins and their complexes: a calibration framework and database for gas-phase structural biology. *Anal Chem* *82*, 9557-9565.

Chirot, F., Calvo, F., Albrieux, F., Lemoine, J., Tsybin, Y.O., and Dugourd, P. (2012). Statistical analysis of ion mobility spectrometry. I. Unbiased and guided replica-exchange molecular dynamics. *J Am Soc Mass Spectrom* *23*, 386-396.

Fischer, H., Polikarpov, I., and Craievich, A.F. (2004). Average protein density is a molecular-weight-dependent function. *Protein Sci* *13*, 2825-2828.

Gutmanas, A., Alhroub, Y., Battle, G.M., Berrisford, J.M., Bochet, E., Conroy, M.J., Dana, J.M., Fernandez Montecelo, M.A., van Ginkel, G., Gore, S.P., *et al.* (2014). PDBe: protein data bank in europe. *Nucleic Acids Res* *42*, D285-D291.

Hewitt, D., Marklund, E., Scott, D.J., Robinson, C.V., and Borysik, A.J. (2014). A hydrodynamic comparison of solution and gas phase proteins and their complexes. *J Phys Chem B* *118*, 8489-8495.

Hogan, C.J., Ruotolo, B.T., Robinson, C.V., and De la Mora, J.F. (2011). Tandem differential mobility analysis-mass spectrometry reveals partial gas-phase collapse of the GroEL complex. *J Phys Chem B* *115*, 3614-3621.

Hopper, J.T.S., and Oldham, N.J. (2009). Collision induced unfolding of protein ions in the gas phase studied by ion mobility-mass spectrometry: the effect of ligand binding on conformational stability. *J Am Soc Mass Spectrom* *20*, 1851-1858.

Jurneczko, E., and Barran, P.E. (2011). How useful is ion mobility mass spectrometry for structural biology? The relationship between protein crystal structures and their collision cross sections in the gas phase. *Analyst* *136*, 20-28.

Kaddis, C.S., Lomeli, S.H., Yin, S., Berhane, B., Apostol, M.I., Kickhoefer, V.A., Rome, L.H., and Loo, J.A. (2007). Sizing large proteins and protein complexes by electrospray ionization mass spectrometry and ion mobility. *J Am Soc Mass Spectrom* 18, 1206-1216.

Karaca, E., and Bonvin, A.M.J.J. (2013). On the usefulness of ion-mobility mass spectrometry and SAXS data in scoring docking decoys. *Acta Crystallogr Sect D: Biol Crystallogr* 69, 683-694.

Koeniger, S.L., Merenbloom, S.I., and Clemmer, D.E. (2006). Evidence for many resolvable structures within conformation types of electrosprayed ubiquitin ions. *J Phys Chem B* 110, 7017-7021.

Konijnenberg, A., Butterer, A., and Sobott, F. (2013). Native ion mobility-mass spectrometry and related methods in structural biology. *Biochim Biophys Acta* 1834, 1239-1256.

Krissinel, E., and Henrick, K. (2007). Inference of macromolecular assemblies from crystalline state. *J Mol Biol* 372, 774-797.

Laganowsky, A., Reading, E., Allison, T.M., Ulmschneider, M.B., Degiacomi, M.T., Baldwin, A.J., and Robinson, C.V. (2014). Membrane proteins bind lipids selectively to modulate their structure and function. *Nature* 510, 172-175.

Lange, O.F., Lakomek, N.-A., Farès, C., Schröder, G.F., Walter, K.F.A., Becker, S., Meiler, J., Grubmüller, H., Griesinger, C., and Groot, B.L.d. (2008). Recognition dynamics up to microseconds revealed from an rdc-derived ubiquitin ensemble in solution. *Science* 320, 1471-1475.

Larriba, C., and Hogan Jr, C.J. (2013). Free molecular collision cross section calculation methods for nanoparticles and complex ions with energy accommodation. *J Comput Phys* 251, 344-363.

- Levy, E.D. (2007). PiQSi: protein quaternary structure investigation. *Structure* *15*, 1364-1367.
- Mack, E. (1925). Average cross-sectional areas of molecules by gaseous diffusion methods. *J Am Chem Soc* *47*, 2468-2482.
- Marklund, E.G., Larsson, D.S.D., van der Spoel, D., Patriksson, A., and Caleman, C. (2009). Structural stability of electrosprayed proteins: temperature and hydration effects. *Phys Chem Chem Phys* *11*, 8069-8078.
- Meagher, D. (1982). Geometric modeling using octree encoding. *Comput Gr Image Process* *19*, 129-147.
- Mesleh, M.F., Hunter, J.M., Shvartsburg, A.A., Schatz, G.C., and Jarrold, M.F. (1996). Structural information from ion mobility measurements: effects of the long-range potential. *J Phys Chem* *100*, 16082-16086.
- Pagel, K., Natan, E., Hall, Z., Fersht, A.R., and Robinson, C.V. (2013). Intrinsically disordered p53 and its complexes populate compact conformations in the gas phase. *Angew Chem, Int Ed* *52*, 361-365.
- Paizs, B. (2014). A divide-and-conquer approach to compute collision cross sections in the projection approximation method. *Int J Mass Spectrom*, [doi:10.1016/j.ijms.2014.10.005](https://doi.org/10.1016/j.ijms.2014.10.005).
- Perrin, F. (1936). Mouvement Brownien d'un ellipsoïde (II). Rotation libre et dépolarisation des fluorescences. Translation et diffusion de molécules ellipsoïdales. *J Phys Radium* *7*, 1-11.
- Robinson, C.V., Sali, A., and Baumeister, W. (2007). The molecular sociology of the cell. *Nature* *450*, 973-982.

Ruotolo, B.T., Benesch, J.L., Sandercock, A.M., Hyung, S.J., and Robinson, C.V. (2008). Ion mobility-mass spectrometry analysis of large protein complexes. *Nat Protoc* *3*, 1139-1152.

Ruotolo, B.T., Hyung, S.-J., Robinson, P.M., Giles, K., Bateman, R.H., and Robinson, C.V. (2007). Ion mobility-mass spectrometry reveals long-lived, unfolded intermediates in the dissociation of protein complexes. *Angew Chem, Int Ed* *46*, 8001-8004.

Ruotolo, B.T., and Robinson, C.V. (2006). Aspects of native proteins are retained in vacuum. *Curr Op Chem Biol* *10*, 402-408.

Russel, D., Lasker, K., Phillips, J., Schneidman-Duhovny, D., Velazquez-Muriel, J.A., and Sali, A. (2009). The structural dynamics of macromolecular processes. *Curr OpCell Biol* *21*, 97-108.

Sharon, M. (2013). Structural MS Pulls Its Weight. *Science* *340*, 1059-1060.

Shvartsburg, A.A., and Jarrold, M.F. (1996). An exact hard-spheres scattering model for the mobilities of polyatomic ions. *Chem Phys Lett* *261*, 86-91.

Stengel, F., Aebersold, R., and Robinson, C.V. (2012). Joining Forces: Integrating Proteomics and Cross-linking with the Mass Spectrometry of Intact Complexes. *Mol Cell Proteomics* *11*, R111.014027.

Svergun, D., Barberato, C., and Koch, M.H.J. (1995). CRY SOL - a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. *J Appl Crystallogr* *28*, 768-773.

Thalassinos, K., Pandurangan, Arun P., Xu, M., Alber, F., and Topf, M. (2013). Conformational states of macromolecular assemblies explored by integrative structure calculation. *Structure* *21*, 1500-1508.

Utrecht, C., Rose, R.J., van Duijn, E., Lorenzen, K., and Heck, A.J. (2010). Ion mobility mass spectrometry of proteins and protein assemblies. *Chem Soc Rev* *39*, 1633-1655.

Volkov, V.V., and Svergun, D.I. (2003). Uniqueness of ab initio shape determination in small-angle scattering. *J Appl Crystallogr* 36, 860-864.

Von Helden, G., Hsu, M., Gotts, N., and Bowers, M. (1993). Carbon cluster cations with up to 84 atoms - structures, formation mechanism, and reactivity. *J Phys Chem* 97, 8182-8192.

Ward, A.B., Sali, A., and Wilson, I.A. (2013). Biochemistry. Integrative structural biology. *Science* 339, 913-915.

Williams, J.P., Lough, J.A., Campuzano, I., Richardson, K., and Sadler, P.J. (2009). Use of ion mobility mass spectrometry and a collision cross-section algorithm to study an organometallic ruthenium anticancer complex and its adducts with a DNA oligonucleotide. *Rapid Commun Mass Spectrom* 23, 3563-3569.

Wytttenbach, T., and Bowers, M.T. (2011). Structural stability from solution to the gas phase: native solution structure of ubiquitin survives analysis in a solvent-free ion mobility-mass spectrometry environment. *J Phys Chem B* 115, 12266-12275.

Wytttenbach, T., Pierson, N.A., Clemmer, D.E., and Bowers, M.T. (2014). Ion mobility analysis of molecular dynamics. *Annu Rev Phys Chem* 65, 175-196.

Zhong, Y., Han, L., and Ruotolo, B.T. (2014). Collisional and coulombic unfolding of gas-phase proteins: high correlation to their domain structures in solution. *Angew Chem, Int Ed* 53, 9209-9212.

Zhong, Y., Hyung, S.-J., and Ruotolo, B.T. (2011). Characterizing the resolution and accuracy of a second-generation traveling-wave ion mobility separator for biomolecular ions. *Analyst* 136, 3534-3541.

Zhou, M., Politis, A., Davies, R.B., Liko, I., Wu, K.-J., Stewart, A.G., Stock, D., and Robinson, C.V. (2014). Ion mobility-mass spectrometry of a rotary ATPase reveals ATP-induced reduction in conformational flexibility. *Nat Chem* 6, 208-215.

Zhou, M., and Robinson, C.V. (2014). Flexible membrane proteins: functional dynamics captured by mass spectrometry. *Curr Op Struct Biol* 28, 122-130.