Briefings in Bioinformatics, 2023, **24(5)**, 1–9
https://doi.org/10.1093/bib/bbad308
Advance access publication date 22 August 2023
Problem Solving Protocol

Predicting ion mobility collision cross sections using projection approximation with ROSIE-PARCS webserver

S.M. Bargeen Alam Turzo, Justin T. Seffernick, Sergey Lyskov and Steffen Lindert

Corresponding author. Steffen Lindert, Department of Chemistry and Biochemistry and Resource for Native Mass Spectrometry Guided Structural Biology, Ohio State University, Columbus, OH 43210, USA. E-mail: lindert.1@osu.edu

Abstract

Ion mobility coupled to mass spectrometry informs on the shape and size of protein structures in the form of a collision cross section (CCS_{IM}). Although there are several computational methods for predicting CCS_{IM} based on protein structures, including our previously developed projection approximation using rough circular shapes (PARCS), the process usually requires prior experience with the command-line interface. To overcome this challenge, here we present a web application on the Rosetta Online Server that Includes Everyone (ROSIE) webserver to predict CCS_{IM} from protein structure using projection approximation with PARCS. In this web interface, the user is only required to provide one or more PDB files as input. Results from our case studies suggest that CCS_{IM} predictions (with ROSIE-PARCS) are highly accurate with an average error of 6.12%. Furthermore, the absolute difference between CCS_{IM} and CCS_{PARCS} can help in distinguishing accurate from inaccurate AlphaFold2 protein structure predictions. ROSIE-PARCS is designed with a user-friendly interface, is available publicly and is free to use. The ROSIE-PARCS web interface is supported by all major web browsers and can be accessed via this link (https://rosie.graylab.jhu.edu).

Keywords: ion mobility, protein structure prediction, collision cross section, mass spectrometry, webserver, AlphaFold2

INTRODUCTION

Proteins are key functional units that lay the groundwork for many biological processes such as cell signaling, immune function and metabolism. Therefore, knowledge of protein structure is crucial for understanding their roles in biological functions and developing new therapeutics. Protein structures can be experimentally determined to atomic detail using techniques such as X-ray crystallography, cryo-electron microscopy (cryo-EM) and nuclear magnetic resonance (NMR) spectroscopy. However, these methods are limited by factors such as size, resolution, purification conditions, sample quantity and time to name a few [1]. Due to these limitations, routine analyses are not always able to determine structure to atomistic resolution.

Mass spectrometry (MS) is a widely used analytical technique in structural biology because it can rapidly provide structural information of proteins. Some benefits of MS are that it can be used on a variety of samples, requires minimal sample preparation and can be easily incorporated into various stages of a research pipeline. Although native MS is a gas phase technique, several studies have shown that key aspects of protein structure, including elements of secondary structure, compactness and protein–protein interactions, can be retained when proteins transition from a solution to gas phase [2–5]. There are many techniques that can be used in conjunction with MS to study protein structures [2, 6–14]. Although such experiments can provide rich information about the structure of proteins, the data obtained are often sparse and not enough to resolve the structure to atomistic detail. Thus, computational methods along with sparse experimental data can be used in an integrative approach to further enhance the understanding of protein structures [15–22]. Particularly, several studies [6, 13, 23–47] show that sparse data from various MS methods have played major roles in integrative structural biology frameworks.

Ion mobility (IM) provides structural information in the form of protein shape and size. IM data supplemented with computational methods can improve protein structure prediction [4, 34–36, 46, 48]. Briefly, IM is a separation technique used to measure the movement of ions through a gas under the influence of an electric field [49, 50]. Particularly, in structural biology, it is used in the analysis of proteins and other biomolecules [51], where it can provide information about the shape and size of the analyte. IM provides this shape and size information in the form of a collision cross section (CCS) of a protein. The CCS_{IM} value quantifies the amount of momentum that is exchanged between the ion and the buffer gas during collisions, and can be thought of as the rotationally averaged cross-sectional area [52].

There are several methods for predicting CCS_{IM} from structure [46, 52–59], varying widely with respect to accuracy and speed. Among these methods, the projection approximation (PA) is the simplest and fastest method to predict CCS_{IM} [52]. In the PA method, atoms in the protein are treated as hard spheres with a predetermined radius. A rotation matrix is employed to randomly sample various orientations, and subsequently a Monte Carlo integration method is used to determine the projection area of the protein structure at each orientation. The projection area is calculated from the ratio of probe particles that are directly in contact with the projected atoms to the total number of probe particles. Usually, a large number of probe particles are required to obtain an accurate projection area. CCS_{PA} is then obtained from averaging the projection areas across a large number of random orientations [59]. Because the PA method relies on a basic hard-sphere surface, it is impossible to incorporate temperature and

Received: April 20, 2023. **Revised:** July 3, 2023. **Accepted:** August 8, 2023

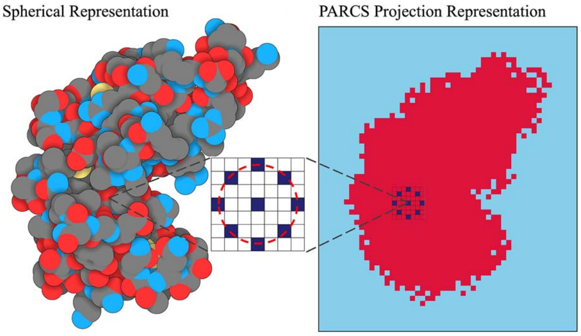


Figure 1. Illustration of the PARCS algorithm in Rosetta. In PARCS, each protein atom is estimated as a rough circle using a 9-point approximation based on the respective atomic and buffer gas radii. The PARCS projection representation is obtained after iterating through every single atom in the structure.

charge state effects into the CCS calculation [60]. Additionally, the PA method cannot accurately account for the concavity in protein structures [60]. However, due to its speed and accuracy, CCS_{PA} is a very popular method with integrative modeling [4, 29, 34–36, 38, 45, 46, 61–63], which often requires fast calculations for on-the-fly structure assessment. Recently, we developed the projection approximation using rough circular shapes [46] (PARCS), a novel approach to PA, to accurately predict CCS_{IM} in the Rosetta molecular modeling suite [18, 64, 65]. PARCS calculates the projection area by first casting the protein structure onto a 2D grid. PARCS then estimates the area of the 2D projection by filling the grid using a 9-point circle approximation for each atom (Figure 1). Finally, the CCS_{IM} prediction from PARCS (CCS_{PARCS}) is obtained by performing this calculation several times at different random rotations and averaging the projection area. Although our PARCS method (developed within the Rosetta framework) is highly accurate and successful in predicting CCS_{IM}, up to this point, usage of PARCS has required users to be familiar with the command-line interface in a Unix environment. Often it may also have required access to high-performance computing and being proficient in programming languages for data analysis. Thus, usage of PARCS may have been intimidating and difficult for non-technical users. On the contrary, a web server interface provides a more user-friendly and accessible way to interact with such applications for a wide range of users. The availability of web interfaces for predicting CCS_{IM} is currently limited. Notably, it is possible to predict CCS_{IM} via the popular Projection Super Approximation method through their intuitive webserver [57]. However, at the time of this study, we could not find any active web interface to predict CCS_{IM} for protein structures using the PA method.

Therefore to bridge this gap, in this study, we introduced a web server interface for our PARCS method within the Rosetta Online Server that Includes Everyone [66] (ROSIE). Briefly, ROSIE is an online platform that allows researchers to create, host and/or use web-based tools and interfaces for protein structure prediction and analysis using the Rosetta [18] software package. Here, we first showcased the simplicity of the PARCS web interface within ROSIE. We then verified the results of ROSIE-PARCS against the command-line interface of PARCS (CLI-PARCS). Upon successful verification, we then highlighted the effectiveness of the PARCS web application with two different case studies. In the first case study, we showed the agreement of the CCS_{IM} and the CCS_{PARCS} for known protein structures from the Protein Data Bank [67] (PDB) with various structural complexity. In the second case study, we showed that the absolute difference between CCS_{IM} and CCS_{PARCS} helped in distinguishing accurate from inaccurate AlphaFold2 protein structure predictions. To do this, we first predicted protein structures with AlphaFold2 [68] (AF2) and calculated the absolute difference (ΔCCS) between the CCS_{IM} and the CCS_{PARCS} for the AF2 predicted structures. We then compared the ΔCCS to the root mean squared deviation (RMSD) of the AF2 predicted structure from its native protein structure. Our verification tests showed that, as expected, ROSIE-PARCS performed exactly like CLI-PARCS. Additionally, our case studies demonstrated that ROSIE-PARCS can accurately predict IM data and can differentiate between accurate and inaccurate models of protein structures. ROSIE-PARCS, freely available to everybody, can be accessed via the following link (https://rosie.graylab.jhu.edu).

METHODS

IM coupled to MS can provide shape and size information of proteins in their native state. The shape and size information can be derived from IM in the form of collision cross sections [14]. PA is a method that can predict CCS values for a given protein structure. Typically, in the PA method, the projection area of a protein structure is calculated by first placing the structure in a 2D bounding area (most commonly a rectangle or a circle). Then the 2D area is probed with random particles (meant to mimic the buffer gas particles in the IM experiments). These probe particles that are directly in contact with the atoms (within the bounding area) are considered as 'hits'. The ratio of hits to the total number of probe particles is then multiplied by the area of the bounding area to obtain the projection area [60]. To ensure accuracy of the projection area, a large number of probe particles are usually required [60]. In our previous work [46], we developed a method to approximate CCS_{IM} using the PA using Rough Circular Shapes method in the Rosetta molecular modeling suite. In contrast to previous PA implementations (outlined above), in our PARCS algorithm, the projection area is instead calculated by approximating the projection of each atom as a rough circle with nine points. No probe particles are necessary. More specifically, the PARCS algorithm first takes 3D atomic protein coordinates as input, randomly rotates the structure and projects it onto a 2D grid. For each atom on the grid, the center cell and eight surrounding cells (based on the radii of the projected atom and a buffer gas) are filled. The process is repeated for all atoms in the protein and the projection area is derived by summing the areas of the filled grid cells. From the x-y, y-z and x-z projections at each random rotation, three projection areas are obtained. The CCS_{PARCS} of the structure is then acquired from the average projection area across the total number of random rotations (NRRs). A simplified illustration of PARCS is shown in Figure 1. Here, we developed a PARCS web interface on ROSIE, to make the application more easily accessible and convenient for users to calculate CCS_{PARCS}. In the subsequent sections, we first describe the dataset of protein structures used to test the ROSIE-PARCS web interface. Then, we describe the design and usage of the ROSIE-PARCS web interface.

CCS_{IM} dataset

To test the ROSIE-PARCS web interface, we selected a set of 13 proteins with experimentally determined protein structures available in the PDB as well as experimentally determined CCS_{IM} (for the lowest charge states) [14, 69, 70]. This dataset will be referred to as the CCS_{IM} dataset. The sequence length for these proteins varied from 123 to 4096. Additionally, the CCS_{IM} dataset consisted of five monomers, two dimers (1 homodimer and 1 heterodimer), four tetramers (1 dimer of heterodimers and 3 hometetramers), one homopentamer and one heterohexamer. The proteins in the CCS_{IM} dataset were mostly globular with an average percent disorder of 6.85% as predicted by the RosettaResidueDisorder [71–73] application. The CCS_{IM} dataset is outlined in Supplementary Table 1.

Structure prediction with AlphaFold2 and evaluation metrics

We selected a subset of the proteins in the CCS_{IM} dataset to assess protein structures predicted with AF2 using IM data. The primary sequence obtained from the PDB structures of this dataset (as shown in Supplementary Table 2) was used to predict the structures of proteins with AF2 version 2.2.2. All AF2 predictions were done by setting the template date for homologs to 1900-01-01. This removed the influence of homologous templates on the AF2 network for structure predictions. Protein complexes within the CCS_{IM} dataset were predicted with the multimer options in AF2 version 2.2.2. RMSD and $RMSD_{100}$ [74] were used as the analysis metrics for evaluating the structures predicted with AF2, similar to the analysis conducted in one of our previous studies [44]. We first aligned the AF2 prediction with the known structure (structure obtained from the PDB) and then calculated the RMSD and the $RMSD_{100}$. All alignment and RMSD calculations were done using PyMOL version 2.5.2 [75]. Additionally, the AF2 confidence metric [average predicted local distance difference test (pLDDT)] and CCS_{PARCS} were calculated for all AF2 predictions with PARCS in Rosetta.

Using the ROSIE-PARCS web interface

In general, to use any Rosetta protocol on ROSIE, users first have to create a GitHub account, since ROSIE uses the GitHub authorization service for secure user login. The ROSIE-PARCS web application was designed with an emphasis on both simplicity and ease of use. In order to use the ROSIE-PARCS web application to calculate CCS_{PARCS}, users must supply a protein structure in PDB file format.

Uploading PDB files on ROSIE-PARCS

Screenshots of the submission page of ROSIE-PARCS are shown in Figure 2. In order to upload PDB file(s) on the input page of ROSIE-PARCS, users may drag their PDB file(s) directly into the '*Input PDB file(s)*' box on the webpage as shown by the black dashed box in Figure 2. Users alternatively have the option to upload their PDB file(s) by clicking the '*Browse*' button next to the '*Input PDB file(s)*' box, which will open a drop-down menu (indicated by the black dashed box around the '*Browse*' button in Figure 2). From there, users can navigate to the desired PDB file(s). Additionally, as illustrated in Figure 2, users have a third alternative option to directly enter a PDB ID into the 'PDB Code' box (indicated by the green dashed box), and then click the 'download' button. However, this button should be used with caution when downloading protein complexes. The Supplementary Methods describe how to properly download and save protein complexes that contain all protein subunits. Notably, users are allowed to submit multiple PDB files simultaneously (as exemplified by the orange dashed box in Figure 2) on the ROSIE-PARCS webpage (with an upper-limit of 100 PDB files per job submission). This multiple-PDB option is accessible to any of the methods mentioned above. Once the upload is successful, the webpage will display information about the number of residues, atoms and the chain identifiers in the PDB files that were uploaded.

Option to calculate CCS_{PARCS} by varying the NRRs using ROSIE-PARCS

The NRRs determine the number of projections that are being used to average the projection areas. For example, by default, PARCS uses 300 random rotations to calculate the CCS_{PARCS}. This means that 900 projections are being used to average the projection areas to calculate the CCS_{PARCS}. This default setting was based on the dataset in our previous study [46]. However, for smaller proteins, users may be able to obtain comparable results with a lower NRR. Conversely, for large structures, users may find it necessary to increase the NRR to calculate reliable CCS values. Users can set this variable by changing the input in 'Input number of random rotations. Minimum: 100, Maximum: 1000, Default: 300.' as indicated by the red dashed box in Figure 2. The NRR has an upper limit of 1000 (3000 projections). This upper limit for NRR is adequate based on our previous study [46].

Option to calculate CCS_{PARCS} in different buffer gas conditions using ROSIE-PARCS

The CCS_{IM} varies depending on the buffer gas conditions in which the IM experiment is carried out [76]. PARCS includes the option to predict CCS_{IM} for different buffer gas conditions. This is achieved by controlling the probe radius (PR) option within PARCS. Currently PARCS has been demonstrated to successfully predict CCS_{IM} in helium and nitrogen buffer gas conditions. Therefore, we have included this option as '*Input probe radius in Angstrom. He buffer gas: 1.0 (default), N2 buffer gas: 1.81.*' on the input page as indicated by the blue dashed box shown in Figure 2. By default, PARCS calculates the CCS in helium gas conditions.

Verification of ROSIE-PARCS web application results against command-line interface PARCS

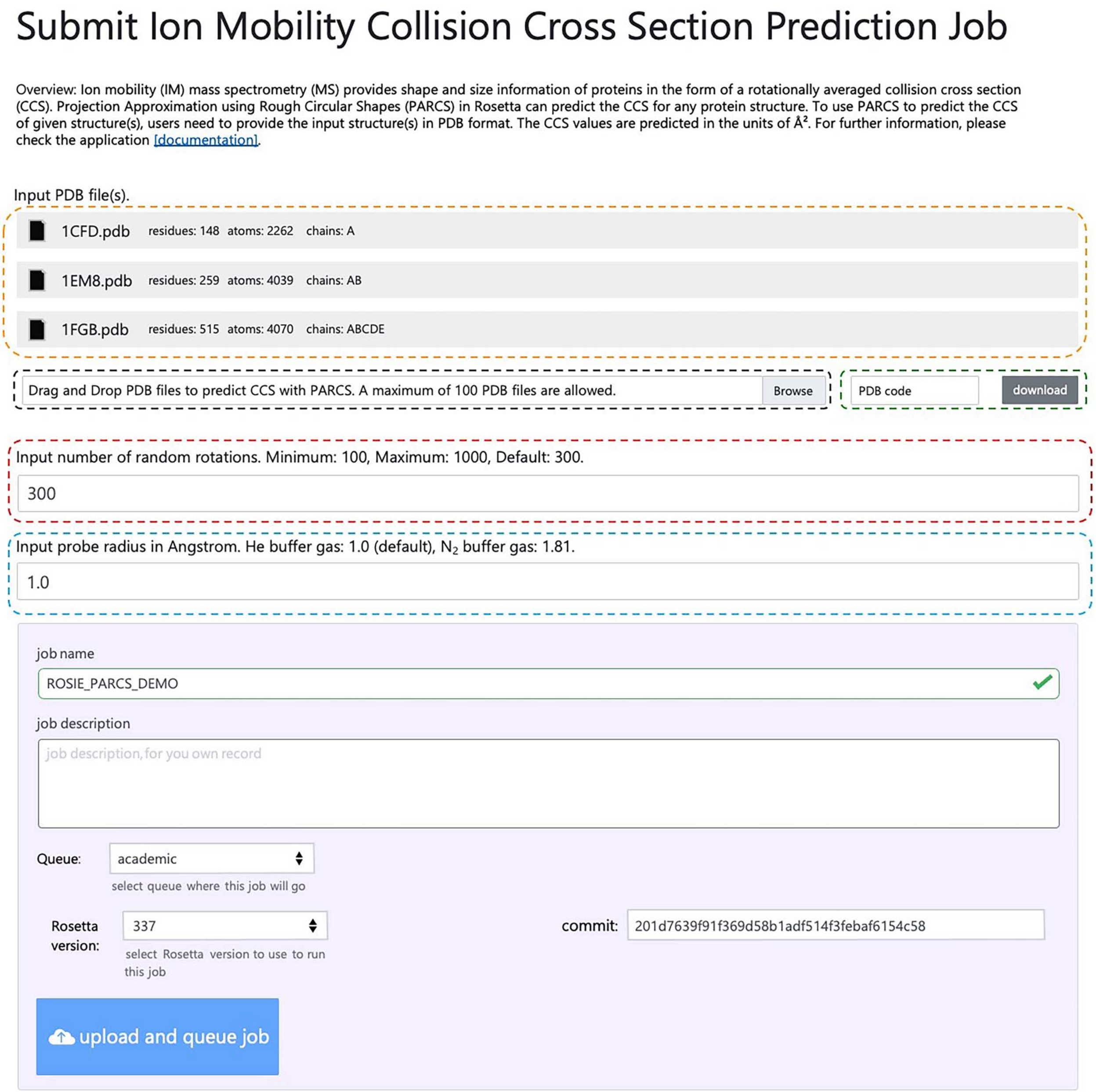
We first set out to verify that the results of ROSIE-PARCS under various parameter settings agreed with the results obtained from the command line version of PARCS (CLI-PARCS). We varied the PR and the NRRs and calculated CCS_{PARCS} for the proteins in the CCS_{IM} dataset. As part of this, we carried out three verification tests. First, the PR for both ROSIE-PARCS and CLI-PARCS was varied from 1.0 to 2.0 Å in increments of 0.2 Å, while keeping the NRR parameter fixed at 300. Next, we varied the NRR from 100 to 600 in increments of 100, while keeping the PR fixed to its default setting (PR = 1.0 Å). Finally, we randomized both input parameters, the PR (within a range of 1.0 –2.0 Å) and the NRR (within a range of 100–600). Using this randomization, we obtained six sets of input parameters (PR, NRR). These were (1.45 Å, 432), (1.22 Å, 218), (1.37 Å, 470), (1.29 Å, 322), (1.24 Å, 160) and (1.78 Å, 231). Using these input parameters, we calculated the respective CCS_{PARCS} for all the proteins in the CCS_{IM} dataset.

Software usage for data analysis

Data analysis was conducted using Python version 3.7.3. Matplotlib version 3.1.2 was utilized to create scatter plots, line plots and violin distributions. The spherical representation of the protein in Figure 1 was generated with the online web application Illustrate [77].

Results and discussions

In this study, we developed a web application on the ROSIE webserver to predict the collision cross section obtained from IM-MS experiments. This web application (ROSIE-PARCS) incorporates our previously developed CCS_{IM} prediction algorithm, PARCS [46],



Submit Ion Mobility Collision Cross Section Prediction Job

Overview: Ion mobility (IM) mass spectrometry (MS) provides shape and size information of proteins in the form of a rotationally averaged collision cross section (CCS). Projection Approximation using Rough Circular Shapes (PARCS) in Rosetta can predict the CCS for any protein structure. To use PARCS to predict the CCS of given structure(s), users need to provide the input structure(s) in PDB format. The CCS values are predicted in the units of Å². For further information, please check the application [documentation].

Input PDB file(s).

1CFD.pdb	residues: 148	atoms: 2262	chains: A
1EM8.pdb	residues: 259	atoms: 4039	chains: AB
1FGB.pdb	residues: 515	atoms: 4070	chains: ABCDE

Drag and Drop PDB files to predict CCS with PARCS. A maximum of 100 PDB files are allowed. Browse

PDB code download

Input number of random rotations. Minimum: 100, Maximum: 1000, Default: 300.

300

Input probe radius in Angstrom. He buffer gas: 1.0 (default), N_2 buffer gas: 1.81.

1.0

job name

ROSIE_PARCS_DEMO ✓

job description

job description, for you own record

Queue: academic

select queue where this job will go

Rosetta version: 337

select Rosetta version to use to run this job

commit: 201d7639f91f369d58b1adf514f3febaf6154c58

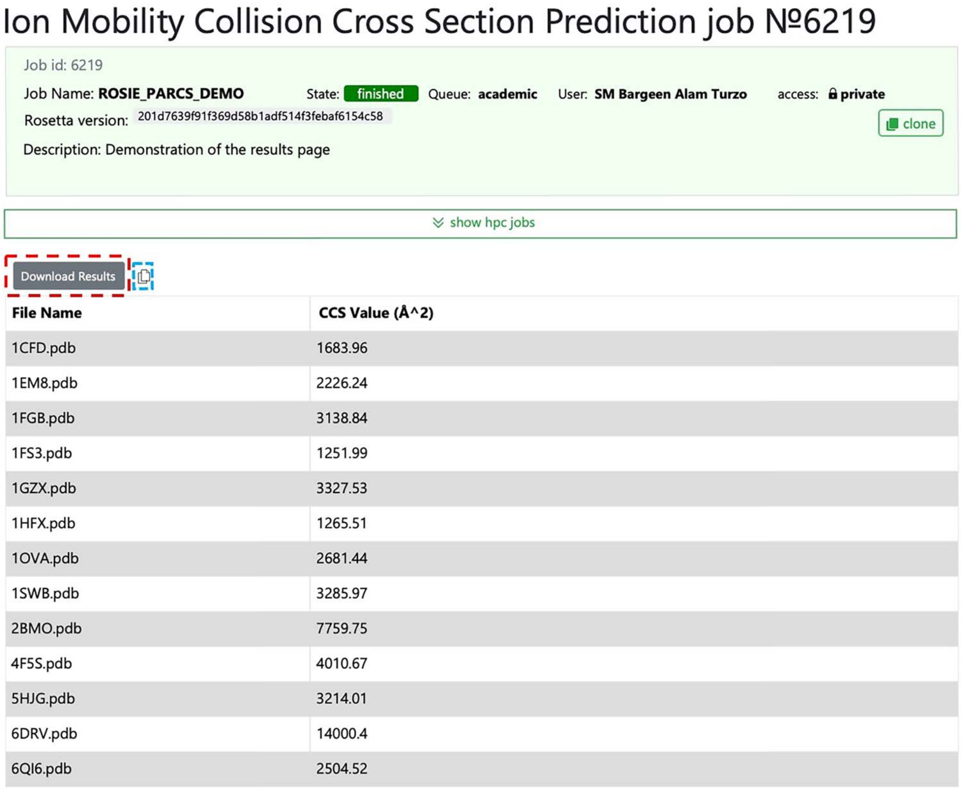
upload and queue job

Figure 2. Input page of the ROSIE-PARCS webpage. PDB file(s) can be dragged into the drop box. Alternatively, users can use the '*Browse*' button (shown within the back dashed box) to navigate their filesystem and/or directly obtain a structure from the PDB by providing the PDB code and pressing the download button (as indicated by the green dashed box). The number of residues, atoms and the chain identifiers are displayed for successfully uploaded PDB files (orange dashed box). The two input parameters, '*number of random rotations*' and '*probe radius in Angstroms*' are shown in the red and blue dashed boxes.

and has been designed to be more user-friendly than its CLI counterpart. ROSIE-PARCS offers several capabilities such as: it can handle multiple PDB files, has options to control application parameters through its user-friendly interface and presents the results in a clear and intuitive manner. Furthermore, ROSIE-PARCS enables users to easily copy or download the $\mathrm{CCS}_{\mathrm{PARCS}}$ results for further analysis. In the following subsections, we first discuss the results display page of ROSIE-PARCS. Next, we verify ROSIE-PARCS against the results obtained from the CLI-PARCS. Then we discuss two case studies of calculating $\mathrm{CCS}_{\mathrm{PARCS}}$ in different scenarios. In the first case study, we show the agreement of the $\mathrm{CCS}_{\mathrm{IM}}$ and the $\mathrm{CCS}_{\mathrm{PARCS}}$ for known protein structures. In the second case study, we calculate the $\mathrm{CCS}_{\mathrm{PARCS}}$ of several AF2-predicted structures and analyze their agreement with $\mathrm{CCS}_{\mathrm{IM}}$ (ΔCCS), as well as the RMSD of the prediction from the known PDB structure (i.e. the ΔCCS versus RMSD).

Efficient results display and management for $\mathrm{CCS}_{\mathrm{PARCS}}$ calculations with ROSIE-PARCS

During the development stage, we focused on visualizing the results from ROSIE-PARCS in a clear and intuitive manner. We developed a clear and user-friendly interface for the ROSIE-PARCS results display page, which enables users to understand and make informed decisions based on the calculation, thus improving their overall experience and efficiency. To achieve this, a successful $\mathrm{CCS}_{\mathrm{PARCS}}$ calculation is always indicated by the presence of the green 'finished' *State* indicator as shown in Figure 3. Following this, the $\mathrm{CCS}_{\mathrm{PARCS}}$ values obtained from ROSIE-PARCS are then



Ion Mobility Collision Cross Section Prediction job №6219

Job id: 6219

Job Name: **ROSIE_PARCS_DEMO** State: finished Queue: **academic** User: **SM Bargeen Alam Turzo** access: **private**

Rosetta version: 201d7639f91f369d58b1adf514f3febaf6154c58

Description: Demonstration of the results page

File Name	CCS Value (Å^2)
1CFD.pdb	1683.96
1EM8.pdb	2226.24
1FGB.pdb	3138.84
1FS3.pdb	1251.99
1GZX.pdb	3327.53
1HFX.pdb	1265.51
1OVA.pdb	2681.44
1SWB.pdb	3285.97
2BMO.pdb	7759.75
4F5S.pdb	4010.67
5HJG.pdb	3214.01
6DRV.pdb	14000.4
6QI6.pdb	2504.52

Figure 3. Results page of the ROSIE-PARCS application. Successful completion of the CCS_{PARCS} calculation is indicated by the green 'finished' *State*. CCS_{PARCS} values are displayed in a tabular format. The table headers are labeled as 'File Name' and 'CCS Value (Å^2)'. The red and blue box are highlighting the 'Download Results' and copy to clipboard buttons, respectively.

displayed in a tabular format. The table displays 'File Name' and 'CCS Value (Å^2)' as its header values (Figure 3). The 'File Name' corresponds to name of the PDB files uploaded by the user. The 'CCS Value (Å^2)' corresponds to the CCS values calculated by PARCS for the respective 'File Name'. In Figure 3, we show the CCS_{PARCS} for all PDB files in the CCS_{IM} dataset. Additionally, the results page also contains two other important features. The first is the download button, 'Download Results' (shown with red dashed box in Figure 3), that allows the user to directly download the displayed tabular data (tab-separated values file format) onto their computer. The second feature is the copy to clipboard button (shown with blue dashed box in Figure 3). This feature allows users to directly copy the table and paste it into other applications (such as text editors, excel sheets, etc.) for further analysis. Therefore, ROSIE-PARCS allows efficient analysis and access to the CCS_{PARCS} values in a clear and intuitive format.

CCS_{PARCS} calculations of ROSIE-PARCS are identical to CLI-PARCS

To test the implementation of the ROSIE-PARCS web application, we first verified that the CCS_{PARCS} calculations from ROSIE-PARCS matched those of CLI-PARCS, from our previous study [46]. This was verified for a range of different probe radii and NRRs in the PARCS algorithm. We used the CCS_{IM} dataset for this test. In the first verification test, we examined the effect of varying PR (while keeping NRR fixed at 300) on both ROSIE-PARCS and CLI-PARCS. We obtained the CCS_{PARCS} from both ROSIE-PARCS and CLI-PARCS by increasing the PR from 1.0 to 2.0 Å by 0.2 Å increments for all the proteins in the CCS_{IM} dataset. As expected, we observed that increasing the PR increased the CCS_{PARCS} for both ROSIE-PARCS (blue) and CLI-PARCS (orange) as shown in Figure 4A. The average sequence-length-normalized percent difference between ROSIE-PARCS and CLI-PARCS in calculated CCS_{PARCS} was 0.033, 0.023, 0.022, 0.016, 0.071 and 0.068% for the PR 1.0 , 1.2 , 1.4 , 1.6 , 1.8 and 2.0 Å, respectively. The slight variations in CCS_{PARCS} were due to stochastic features in PARCS, namely the random rotation matrices used to calculate CCS_{PARCS}. These results (as outlined in Supplementary Table 3) were expected and indicated that the predictions from ROSIE-PARCS under various PR settings matched those of CLI-PARCS. In the second verification test, we examined the effect of varying NRR (while keeping the PR constant at 1.0 Å) from 100 to 600, by increments of 100. We again tested this on the CCS_{IM} dataset. In this verification test, we again observed the expected trend, where CCS_{PARCS} neither increased nor decreased across different NRR. This trend was observed for both ROSIE-PARCS and CLI-PARCS as shown in Figure 4B. On average, the normalized CCS_{PARCS} from ROSIE-PARCS and CLI-PARCS differed by only 0.02%. These results are outlined in detail in Supplementary Table 4. For the final verification test, we randomized both the input parameters (PR and NRR). PR was randomized within the range of 1.0 –2.0 Å, and NRR was randomized within the range of 100–600. From this strategy, we obtained six random sets of PR and NRR as outlined in Supplementary Table 5. Using these random sets of input parameters, we again calculated CCS_{PARCS} with ROSIE-PARCS and CLI-PARCS for all proteins in the CCS_{IM} dataset. As shown in Supplementary Figure 1, the normalized CCS_{PARCS} calculated using ROSIE-PARCS and CLI-PARCS remained virtually identical. The average percent difference (normalized CCS_{PARCS} from ROSIE-PARCS compared to that of CLI-PARCS) across the six random sets of input parameters was 0.02%. Our results from the

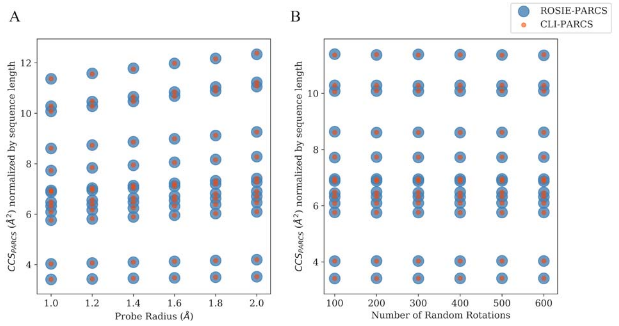


Figure 4. ROSIE-PARCS (blue) and CLI-PARCS (orange) produce virtually identical CCS_{PARCS} predictions at various parameter settings. (**A**) CCS_{PARCS} (normalized by sequence length) for both ROSIE-PARCS and CLI-PARCS at probe radii between 1.0 and 2.0 Å and at a fixed NRRs of 300. (**B**) CCS_{PARCS} (normalized by sequence length) for both ROSIE-PARCS and CLI-PARCS using 100–600 random rotations (at a fixed probe radius of 1.0 Å).

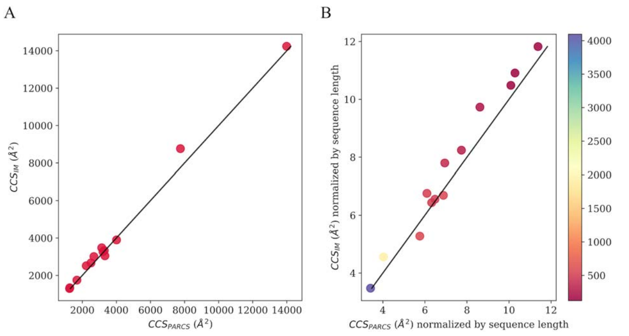


Figure 5. Comparison of CCS_{PARCS} from ROSIE-PARCS to CCS_{IM}. (**A**) The CCS values without normalization are shown. (**B**) CCS values normalized by the sequence length for all proteins in the CCS_{IM} dataset are shown.

three different verification tests indicated that calculations from ROSIE-PARCS are just as consistent and reliable when compared to those of CLI-PARCS.

Case study 1: Experimental CCS_{IM} closely agrees with CCS_{PARCS} calculated by ROSIE-PARCS for known structures

This case study provides an example of basic CCS_{PARCS} calculation with ROSIE-PARCS. In this study, we calculated the CCS_{PARCS} (with ROSIE-PARCS) for all proteins in the CCS_{IM} dataset (13 experimentally determined protein structures deposited in the PDB). We then compared the CCS_{PARCS} with the CCS_{IM}. As shown in Figure 5A, a strong correlation ($R^2 = 0.994$) and an average percent error of 6.12% were observed between CCS_{PARCS} and CCS_{IM} for the proteins in the CCS_{IM} dataset. The CCS_{IM} dataset consisted of proteins of varying sequence length. Therefore, to better compare CCS_{PARCS} with CCS_{IM}, we normalized both CCS_{PARCS} and CCS_{IM} by their respective sequence length as shown in Figure 5B. We still observed a strong correlation ($R^2 = 0.975$).

Case study 2: Inaccurate models predicted with AF2 generally correspond to high predicted ΔCCS values

The recent advances in deep learning and the development of AF2 have led to highly successful prediction of protein tertiary and quaternary structures from amino acid sequence. However, there are still cases where AF2 predicts inaccurate models yet assigns those high confidence scores. To test the ability of ROSIE-PARCS to assess AF2 structures, we predicted the structures of six proteins (a total of 30 predicted structures) shown in Supplementary Table 2. We than calculated the RMSD of the AF2 models to their known structure and obtained the average AF2 pLDDT (AF2 confidence metric) for all AF2 predictions as shown in Figure 6A. In AF2, a high pLDDT signifies higher confidence. Therefore, we explored whether there was a correlation between the average pLDDT and the RMSD. However, we show in Figure 6A that the observed correlation was not very strong, with several notable outliers. We hypothesize that this is because pLDDT is a local measure. Additionally, we used PARCS to obtain the CCS_{PARCS}. We then calculated the absolute difference of the CCS_{PARCS} of the structures predicted with AF2 to the CCS_{IM} (ΔCCS). We also calculated the RMSD of the predicted protein structures to their native structures. A comparison of ΔCCS versus RMSD is shown in Figure 6B. In the context of protein structure prediction, a high ΔCCS indicates high disagreement with experimental data and vice versa [46]. And indeed, we observed a much stronger correlation when comparing the ΔCCS to the RMSD, as shown in Figure 6B. This suggests that, at least for our subset of six proteins, ΔCCS was more effective in assessing predictions with different RMSD values when compared to pLDDT (Figure 6). In this dataset, ΔCCS using PARCS was a better measure of confidence than average pLDDT. Furthermore, we also explored the correlation between normalized ΔCCS and RMSD by accounting for protein size. To do this, we normalized the ΔCCS by the amino acid sequence length (see Supplementary Methods for additional details). Additionally, we normalized the corresponding RMSD values to $RMSD_{100}$ [74] values such that every protein was projected to have the same size. As expected, ΔCCS was sensitive to the protein size information. Although small deviations from the specific trend seen in Figure 6B were observed upon normalization, the overall correlation pattern remained consistent, as shown in Supplementary Figure 2. Thus, in cases where experimental IM CCS data are available, ΔCCS (as calculated by ROSIE-PARCS) might serve as a metric for evaluating the accuracy of predicted protein structures.

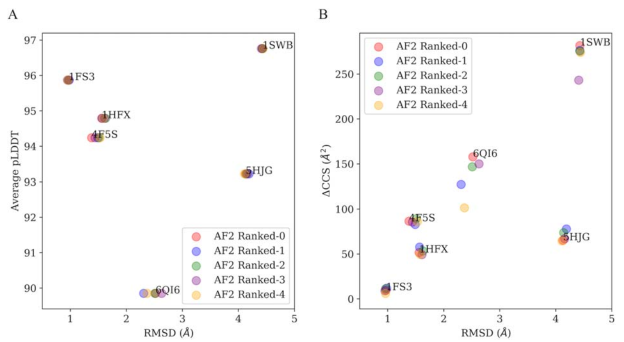


Figure 6. Analysis of average pLDDT, ΔCCS and RMSD for all predicted structures from AF2. Comparison of (**A**) the AF2 confidence metric (average pLDDT) and (**B**) the absolute difference between CCS_{IM} and CCS_{PARCS} (ΔCCS) against the RMSD of the predicted structures.

Conclusion

IM is an experimental technique to investigate the structures of proteins. Specifically, IM coupled to native MS techniques can provide structural information about the shape and size of proteins. Computational modeling to predict IM CCS data further enhances the understanding of protein structures in their native state. Additionally, integrative modeling with IM data has garnered significant attention and interest from the broader structural biology community. We have developed the ROSIE-PARCS web interface which allows prediction of CCS from protein structure. It can be accessed via this link (https://rosie.graylab.jhu.edu). The ROSIE-PARCS interface is intuitive, easy and free to use. Additionally, our case studies indicated that ROSIE-PARCS accurately predicts CCS (within the limitations of the PA) and can aid in distinguishing accurate and inaccurate models of protein structures predicted with AlphaFold2. We believe ROSIE-PARCS can help researchers gain a more comprehensive understanding of protein structures and has the potential to be applied in a variety of IM applications. Furthermore, IM-MS has applicability not only to proteins but can also be used, among other things, to measure the CCS values of polymers, nucleic acids and protein drug complexes. Given this versatility, in future work, we aim to expand our web server application to be able to calculate CCS for a broader range of systems. To further cater to the diverse needs of our users, in future versions of our web server application, we aim to expand ROSIE-PARCS functionality by incorporating support for a wider range of molecular file formats, in addition to the PDB file format.

Key Points

- IM-MS provides shape and size information of protein structure through collision cross section (CCS_{IM}), but existing computational methods for predicting CCS_{IM} are not user-friendly.
- ROSIE-PARCS is a new web application on the ROSIE webserver and allows users to predict CCS_{IM} in a user-friendly manner.
- Case study 1 shows that CCS_{IM} predictions using ROSIE-PARCS are highly accurate with an average error of only 6.1%, while case study 2 shows that CCS_{IM} predictions can help distinguish accurate from inaccurate AlphaFold2 protein structure predictions.
- ROSIE-PARCS is publicly and freely available at https://rosie.graylab.jhu.edu

Acknowledgements

We thank the members of the Lindert lab and Ameya Harmalkar from the Gray Lab at Johns Hopkins University for many useful discussions. We would like to thank the Ohio Supercomputer Center [78] for their valuable computational resources as well as their computational support team including Summer Wang, Zhiqiang You and Samuel Khuvis along with many others for their help with issues related to specific software installations and usage of the supercomputer. We also thank the Wysocki Lab at Ohio State University for their valuable feedback on the ROSIE-PARCS web interface. The ROSIE framework is supported and maintained by the Rosetta Commons.

Funding

This work was supported by National Institutes of Health (P41 GM128577) and a Sloan Research Fellowship to S.L.

Supplementary Data

Supplementary data are available online at https://academic.oup.com/bib.

Author Contributions

SM.B.A.T. performed the simulations, data collection, the analyses of the data, code development and the preparation of the manuscript along with its supplementary information. J.T.S. and St. L. contributed to the development of this manuscript as well as supervising this project. Se. L. provided guidance in code development as well as in the preparation of the manuscript.

Code and data availability

The ROSIE-PARCS web interface is free for academic use and can be accessed via this link (https://rosie.graylab.jhu.edu). In order to access any applications on ROSIE (including PARCS), users will need a GitHub account. All python scripts and data related to the manuscript and its SI are available as part of a supplementary zip file and can be readily used to recreate all figures and tables shown in the manuscript and its SI. Example scripts to run AF2, CLI-PARCS and RRD are available upon request. The source code related to creating the ROSIE-PARCS web interface is available upon request.

REFERENCES

1. Seffernick JT, Lindert S. Hybrid methods for combined experimental and computational determination of protein structure. *J Chem Phys* 2020;**153**:240901.
2. Wyttenbach T, Bowers MT. 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* 2011;**115**:12266–75.
3. Ruotolo BT, Robinson CV. Aspects of native proteins are retained in vacuum. *Curr Opin Chem Biol* 2006;**10**:402–8.
4. Bleiholder C, Liu FC. Structure relaxation approximation (SRA) for elucidation of protein structures from ion mobility measurements. *J Phys Chem B* 2019;**123**:2756–69.
5. Leney AC, Heck AJR. Native mass spectrometry: what is in the name? *J Am Soc Mass Spectrom* 2017;**28**:5–13.
6. Matthew Allen Bullock J, Schwab J, Thalassinos K, *et al.* The importance of non-accessible crosslinks and solvent accessible surface distance in Modeling proteins with restraints from crosslinking mass spectrometry. *Mol Cell Proteomics* 2016;**15**: 2491–500.
7. Mendoza VL, Vachet RW. Probing protein structure by amino acid-specific covalent labeling and mass spectrometry. *Mass Spectrom Rev* 2009;**28**:785–815.
8. Zhou M, Wysocki VH. Surface induced dissociation: dissecting noncovalent protein complexes in the gas phase. *Acc Chem Res* 2014;**47**:1010–8.
9. Medzihradszky KF, Burlingame AL. The advantages and versatility of a high-energy collision-induced dissociation-based strategy for the sequence and structural determination of proteins. *Methods* 1994;**6**:284–303.
10. Dixit SM, Polasky DA, Ruotolo BT. Collision induced unfolding of isolated proteins in the gas phase: past, present, and future. *Curr Opin Chem Biol* 2018;**42**:93–100.
11. Hart-Smith G. A review of electron-capture and electron-transfer dissociation tandem mass spectrometry in polymer chemistry. *Anal Chim Acta* 2014;**808**:44–55.
12. Brodbelt JS, Morrison LJ, Santos I. Ultraviolet photodissociation mass spectrometry for analysis of biological molecules. *Chem Rev* 2020;**120**:3328–80.

13. Roberts VA, Pique ME, Hsu S, Li S. Combining H/D exchange mass spectrometry and computational docking to derive the structure of protein-protein complexes. *Biochemistry* 2017;**56**:6329–42.
14. Jurneczko E, Barran PE. 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* 2011;**136**:20–8.
15. Terwilliger TC, Poon BK, Afonine PV, *et al.* Improved AlphaFold modeling with implicit experimental information. *Nat Methods* 2022;**19**:1376–82.
16. Alber F, Forster F, Korkin D, *et al.* Integrating diverse data for structure determination of macromolecular assemblies. *Annu Rev Biochem* 2008;**77**:443–77.
17. Saltzberg DJ, Viswanath S, Echeverria I, *et al.* Using integrative Modeling platform to compute, validate, and archive a model of a protein complex structure. *Protein Sci* 2021;**30**:250–61.
18. Leman JK, Weitzner BD, Lewis SM, *et al.* Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nat Methods* 2020;**17**:665–80.
19. Marzolf DR, Seffernick JT, Lindert S. Protein structure prediction from NMR hydrogen-deuterium exchange data. *J Chem Theory Comput* 2021;**17**:2619–29.
20. Nguyen TT, Marzolf DR, Seffernick JT, *et al.* Protein structure prediction using residue-resolved protection factors from hydrogen-deuterium exchange NMR. *Structure* 2022;**30**:313–320.e3.
21. Leelananda SP, Lindert S. Iterative molecular dynamics-Rosetta membrane protein structure refinement guided by Cryo-EM densities. *J Chem Theory Comput* 2017;**13**:5131–45.
22. Leelananda SP, Lindert S. Using NMR chemical shifts and Cryo-EM density restraints in iterative Rosetta-MD protein structure refinement. *J Chem Inf Model* 2020;**60**:2522–32.
23. McCafferty CL, Papoulas O, Jordan MA, *et al.* Integrative modeling reveals the molecular architecture of the intraflagellar transport a (IFT-A) complex. *Elife* 2022;**11**:11.
24. Rajabi K, Ashcroft AE, Radford SE. Mass spectrometric methods to analyze the structural organization of macromolecular complexes. *Methods* 2015;**89**:13–21.
25. Seffernick JT, Harvey SR, Wysocki VH, Lindert S. Predicting protein complex structure from surface-induced dissociation mass spectrometry data. *ACS Cent Sci* 2019;**5**:1330–41.
26. Biehn SE, Lindert S. Accurate protein structure prediction with hydroxyl radical protein footprinting data. *Nat Commun* 2021;**12**:341.
27. Aprahamian ML, Chea EE, Jones LM, Lindert S. Rosetta protein structure prediction from hydroxyl radical protein footprinting mass spectrometry data. *Anal Chem* 2018;**90**:7721–9.
28. Aprahamian ML, Lindert S. Utility of covalent Labeling mass spectrometry data in protein structure prediction with Rosetta. *J Chem Theory Comput* 2019;**15**:3410–24.
29. Hall Z, Politis A, Robinson CV. Structural modeling of heteromeric protein complexes from disassembly pathways and ion mobility-mass spectrometry. *Structure* 2012;**20**: 1596–609.
30. Hauri S, Khakzad H, Happonen L, *et al.* Rapid determination of quaternary protein structures in complex biological samples. *Nat Commun* 2019;**10**:192.
31. Saltzberg DJ, Broughton HB, Pellarin R, *et al.* A residue-resolved Bayesian approach to quantitative interpretation of hydrogen-deuterium exchange from mass spectrometry: application to characterizing protein-ligand interactions. *J Phys Chem B* 2017;**121**:3493–501.
32. Zhang MM, Beno BR, Huang RY, *et al.* An integrated approach for determining a protein-protein binding interface in solution and an evaluation of hydrogen-deuterium exchange kinetics for adjudicating candidate docking models. *Anal Chem* 2019;**91**: 15709–17.
33. Xie B, Sood A, Woods RJ, Sharp JS. Quantitative protein topography measurements by high resolution hydroxyl radical protein footprinting enable accurate molecular model selection. *Sci Rep* 2017;**7**:4552.
34. Politis A, Park AY, Hall Z, *et al.* Integrative modelling coupled with ion mobility mass spectrometry reveals structural features of the clamp loader in complex with single-stranded DNA binding protein. *J Mol Biol* 2013;**425**:4790–801.
35. Eschweiler JD, Frank AT, Ruotolo BT. Coming to grips with ambiguity: ion mobility-mass spectrometry for protein quaternary structure assignment. *J Am Soc Mass Spectrom* 2017;**28**:1991–2000.
36. Eschweiler JD, Farrugia MA, Dixit SM, *et al.* A structural model of the urease activation complex derived from ion mobility-mass spectrometry and integrative modeling. *Structure* 2018;**26**:599–606.e3.
37. Harvey SR, Seffernick JT, Quintyn RS, *et al.* Relative interfacial cleavage energetics of protein complexes revealed by surface collisions. *Proc Natl Acad Sci U S A* 2019;**116**:8143–8.
38. Wang H, Eschweiler J, Cui W, *et al.* Native mass spectrometry, ion mobility, electron-capture dissociation, and Modeling provide structural information for gas-phase apolipoprotein E oligomers. *J Am Soc Mass Spectrom* 2019;**30**:876–85.
39. Drake ZC, Seffernick JT, Lindert S. Protein complex prediction using Rosetta, AlphaFold, and mass spectrometry covalent labeling. *Nat Commun* 2022;**13**:7846.
40. Seffernick JT, Turzo SMBA, Harvey SR, *et al.* Simulation of energy-resolved mass spectrometry distributions from surface-induced dissociation. *Anal Chem* 2022;**94**:10506–14.
41. Biehn SE, Picarello DM, Pan X, *et al.* Accounting for Neighboring residue hydrophobicity in diethylpyrocarbonate labeling mass spectrometry improves Rosetta protein structure prediction. *J Am Soc Mass Spectrom* 2022;**33**:584–91.
42. Biehn SE, Lindert S. Protein structure prediction with mass spectrometry data. *Annu Rev Phys Chem* 2022;**73**:1–19.
43. Biehn SE, Limpikirati P, Vachet RW, Lindert S. Utilization of hydrophobic microenvironment sensitivity in diethylpyrocarbonate labeling for protein structure prediction. *Anal Chem* 2021;**93**:8188–95.
44. Seffernick JT, Canfield SM, Harvey SR, *et al.* Prediction of protein complex structure using surface-induced dissociation and cryo-electron microscopy. *Anal Chem* 2021;**93**:7596–605.
45. Landreh M, Sahin C, Gault J, *et al.* Predicting the shapes of protein complexes through collision cross section measurements and database searches. *Anal Chem* 2020;**92**:12297–303.
46. Turzo SMBA, Seffernick JT, Rolland AD, *et al.* Protein shape sampled by ion mobility mass spectrometry consistently improves protein structure prediction. *Nat Commun* 2022;**13**:4377.
47. Kahraman A, Herzog F, Leitner A, *et al.* Cross-link guided molecular modeling with ROSETTA. *PLoS One* 2013;**8**:e73411–1.
48. Ruotolo BT, Benesch JL, Sandercock AM, *et al.* Ion mobility-mass spectrometry analysis of large protein complexes. *Nat Protoc* 2008;**3**:1139–52.
49. Graves DB. In: Mason EA, McDaniel EW (eds). *Transport Properties of Ions in Gases: Kinetic Theory of Mobility and Diffusion.* New York: John Wiley and Sons, 1988, 560 + xvi pp, AIChE Journal 1989;**35**:701-701.
50. Mason EA, Schamp HW. Mobility of gaseous lons in weak electric fields. *Ann Phys Rehabil Med* 1958;**4**:233–70.
51. Morris CB, Poland JC, May JC, *et al.* Fundamentals of ion mobility-mass spectrometry for the analysis of biomolecules. *Methods Mol Biol* 2020;**2084**:1–31.

52. Marklund Erik G, Degiacomi Matteo T, Robinson Carol V, *et al.* Collision cross sections for structural proteomics. *Structure* 2015;**23**:791–9.
53. Shvartsburg AA, Jarrold MF. An exact hard-spheres scattering model for the mobilities of polyatomic ions. *Chemical Physics Letters* 1996;**261**:86–91.
54. Mesleh MF, Hunter JM, Shvartsburg AA, *et al.* Structural information from ion mobility measurements: effects of the long-range potential. *J Phys Chem* 1996;**100**:16082–6.
55. Ewing SA, Donor MT, Wilson JW, Prell JS. Collidoscope: an improved tool for computing collisional cross-sections with the trajectory method. *J Am Soc Mass Spectrom* 2017;**28**:587–96.
56. Larriba C, Hogan CJ. Free molecular collision cross section calculation methods for nanoparticles and complex ions with energy accommodation. *J Comput Phys* 2013;**251**:344–63.
57. Bleiholder C, Wyttenbach T, Bowers MT. A novel projection approximation algorithm for the fast and accurate computation of molecular collision cross sections (IV). Applications to polypeptides. *Int J Mass Spectrom* 2011;**308**:1–10.
58. Bleiholder C. A local collision probability approximation for predicting momentum transfer cross sections. *Analyst* 2015;**140**: 6804–13.
59. Mack E. Average cross-sectional areas of molecules by gaseous diffusion methods. *J Am Chem Soc* 1925;**47**:2468–82.
60. Prell JS. Chapter one – modelling collisional cross sections. In: Donald WA, Prell JS (eds). *Comprehensive Analytical Chemistry*. Amsterdam, Netherlands: Elsevier, 2019, 1–22.
61. Degiacomi MT. On the effect of sphere-overlap on super coarse-grained models of protein assemblies. *J Am Soc Mass Spectrom* 2019;**30**:113–7.
62. Webb B, Viswanath S, Bonomi M, *et al.* Integrative structure modeling with the integrative modeling platform. *Protein Sci* 2018;**27**:245–58.
63. Kaldmäe M, Sahin C, Saluri M, *et al.* 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**: 1024–30.
64. Leaver-Fay A, Tyka M, Lewis SM, *et al.* ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol* 2011;**487**:545–74.
65. Raman S, Vernon R, Thompson J, *et al.* Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins* 2009;**77**(Suppl 9):89–99.
66. Lyskov S, Chou FC, Conchúir S, *et al.* Serverification of molecular modeling applications: the Rosetta online server that includes everyone (ROSIE). *PLoS One* 2013;**8**:e63906.
67. Berman HM, Battistuz T, Bhat TN, *et al.* The protein data Bank. *Acta Crystallographica Section D* 2002;**58**:899–907.
68. Jumper J, Evans R, Pritzel A, *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**:583–9.
69. Stiving AQ, Jones BJ, Ujma J, *et al.* Collision cross sections of charge-reduced proteins and protein complexes: a database for CCS calibration. *Anal Chem* 2020;**92**:4475–83.
70. Allen SJ, Giles K, Gilbert T, Bush MF. Ion mobility mass spectrometry of peptide, protein, and protein complex ions using a radio-frequency confining drift cell. *Analyst* 2016;**141**:884–91.
71. Kim SS, Seffernick JT, Lindert S. Accurately predicting disordered regions of proteins using Rosetta ResidueDisorder application. *J Phys Chem B* 2018;**122**:3920–30.
72. Seffernick JT, Ren H, Kim SS, Lindert S. Measuring intrinsic disorder and tracking conformational transitions using Rosetta ResidueDisorder. *J Phys Chem B* 2019;**123**:7103–12.
73. He J, Turzo SBA, Seffernick JT, *et al.* Prediction of intrinsic disorder using Rosetta ResidueDisorder and AlphaFold2. *J Phys Chem B* 2022;**126**:8439–46.
74. Carugo O, Pongor S. A normalized root-mean-square distance for comparing protein three-dimensional structures. *Protein Sci* 2001;**10**:1470–3.
75. *The PyMOL Molecular Graphics System, Version* 2.5.2. Schrödinger, LLC.
76. Gabelica V, Marklund E. Fundamentals of ion mobility spectrometry. *Curr Opin Chem Biol* 2018;**42**:51–9.
77. Goodsell DS, Autin L, Olson AJ. Illustrate: software for biomolecular illustration. *Structure* 2019;**27**:1716–1720.e1.
78. Ohio Supercomputer Center. Ohio Supercomputer Center. 1987.