




Protein shape sampled by ion mobility mass spectrometry consistently improves protein structure prediction

Received: 27 July 2021

Accepted: 14 July 2022

Published online: 28 July 2022

 Check for updates

SM Bargeen Alam Turzo¹, Justin T. Seffernick¹, Amber D. Rolland ², Micah T. Donor², Sten Heinze¹, James S. Prell², Vicki H. Wysocki ¹ & Steffen Lindert ¹ ✉

Ion mobility (IM) mass spectrometry provides structural information about protein shape and size in the form of an orientationally-averaged collision cross-section (CCS_{IM}). While IM data have been used with various computational methods, they have not yet been utilized to predict monomeric protein structure from sequence. Here, we show that IM data can significantly improve protein structure determination using the modelling suite Rosetta. We develop the Rosetta Projection Approximation using Rough Circular Shapes (PARCS) algorithm that allows for fast and accurate prediction of CCS_{IM} from structure. Following successful testing of the PARCS algorithm, we use an integrative modelling approach to utilize IM data for protein structure prediction. Additionally, we propose a confidence metric that identifies near native models in the absence of a known structure. The results of this study demonstrate the ability of IM data to consistently improve protein structure prediction.

Proteins are at the core of virtually all cellular processes. Therefore, comprehensive knowledge of protein structures with atomistic detail can be beneficial for several pharmaceutical applications such as vaccine design¹, drug discovery^{2,3}, enzyme design⁴, self-assembling molecular machines⁵, and many more⁶. Mass spectrometry (MS) has become a prominent technique in the field of structural biology due to its ability to provide structural information for proteins and protein complexes. MS can be particularly beneficial because it is faster, can work for heterogeneous samples, can be used routinely at all stages of a project, and has fewer sample preparation complications compared to high-resolution techniques such as X-ray crystallography and cryo-electron microscopy (cryo-EM), and fewer concentration and size limitations compared to nuclear magnetic resonance (NMR) spectroscopy. Several findings for protein structures in the gas phase also suggest that features such as elements of secondary structure, compactness, and quaternary structure can be preserved during the transition from solution to desolvated state^{7–9}. For these reasons, structural

MS can be very beneficial particularly when high-resolution methods are not feasible^{10,11}. Various methods have been developed and coupled to MS to study protein structures^{12,13} including in-solution approaches such as chemical crosslinking¹⁴, covalent labeling¹⁵ and hydrogen-deuterium exchange¹⁶, and gas-phase approaches such as collision-induced dissociation, electron capture/transfer dissociation, ultraviolet photodissociation, surface-induced dissociation¹⁷ and ion mobility (IM) spectrometry¹⁸. While such MS techniques may provide diverse details and routine analysis of structures, experimental data collected from experiments are sparse and cannot unambiguously determine atomic-resolution structure¹⁹.

An alternative approach to experimental structure determination is to use computational modelling methods. These approaches, such as structure prediction from sequence or protein-protein docking, can also provide insight into atomistic details of biomolecules but are frequently limited in accuracy due to the large conformational sampling space among other challenges²⁰. While these methods can be

¹Department of Chemistry and Biochemistry and Resource for Native Mass Spectrometry Guided Structural Biology, Ohio State University, Columbus, OH 43210, USA. ²Department of Chemistry and Biochemistry and Materials Science Institute, University of Oregon, Eugene, OR 97403, USA.

✉ e-mail: lindert.1@osu.edu

successfully utilized in the absence of experimental data, sparse experimental data are often used to guide and improve modelling^{19,21,22}. Experimental data from various MS techniques have recently proved pivotal within integrative structural biology frameworks^{14,16,23–40}.

In IM, ions are transferred into an inert gas chamber at a constant pressure and temperature under the influence of a weak electric field^{41,42}. This technique is regularly utilized to separate protein structures based on their shape and size. IM can also provide a rotationally averaged collision cross section (CCS_{IM}) of the protein which is related to the amount of momentum exchanged between ion and buffer gas over the course of the collisions and can be thought of as somewhat like rotationally averaged cross sectional area⁴³. Several methods have been developed to predict CCS from protein structure. Among these, the most physically realistic algorithms are the trajectory method (TJM)^{44,45} and diffuse trajectory method (DTM)⁴⁶ which integrate Newton's equation of motion to calculate the classical scattering of gas particles. Both TJM and DTM explicitly account for long-range interactions through Lennard-Jones potentials to approximate momentum transfer from each gas particle to the collided ion. CCS obtained from these methods is very accurate⁴⁵, but these calculations can be slow. Due to the high computational cost, prediction methods such as elastic hard sphere scattering⁴⁷, projection superposition approximation (PSA)⁴⁸, local collision probability approximation⁴⁹, and projection approximation (PA)⁴³ make further approximations on TJM, resulting in faster CCS calculations. Among these approximated methods, PA is the simplest and fastest, because it neglects the scattering and long-range interactions^{43,50}. CCS_{PA} only accounts for the collisions of a gas particle with the ion based on hard sphere atomic radii by calculating the average cross-sectional area of the protein structure as experienced by the buffer gas. Using the CCS projection approximation calculation tool IMPACT, calculations are about 10^6 times faster⁴³ than the most rigorous methods and are widely used for comparison with experimental IM data. Therefore, PA approaches are advantageous for use in integrative modelling, where the CCS calculation is required for thousands of structures that are obtained from Monte Carlo sampling.

Several instances of structural modelling in conjunction with IM data have been reported. IM spectra have been successfully predicted with the structure relaxation approximation (SRA) method⁹. This method uses molecular dynamics simulations to model structures in the specific charge states. It then utilizes CCS_{PSA} of the generated structures to predict an overall IM spectrum. The SRA method indicated that systems studied with IM methods are generally consistent with retention of many residue-residue contacts determined by X-ray crystallography. Several studies suggest that during native IM experiments, globular proteins undergo minimal compaction and structural rearrangement upon transfer to the gas phase using nanoelectrospray ionization, assuming they are appropriately charged (with lower charge states often exhibiting patterns consistent with solution structures) and that instrument conditions are kept “soft” to avoid unintentional activation^{51–53}. Thus, the proteins largely retain native-like secondary, tertiary, and quaternary structures during the time-scale of native IM experiments^{9,40,54,55}. Therefore, data from such gas phase studies likely is beneficial for modelling solution phase structures though a rapid, validated way to evaluate plausible candidate structures consistent with IM-MS data would be beneficial to interpreting these results. Furthermore, IM data have been incorporated in computational modelling for protein complex structure prediction. In these methods, coarse-grained models generated using the Integrative Modelling Platform⁵⁶ were ranked and clustered based on the agreement between predicted and experimental CCS measured from IM^{28,35}. CCS_{IM} values for complexes and their individual subunits have also been successfully used to approximate rough intersubunit distance used as restraints in modelling methods to identify coarse-grained

topologies of complexes^{36–39}. Studies have also revealed that shapes and architectures of protein complexes can be determined from CCS_{IM} measurements and database searches^{57,58}. In addition to complex structure prediction, work has also been done to show correlation between IM data and structural similarity (RMSD)⁹. While several studies have demonstrated that IM data can be predicted and utilized with various computational methods, IM data have not yet been utilized to predict monomeric structure from sequence.

Therefore, in this work a non-stochastic grid-based algorithm, PARCS, has been implemented in Rosetta^{59,60} to predict CCS from structure. It has been demonstrated that PARCS yields comparable results to IMPACT in terms of speed and accuracy. Next an IM score term has been developed for use in the *ab initio*^{61–63} and comparative modelling (CM)⁶⁴ protocols in Rosetta, in combination with the Rosetta all-atom scoring function⁶⁵. This score term scored structures based on their (dis)agreement with experimental IM data. When this score term was included, the prediction of structures improved for a benchmark of 25 proteins: the RMSD improved by an average of 2.0 Å and 17/25 structures were predicted accurately.

Results

In this study, to utilize IM data to predict tertiary (monomeric) structures in Rosetta, an algorithm designed for rapid prediction of CCS has been developed and implemented. This method uses Projection Approximation (Eq. 1) via a grid-based calculation of Rough Circular Shapes (PARCS). Subsequently, a score function was developed (Eq. 2 and Eq. 3) that assessed the agreement of Rosetta-generated models to the CCS_{IM} for tertiary structure prediction.

CCS Calculations by PARCS are fast, accurate and comparable to existing software

Area calculation in projection approximation methods is typically performed using Monte Carlo integration methods. In such an approach, probes representing the buffer gas particle are fired upon the randomly oriented 2D-projected target structure to calculate the area of the projection. A large number of probes is usually required for CCS calculations to converge. However, when a large number of probes is used, random probes frequently survey areas with no protein present, resulting in unnecessary calculations and thus adding to the computational cost⁶⁶. Therefore, run-to-run variability in probe-based projection area calculation per rotation is common. To circumvent this issue, in PARCS, the projection area is calculated by projecting the structure on a 2D grid and then geometrically estimating the projection area directly (by geometrically filling the grid based on locations of atoms and radii of atoms and probes). This approach eliminates the variability in projection area calculation. Therefore, the only attribute contributing to the variability in CCS calculations using PARCS is the random rotations (Eq. 1).

To benchmark our PARCS algorithm, CCS values for 4465 non-homologous protein structures in the PARCS evaluation dataset were calculated. Results for convergence of CCS calculations at varying number of random rotations on the PARCS evaluation dataset are shown in Fig. 1a. The average standard deviation of the CCS distributions for 100 rotations was only 2.26 Å² (which was less than 0.2% of the CCS_{PARCS} on average) and decreased as the number of rotations increased. The average of the standard deviations of the CCS distributions was well below 2.0 Å² for more than 100 rotations as shown in Fig. 1a. For CCS_{PARCS} , the default number of rotations was set to 300, where the average standard deviation of the distribution was 1.31 Å².

For all proteins in the PARCS evaluation dataset, CCS calculated by PARCS was compared to CCS calculated by IMPACT, one of the most widely used CCS calculation methods, as shown in Fig. 1b. A strong correlation ($R^2 = 0.9996$) was observed between CCS_{PARCS} and

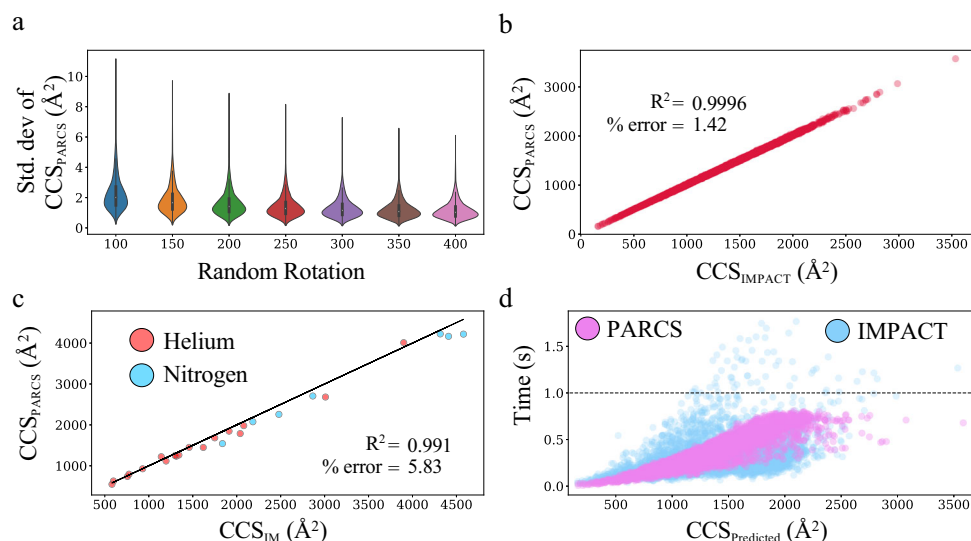


Fig. 1 | Analysis of projection area using rough circular shapes (PARCS) algorithm. **a** Convergence of collision cross section (CCS) calculation using PARCS was tested on, $n = 4465$, biologically independent samples (crystal structures deposited in the Protein Data Bank [PDB]) over 100 independent runs for 7 separate categories of random rotations ranging from 100R400 random rotations. The mean and the standard error of the mean for these distributions of 100R400 random rotations were $2.257 \pm 0.016 \text{ \AA}^2$, $1.852 \pm 0.013 \text{ \AA}^2$, $1.598 \pm 0.011 \text{ \AA}^2$, $1.415 \pm 0.010 \text{ \AA}^2$, $1.306 \pm 0.009 \text{ \AA}^2$, $1.222 \pm 0.009 \text{ \AA}^2$, $1.162 \pm 0.008 \text{ \AA}^2$ respectively. The white dots represent the median in each violin distribution. The black bar in the center of the distribution is the interquartile range (IQR). The black stretched line extends from

the “first quartile – 1.5 IQR” to the “third quartile + 1.5 IQR”. Values beyond this range are considered outliers. **b** Comparison of $\text{CCS}_{\text{PARCS}}$ to that of $\text{CCS}_{\text{IMPACT}}$ (of these $n = 4465$ biologically independent samples) exhibited a very strong correlation (R^2 of 0.9996). **c** A strong correlation (R^2 of 0.991) was observed for predicted $\text{CCS}_{\text{PARCS}}$ values of PARCS when compared with CCS_{IM} from nitrogen (blue) and helium (red) buffer gas for the experimental dataset ($n = 25$ biologically independent samples with ion mobility data). **d** Comparison of CCS calculation time of PARCS and IMPACT ($n = 4465$ biologically independent samples) showed that PARCS and IMPACT performed equally well in terms of speed. Source data are provided as a Source Data file.

$\text{CCS}_{\text{IMPACT}}$ with a root mean squared error (RMSE) of 21.03 \AA^2 and an average percent error of 1.42%. The results demonstrate that PARCS calculates CCS values as accurately as other projection approximation methods. $\text{CCS}_{\text{PARCS}}$ were then compared to CCS_{IM} for the experimental dataset. We observed a strong correlation ($R^2_{\text{PARCS}} = 0.991$) between $\text{CCS}_{\text{PARCS}}$ and CCS_{IM} values as shown in Fig. 1c, where IM data collected in nitrogen and helium buffer gas are shown in blue and red respectively. We observed an average percent error of only 5.83% (similar to that of IMPACT at 5.61%) from CCS_{IM} . To use IM data in computational structure prediction methods (where CCS prediction is required on a large number of decoy structures), the speed of CCS calculations should be within about a second. Therefore, calculation times of PARCS were compared to that of IMPACT as presented in Fig. 1d. Using the PARCS evaluation dataset, PARCS took an average of 0.40 seconds to calculate the CCS of proteins when 300 random rotations were used compared to 0.32 seconds for IMPACT. Thus, the timing of PARCS was comparable to IMPACT. We note that the slightly longer average time for PARCS was due to additional steps performed by Rosetta when reading in a PDB file (such as checks for correctness and adding missing hydrogens⁵⁹). For all 4465 proteins, calculations for PARCS completed in under 1.0 second as shown in Fig. 1d. These results indicate that PARCS in Rosetta offers similar speed and accuracy to established PA algorithms. We hypothesized that the information contained within CCS_{IM} may be sufficient to predict structures using $\text{CCS}_{\text{PARCS}}$.

PARCS in IM score function improves model selection in an ideal dataset

In this study we sought to investigate the usefulness of the structural information encoded in IM data for predicting the complete structure of single-subunit proteins. However, it was unclear whether a single CCS value, encoding overall size and shape, was sufficient to distinguish near-native from incorrect protein models. To test how useful the information in CCS was for structure prediction, an IM score

function (Eq. 2) has been developed to score structures based on the (dis)agreement with experimental IM data (Eq. 3). To assess the capabilities of this score function to adequately distinguish good from bad models, we first tested it on the ideal dataset (where the experimental CCS_{IM} was replaced with $\text{CCS}_{\text{PARCS}}$ of the native structure for a set of proteins representative of all unique architectures in the CATH database, $\text{CCS}_{\text{Ideal}}$). For each protein in the ideal dataset, 10,000 potential structures were generated (decoy structures), using the protocol outlined in Supplementary Methods (modelling protocol explicitly noted for each protein in Supplementary Data 1) and scored using the developed IM and RG score functions. Prediction results from the RG, RS, and IM score functions were evaluated and compared based on agreement with experimental structures (using the RMSD and TM-Score of the lowest scoring model, i.e., the predicted structure, Supplementary Data 2). We observed a significant improvement in model quality upon the inclusion of ideal IM data. The predicted structures with the IM score function were close to their native structures with an average RMSD of only 3.7 \AA . The average TM-Score of these predicted models was 0.86. The models predicted with the RG score function (a proxy score function that only favors compact models) had an average RMSD and TM-Score of 5.7 \AA and 0.80 respectively. As highlighted in Fig. 2a, the models predicted with the IM score function generally had lower RMSD (i) and higher TM-Score (ii) compared to those predicted by the RG score function. These results suggest that the two quantities (collision cross section and radius of gyration) do not provide the same structural information and that CCS-based scoring far outperforms modelling based solely on radius of gyration. Thus, RG cannot be used as a substitute for the IM score function. Structures predicted with the IM score function were then compared to those predicted with the RS score function (default Rosetta score function that did not utilize IM data). As shown in Fig. 2b (i), the RMSD of the predicted structures improved or remained unchanged for 58 out of 60 proteins, with an average RMSD improvement of 0.8 \AA . The TM-Score also improved or remained unchanged for 58 proteins as shown in Fig. 2b (ii). The

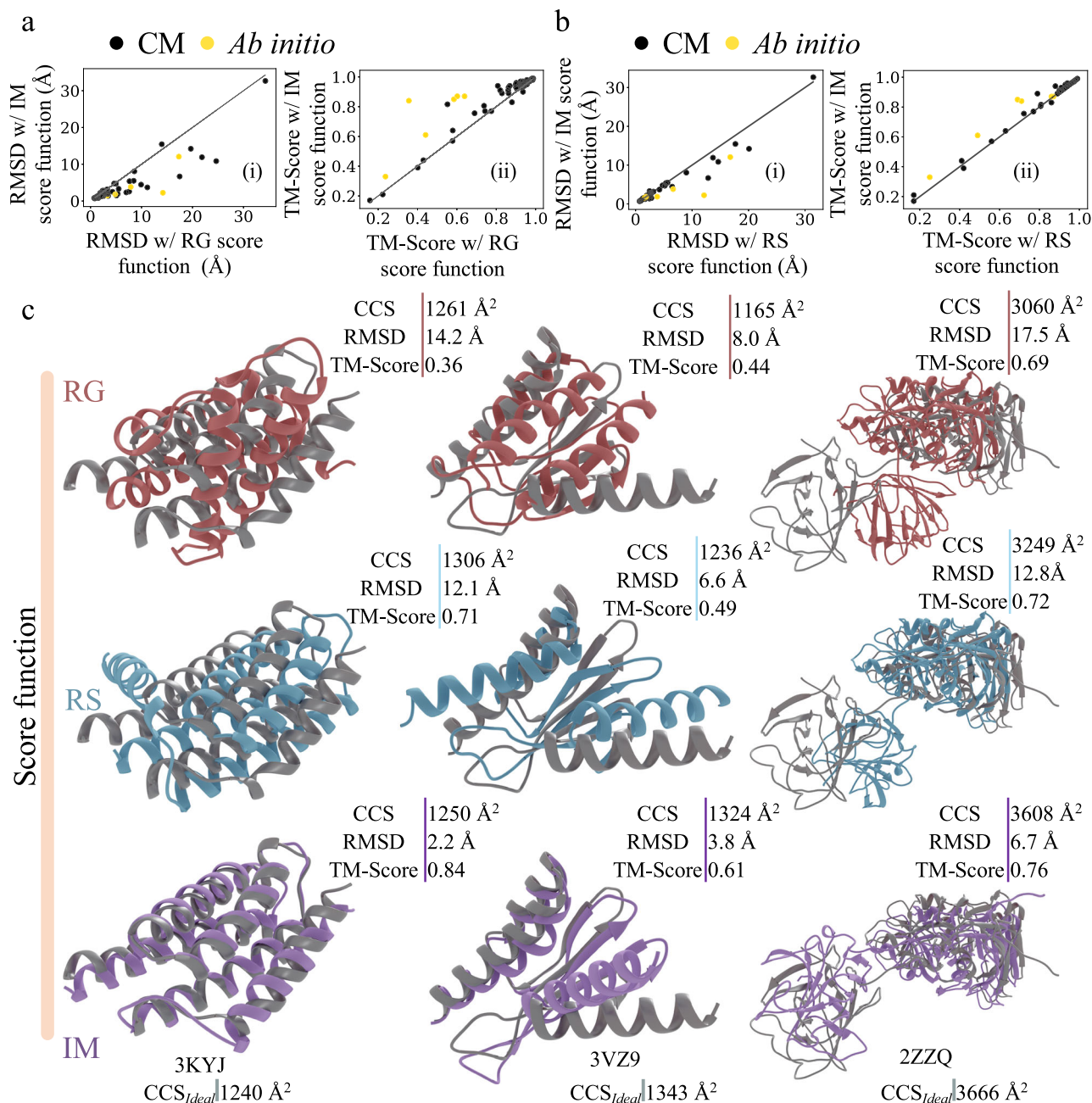


Fig. 2 | Protein structure prediction with and without ideal ion mobility (IM) data. Consistent improvement in model selection was observed when using the IM score function for the 60 proteins in the ideal dataset. The predicted models from the IM score function were compared to that of the a radius of gyration (RG) and b Rosetta (RS) score functions in terms of their respective (i) root mean square

deviation (RMSD) and (ii) template modelling score (TM-Score). For both a, b models predicted with comparative modelling (CM) and ab initio are shown with black and yellow circles respectively. c Comparison of predicted structures with RG (pink), RS (blue) and IM (purple) score function to their native structures (grey) for three members of the ideal dataset. Source data are provided as a Source Data file.

average TM-Score improvement over the RS score function was 0.02. Three of the predicted structures (that showed significant above-average improvement) from each the score function (RG, RS, and IM) are shown in Fig. 2c, where the native structure (grey) was compared to the predicted model of the RG (red), RS (blue), and IM (purple) score function.

The ability of the IM score function to predict structures as a function of sample size was also tested by varying the number of scored decoy structures from 100 to 10,000 for the ideal dataset. The average RMSD and TM-Score effectively improved for the IM score function as the number of predicted decoys increased (Supplementary Fig. 1a [i] and Supplementary Fig. 1a [ii] respectively). However, after

about 7000 structures this improvement was negligible, suggesting that users should generate at least 7000 decoys.

Along with the improvement in model selection, the IM score exhibited a more well-defined energy funnel when compared to the RG and RS score functions. We saw a 9.5-fold increase in P_{near} when we used the IM score function over RS. We also observed a 22.4-fold increase in P_{near} with the IM over RG score function. This suggests that inclusion of IM data significantly improved the goodness of the score versus RMSD and TM-Score funnel. Finally, to test the robustness of the IM score function in the presence of experimental uncertainty, noise was introduced to the ideal CCS data (as outlined in Supplementary Methods). These “noisy” ideal CCS data were then used with

the IM score function to predict structures. As shown in Supplementary Fig. 2a, the IM_{score} versus TM-Score distributions for all generated structures of the ideal dataset (600,000 structures) at 0% noise (blue, [i]), 15% noise (orange, [ii]), and 30% noise (green, [iii]) were largely identical. Furthermore, no significant change in the global fold was observed for the 60 best scoring models that were predicted with 15% and 30% random noise (Supplementary Fig. 2b). The average TM-Score of the predicted structures at 15% and 30% noise differed by only 0.01 and 0.02, respectively, when compared to that in the absence of noise for the ideal dataset. This suggests that the IM score function is not affected by reasonable amounts of random noise.

Due to the possibility of a gas-phase collapse, which would lead to a lower measured CCS than expected based on the protein crystal structures^{67–70}, we also used the IM score function to predict structures based on artificially reduced CCS_{ideal} values (for the ideal dataset) that were significantly lowered (up to 30%) compared to those found in the native structures (Supplementary Data 3). When these adjusted CCS values were used in scoring, the average radius of gyration (of the predicted models) decreased only slightly from 20.28 Å to 19.92 Å (as CCS_{ideal} decreased by 30%) while the average radius of gyration of the native structures was 20.62 Å. This analysis indicated that even significantly smaller than expected CCS values induced only minimal compaction in the global fold of the predicted structures compared to the native models. This compaction is expected, since the IM data yields information about the shape and size of proteins, therefore lowering the CCS_{ideal} will lead to prediction of models that are more compact. Several studies also indicate that minor gas-phase compaction could result due to the self-solvation of surface residue side chains^{55,71–74}. Therefore, to test whether the observed minor compaction was simply due to the exposed side chains self-solvating in the gas phase, we also analyzed the neighbor count (NC) of surface residues (as defined earlier) in the predicted models using the same decreasing CCS_{ideal} (Supplementary Data 3). A low NC generally corresponds to residues that are more solvent exposed and vice versa. Our results show that as CCS_{ideal} was lowered (as described above), the IM models tended to exhibit surface residues that were only slightly less solvent exposed on average, suggesting that the side chains of these predicted models were only slightly collapsed compared to those of the native structures. However, the average difference of the surface residue NC (from that of the native structures) was only 0.78 when surface residues were defined as the residues with the lowest 20% of neighbor counts. Furthermore, the corresponding average RMSD increased to no more than 4.70 Å and the TM-Score decreased to no less than 0.84 when the CCS was artificially lowered up to 30%. These findings indicated that a small amount of side chain collapse for surface residues is the primary reason for the increase in compactness. Additionally, these results were expected since Rosetta used an implicit solvation model during model generation. As a result, gas-phase like structures were not likely to be found within the ensemble. Moreover, the IM_{Score_Term} was determined using an upper bound (UB) and lower bound (LB) as shown in Eq. 3. If the ΔCCS (absolute difference between the predicted and experimental CCS) was higher than the UB, then the structure in question was given a constant maximum penalty (Eq. 3). Effectively, this caused the ranking of structures using IM score function to be the same as the RS score function.

Protein structures in the ideal dataset (60 proteins) were also predicted with AlphaFold2 (AF) and RoseTTAFold (RF). These structures were predicted both with and without templates (as outlined in Supplementary Methods). Furthermore, we benchmarked the prediction results of AF and RF both with and without templates for the ideal dataset as shown in Supplementary Fig. 3. For the ideal dataset, for both AF and RF, the average RMSD decreased by 0.2 Å and 0.33 Å respectively when templates were used. Similarly, the average TM-Score increased by 0.01 and 0.04 when templates were used for predicting the structures of proteins. In Fig. 3a, all three methods (IM,

AF_{with_templates}, and RF_{with_templates}) predicted better structures than RS (Supplementary Data 2). For the ideal dataset, the average RMSD of the predicted models from the IM score function was higher than those predicted with AF_{with_templates} by 0.3 Å. In contrast, the average RMSD of IM predicted models were lower by 0.6 Å when compared to those predicted with RF_{with_templates}. These results are highlighted in Fig. 3a (i). The average TM-Score of the IM predicted models were lower than AF_{with_templates} and RF_{with_templates} by 0.06 and 0.03 respectively as shown in Fig. 3a (ii). However, there were 28 cases where the IM predicted models that were better or the same when compared to AF_{with_templates} models. The average RMSD and TM-Score for this subset improved by 2.7 Å and 0.04, respectively. IM also predicted 37 models that were better or the same when compared to RF_{with_templates} models. For this subset, the average RMSD and TM-Score improved by 2.6 Å and 0.05 respectively. Figure 3b shows two such cases where the predicted structure using IM (purple) matches the native (grey) significantly more closely over the predicted AF_{with_templates} (cyan) and RF_{with_templates} (red) structures. Additionally, the average absolute percent error of CCS_{PARCS} to CCS_{ideal} of the top scoring models of RG, RS, AF_{with_templates}, RF_{with_templates}, and IM were 4.6%, 3.1%, 2.4%, 2.2% and 0.9% (Supplementary Data 4). This is expected because the IM score function predicts models that agree the most with IM data as compared to all other methods that do not use this information for structure prediction. Furthermore, large deviations in normalized ΔCCS (ΔCCS divided by sequence length) were also observed for poorly predicted AF_{with_templates} and RF_{with_templates} structures (TM-Score below 0.5) as shown in Fig. 3c. Similar results were also observed for poorly predicted AF_{without_templates} and RF_{without_templates} structures as shown in Supplementary Fig. 4 (and Supplementary Data 5). This suggested that in the future the IM score function could also be used to assess structures generated with AF and RF (both with and without templates). Given the sparseness of the data (CCS is a single number denoting the average cross-sectional area of the protein) these results indicated that the overall size and shape information contained in the IM data indeed had a strong potential to facilitate the discrimination of good from bad models. While an encouraging proof of principle, these results do not account for the uncertainty associated with real experimental IM data. An average percent error of 5.83% between CCS_{PARCS} and CCS_{IM} was observed for the experimental dataset (Fig. 1c). Thus, when we turn to experimental IM data for the structure prediction, additional uncertainty will be present. Therefore, we next present tests of the effectiveness of IM data to improve structure prediction that are based on use of a dataset with experimental IM data.

IM data improve model selection of protein structures in an experimental dataset

For proteins in the experimental dataset, 10,000 decoy models were generated with either the ab initio or comparative modelling (CM) protocols as specified in Supplementary Data 6. Each of these decoy models was scored with IM data (Eq. 2) and the predicted models were then compared to those scored with the RS and RG score functions (Supplementary Data 7). Again, we saw a notable improvement in model quality upon the inclusion of IM data. In Fig. 4a and b, (i) the RMSDs (and [ii] TM-Score) of the best scoring models with IM data are compared to those from the RG and RS score functions. As highlighted in Fig. 4a (i), the average RMSD of the predicted structures with the IM score function was 5.3 Å, while the average RMSD for those predicted with the RG score function was 10.6 Å. Similarly, the average TM-Score of the IM and RG predicted structures were 0.67 and 0.52 respectively (Fig. 4a [ii]). These results further established that model discrimination using actual experimental IM data significantly outperforms the simple proxy score function (RG) that only ensured protein compactness. Compared to the RS score function as shown in Fig. 4b (i), the RMSDs of the predicted models for proteins in the experimental

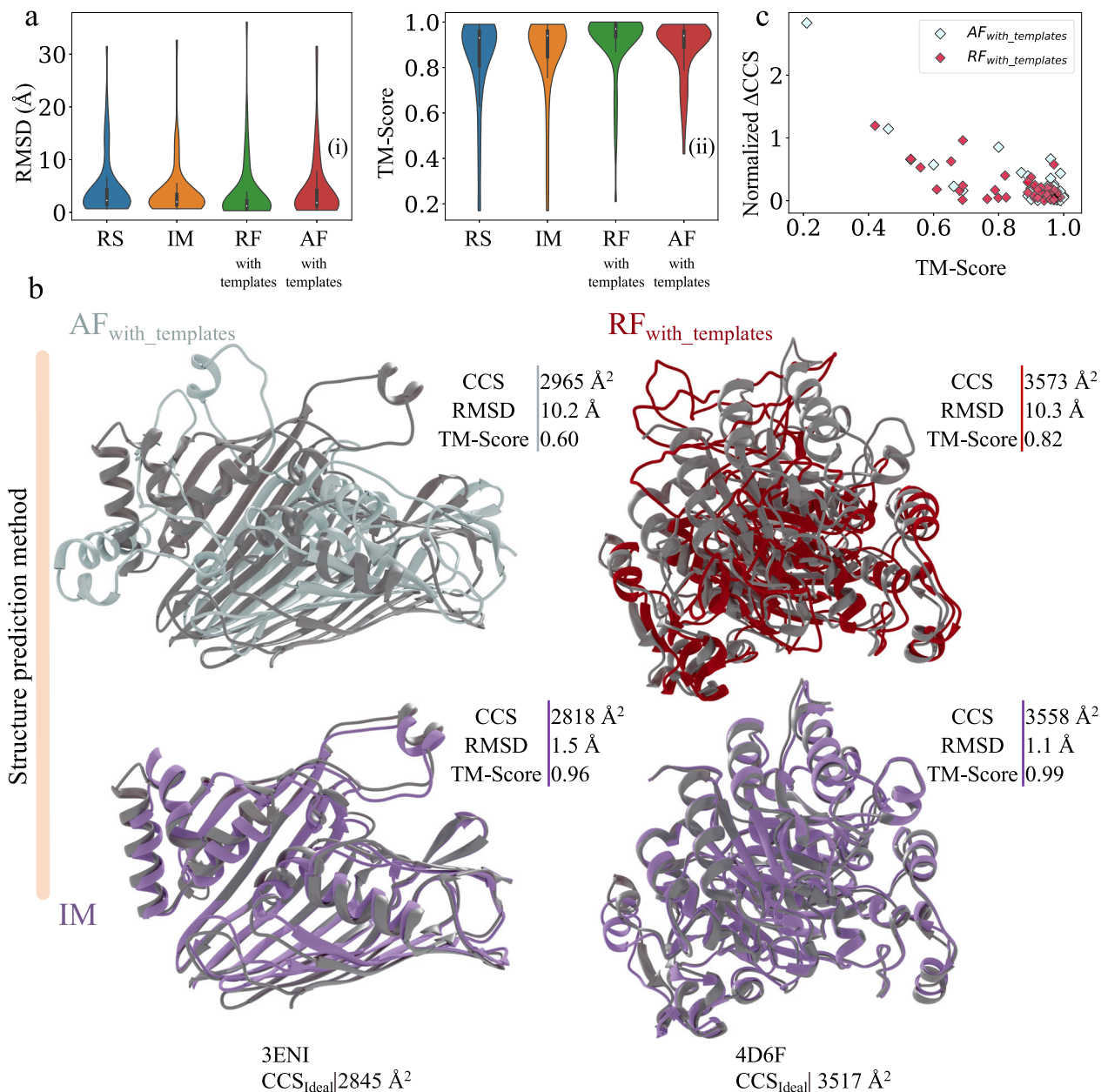


Fig. 3 | Comparison of predictions with ideal ion mobility (IM) data to those of AlphaFold (AF) and RF (RosettaFold). Predictions, for the ideal dataset, using the IM score function were compared to that of Rosetta (RS), AF_{with_templates}, and RF_{with_templates}. **a** Violin distributions ($n = 60$ biologically independent samples over 4 independent modelling approaches), of (i) root mean square deviation (RMSD) and (ii) template modelling score (TM-Score) of the predicted structure using the RS, IM, AF_{with_templates} and RF_{with_templates} score functions. For protein structure predictions with methods shown in **a** the mean and the standard error of mean in (i) are 4.46 ± 0.74 Å, 3.72 ± 0.65 Å, 3.43 ± 0.77 Å, and 4.37 ± 0.80 Å respectively. Similarly the mean and the standard error of mean in (ii) are 0.84 ± 0.03 , 0.86 ± 0.02 , 0.92 ± 0.02 , and 0.88 ± 0.02 respectively. For the violin

distributions in a (i) and a (ii) the white dots represent the median. The black bar in the center of the distribution is the interquartile range (IQR). The black stretched line extends from the “first quartile -1.5 IQR” to the “third quartile $+1.5$ IQR”. Values beyond this range are considered outliers. **b** Comparison of predicted structures with AF_{with_templates} (cyan), RF_{with_templates} (red) and IM (purple) to their native structures (grey) for the ideal dataset. **c** High normalized absolute difference in collision cross section of the predicted structure and the native structure (Δ CCS divided by sequence length) for structures predicted with AF_{with_templates} (cyan) and RF_{with_templates} (red) generally corresponded to structures with low TM-Score as seen for the ideal dataset. Source data are provided as a Source Data file.

dataset either improved or remained unchanged in all 25 cases. Similarly, the TM-Score either improved or remained unchanged for 22/25 proteins (Fig. 4b [iii]). The RMSD improved by an average of 2.0 Å (average TM-Score improvement of 0.03) when IM data were utilized as restraints. Of these 25 cases, 17 proteins were ultimately predicted with an RMSD of less than 5.5 Å, compared to 13 proteins without IM data (RS score function). Figure 4c shows structures of the predicted

models (aligned to the native structures in grey) obtained with the RG (pink), RS (blue), and IM (purple) score functions. The largest RMSD improvement was observed for the system β -crystallin B2 (PDB ID: 1YTQ), whose RMSD decreased from 17.7 Å to 5.0 Å. The TM-Score for this protein improved from 0.46 to 0.70 when the predicted structure from RS was compared to that of the IM score function. Similarly, for the system Hemolysin E (PDB ID: 1QOY), the IM score function

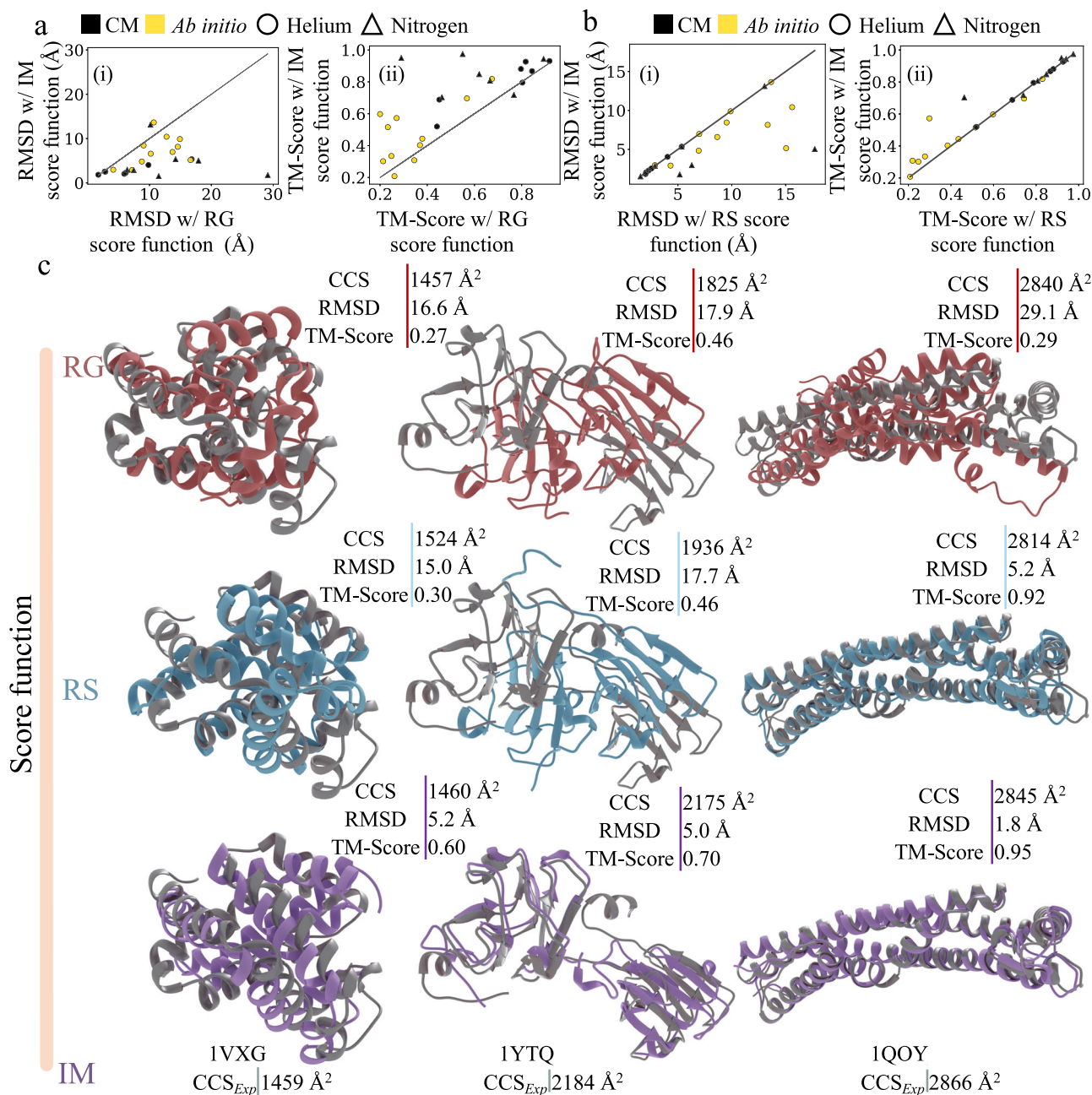


Fig. 4 | Protein structure prediction with and without experimental ion mobility (IM data). Consistent improvement in model selection using the IM score function was observed for the 25 proteins with experimental IM data. The predicted models from the IM score function were compared to those of the **a** radius of gyration (RG) and **b** Rosetta (RS) score functions in terms of their respective (i) root mean square deviation (RMSD) and (ii) template modelling score (TM-Score). For

both **a**, **b**, circle and triangle indicate IM data collected in helium and nitrogen buffer gas respectively; while models predicted with comparative modelling (CM) and *ab initio* are shown in black and yellow respectively. **c** Comparison of predicted structures using the RG (pink), RS (blue) and IM (purple) score function to their native structures (grey). Source data are provided as a Source Data file.

predicted a significantly better model when compared to that predicted by the RG score function. The RMSD and TM-Score differences (between the models selected by the IM and RG score functions) were 27.3 Å and 0.66 respectively. No significant difference was observed in IM score model selection with respect to IM data collected from the two different experimental conditions (helium and nitrogen buffer gas). This suggests that IM data from both buffer gases are equally useful for modelling. The score vs RMSD and TM-Score distributions for several benchmark proteins using the RS (blue) and IM (purple) score functions are shown in Fig. 5a (i) and Fig. 5a (ii) respectively. In these distributions, the predicted models from RS and IM are marked

with a blue and purple star, respectively. We observed a general improvement in P_{near} upon scoring with the IM score function with a 4.6-fold average improvement over that of the RS score function (Supplementary Data 7). The average P_{near} also increased by 43.4-fold when the score distributions were compared to those from the RG score function. This showed that the goodness of the score vs RMSD and TM-Score distribution is generally improved when IM data are included to predict structures. For comparison purposes, protein structures were also predicted with AF_{with_templates} and RF_{with_templates} for the experimental dataset as shown in Fig. 5b. The predicted models were compared to their native structures by RMSD as shown in

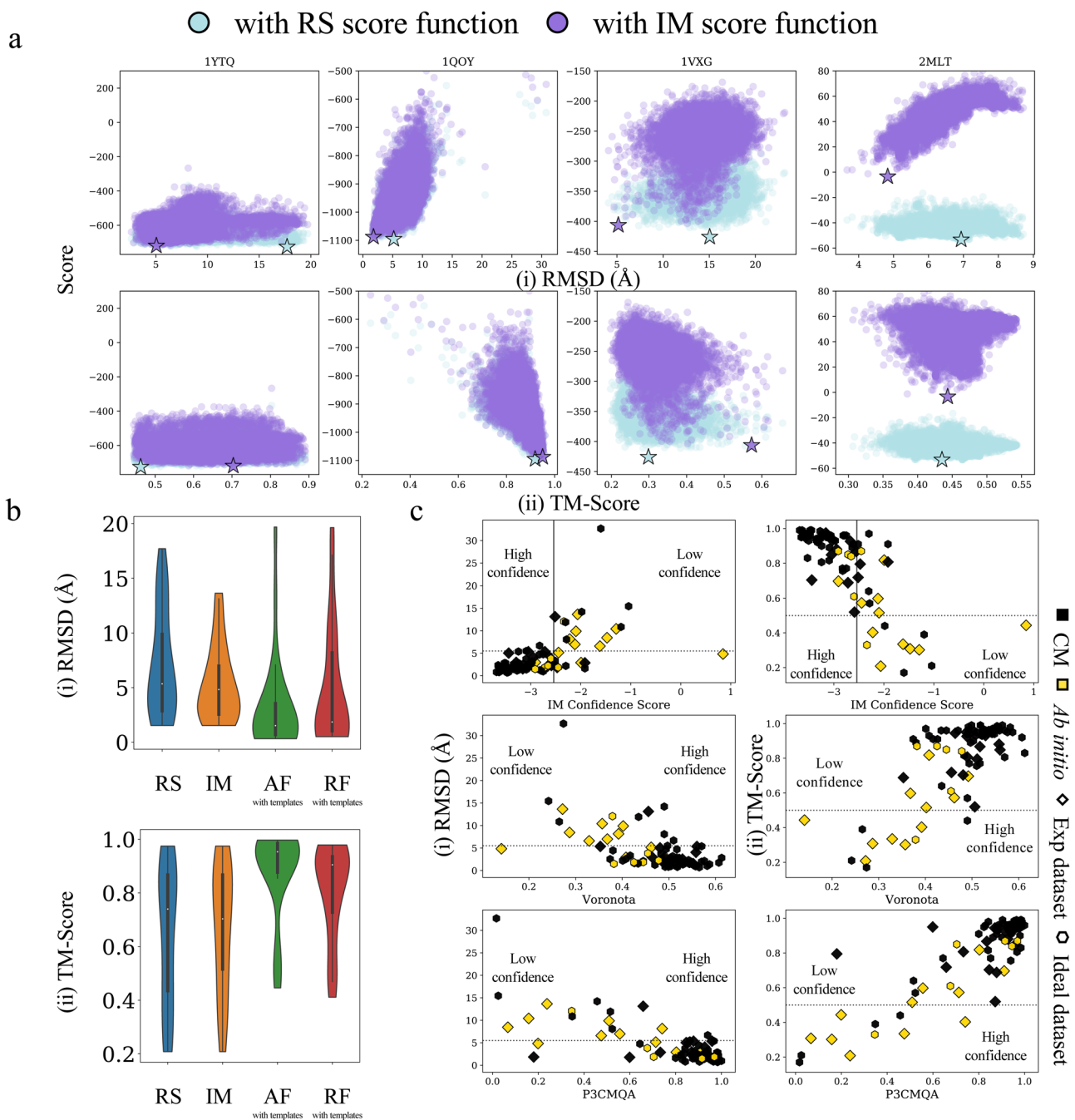


Fig. 5 | Protein structure prediction results for the experimental dataset and model quality assessment. The score distribution improves to a more funnel-like shape with the ion mobility (IM) score function for the experimental dataset. **a** This is highlighted by the score vs (i) root mean square deviation (RMSD) and (ii) template modelling score (TM-Score) distributions of four proteins that showed significant improvement with the IM (purple) score function over Rosetta (RS, blue) score function. The best scoring models from both predictions are marked with a blue and purple star for RS and IM, respectively. **b** Violin distributions, $n = 25$, biologically independent samples over 4 independent modelling approaches of (i) RMSD and (ii) TM-Score of the predicted structures with RS, IM, AlphaFold with templates (AF_{with templates}) and RosettaFold with templates (RF_{with templates}). For protein structure predictions with methods shown in **a** the

mean and the standard error of mean in (i) are $7.18 \pm 0.99 \text{ \AA}$, $5.33 \pm 0.70 \text{ \AA}$, $3.13 \pm 0.90 \text{ \AA}$, and $4.57 \pm 1.08 \text{ \AA}$ respectively. Similarly, the mean and the standard error of mean in (ii) are 0.64 ± 0.05 , 0.67 ± 0.05 , 0.88 ± 0.03 , and 0.81 ± 0.04 respectively. For the violin distributions in **b** (i) and **b** (ii) the white dots represent the median. The black bar in the center of the distribution is the interquartile range (IQR). The black stretched line extends from the “first quartile - 1.5 IQR” to the “third quartile + 1.5 IQR”. Values beyond this range are considered outliers. **c** Comparison of (i) RMSD and (ii) TM-Score of the predicted structures (with IM) from both the experimental (diamond) and ideal (hexagon) dataset vs IM confidence score. Models generated with comparative modelling (CM) and *ab initio* are shown in black and yellow respectively. Similarly, the Voronota and P3CMQA scores are shown. Source data are provided as a Source Data file.

Fig. 5b (i) (and **Fig. 5b** [ii] for TM-Score). On average, AF_{with templates} and RF_{with templates} both predicted structures better than the IM score function. Additionally for the experimental dataset, we observed that use of templates with AF and RF improved the RMSD and TM-Score

difference when compared to prediction results without templates (Supplementary Fig. 5). The RMSD difference improved by 0.58 Å and 1.94 Å for AF and RF respectively. Similarly, the TM-Score difference for AF and RF improved by 0.03 and 0.04 respectively. Similarly,

structures predicted with the IM score function were better than those predicted with RS (without the aid of IM data). Despite the impressive performance of AF and RF, there were several cases where inclusion of IM data outperformed those predictions. There were 6 cases (out of 25 proteins in the experimental dataset) where structure prediction was better or the same, with the IM score function when compared to that of AF_{with_templates} (Supplementary Data 7). The average RMSD and TM-Score difference of this subset were 3.2 Å and 0.04 respectively. Similarly, there were 9 cases (Supplementary Data 7) where predictions from IM outperformed those models predicted by RF_{with_templates}. The average RMSD and TM-Score of this subset differed by 5.4 Å and 0.13. Furthermore, the average absolute percent error of CCS_{PARCS} to CCS_{IM} of the top scoring models of RG, RS, RF_{with_templates}, AF_{with_templates}, and IM were 8.6%, 7.6%, 6.0%, 5.5%, and 5.4%, respectively (Supplementary Data 8). This confirmed that the IM score function predicts models that best agree with the experimental IM data when compared to all other structure prediction methods (RG, RS, RF, and AF). Furthermore, we focused on a subset of 54 proteins from both the experimental and ideal dataset for which either the CM protocol with non-perfect templates or the ab initio protocol (template-free modelling) were performed. For this subset, we defined non-perfect templates as templates with sequence identity and coverage (to target protein) ranging anywhere from 14% to 84% and 6% to 100% respectively. Next, for this subset we compared the best scoring models to their native structures for the radius of gyration (RG), Rosetta (RS), and ion mobility (IM) score functions. Again, we observed consistent improvement in model selection for the IM score function over the RG and RS score functions, for both the ideal and experimental dataset as shown in Supplementary Fig. 6. Compared to the RG score function, the IM score improved the RMSD of the selected model (for all the 54 proteins in this subset, Supplementary Data 9) by 2.72 Å on average. The TM-Score for this comparison also improved by 0.08 on average. Similarly, compared to the RS score function, the IM score improved the RMSD of the selected model by 1.12 Å and the TM-Score by 0.02. A summary of this analysis can be found in Supplementary Data 10. Thus, we conclude that IM data can improve protein structure prediction both in complete absence of templates and in the presence of non-perfect templates. In summary, these results demonstrated that experimental IM data can offer shape and size information that can be used to improve protein structure prediction.

IM confidence score discriminates accurate and inaccurate models

The inclusion of IM data helped improve structure prediction for all 25 proteins in the experimental dataset. However, there were 8 cases where the RMSD of the selected model (even after improvement) was greater than 5.5 Å (five of those eight cases had TM-Scores less than 0.5). This knowledge was available to us since the native structures were known for the models generated within this benchmark dataset. However, in true blind structure prediction protocols, RMSD or TM-Score information is not available. For this reason, we developed a confidence measure, the IM confidence score, that allowed us to selectively flag successful prediction cases in the absence of native structure. The IM confidence score was defined as the average score of the top 100 scoring models divided by sequence length. According to this metric analysis, the high and low confidence structures were separated by a score cutoff of -2.54 . This metric flagged all inaccurate predictions as low confidence, whereas all high confidence predictions were accurate. We tested whether the IM confidence score made predictions that surpassed or at least matched other software that can assess model quality in the absence of IM data. For this purpose, IM confidence score results were benchmarked against two other model quality assessment programs (Voronota and P3CMQA) as shown in Fig. 5c for both the experimental (diamond) and ideal dataset (hexagon). To better

compare the performance of the IM confidence score to Voronota and P3CMQA, the confidence score was scaled (as outlined in Supplementary Methods) such that it ranged from 0 to 1, with 1 being the most confident model and 0 being the least confident (similar to the convention in Voronota and P3CMQA) as shown in Supplementary Fig. 1 (b) with (i) RMSD and (ii) TM-Score. Our results (Fig. 5c and Supplementary Fig. 1 [b]) indicated that the IM confidence score were comparable to that of Voronota and P3CMQA.

Discussion

Ion mobility (IM) has emerged as a prime tool to study proteins in their native states using mass spectrometry (MS) due its ability to conserve native-like structural information in the gas phase. Furthermore, native IM-MS measurements are relatively fast, use very little sample, and are highly chemically specific, making them both relatively easy and informative as compared to many other types of structural biology techniques. Additionally, native IM-MS measurements are not limited by the size of the system. For these reasons, IM-MS provides a wealth of structural information and can be used as routine analysis when compared to many other types of experimental techniques. However, the information obtained is sparse, not directly allowing for full structure elucidation. Thus, computational techniques are needed to deduce structural information from IM data. In this study we developed a new algorithm for structure prediction of single subunit proteins from IM data. To achieve this, we first developed a method (PARCS) that could predict collision cross section (CCS) from structures, which has been implemented in Rosetta as a stand-alone application. Following the successful benchmarking of this application, a score term, based on restraints derived from IM data, has been developed to predict native-like structures. This score term was tested on a set of 60 structures from the PDB, where CCS_{PARCS} (with simulated noise) of the native structure was treated as the experimental CCS_{IM}. This was done, as a proof of principle, to check whether the score function could translate the structural information (encoded in IM data) into spatial restraints in the absence of model error. Based on RMSD and TM-Score analysis, we observed that the inclusion of IM data (to the Rosetta score function) improved structure prediction results for 58 out of 60 structures. Since radius of gyration (R_g) has previously been used as a simple substitute for IM data, we tested the IM score function against a proxy score function that utilized R_g to favor compact models. From our results, we conclude that the two quantities (CCS and R_g) point to different structures and thus a simple compactness-based proxy score function cannot be used as a substitute for the IM score function. Following this positive validation, the score function was tested on a benchmark set of 25 proteins with experimental IM data. We showed that IM data improved model selection, as demonstrated by analyzing the best scoring models with several metrics. Next, we benchmarked our method against recently developed structure prediction methods, AlphaFold2 (AF), and RoseTTAFold (RF), for both the ideal and experimental datasets. Despite the remarkable accuracy of these deep learning methods, our results show that there were several cases where the IM score function could improve structure prediction over AF and RF. Furthermore, our results also suggest that large deviations of predicted CCS from experimental CCS for AF and RF structures are indicative of models with poor fold, further underscoring the usefulness of IM data at aiding accurate structure identifications. We also developed a confidence metric (IM confidence score) to successfully separate good predictions from bad predictions in the absence of native structure. Our current computational workflow illustrates that CCS obtained from IM experiments, despite its sparseness, provides sufficient information on the overall shape and size of proteins to be used as restraints to improve model selection in protein structure prediction. Furthermore, our results also suggest that, despite its extensive assumptions and approximations, the projection approximation method is sufficient in

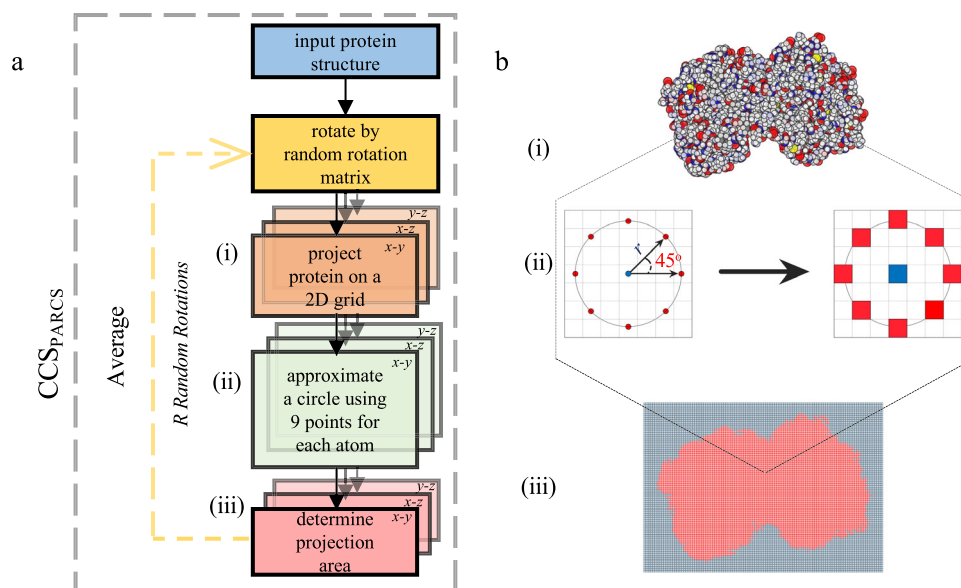


Fig. 6 | Overview of collision cross section (CCS) calculation using rough circular shapes (PARCS) algorithm. a Schematic of the PARCS algorithm to predict CCS from structure. (i) Three projections are obtained from each rotation. (ii) For each atom in each projection the 2 dimensional (2D) grid is filled according to a 9-point circle approximation. (iii) The projection area is determined from the number of filled grid cells. **b** Illustration of a (i) 2D projection of a single random

rotation where the carbon, sulfur, oxygen, nitrogen, and hydrogen are colored grey, yellow, red, blue, and white respectively. (ii) Each atom is projected on a grid with a cell size of 1 \AA^2 . The center grid cell and eight other grid cells at a distance r (based on the radii of the given atom and the buffer gas) from the center of the atom are filled. (iii) Projection of the randomly rotated protein after the grid cells are filled according to the PARCS algorithm.

an integrative modelling pipeline and provides the ability to rapidly compare large numbers of computed results to experiment, which can be very time-consuming with more physically explicit methods such as the Trajectory Method. With these benchmarks in place, future investigation using these higher-level CCS computation methods could help refine the model further. This study also demonstrates how information from native IM techniques (gas phase) can be used to successfully infer solution structure. Our developed CCS calculation method and score function are freely and easily accessible through Rosetta Commons. Supplementary Note 1 shows examples on how to use the PARCS application and Supplementary Note 2 contains instructions on the use of IM data in structure prediction. Furthermore, all related data (including all models generated) can be found in our GitHub repository⁷⁵ (<https://doi.org/10.5281/zenodo.6726418>). Further work will focus on improving methods to incorporate CCS data for protein complexes using RosettaDock⁷⁶ and on the use of multiple complementary data types (such as the combination of covalent labeling^{77–79}, surface-induced dissociation⁸⁰, cryo-EM^{81,82} and/or NMR^{83,84} with IM data) for protein and protein complex structure prediction in Rosetta.

Methods

Projection approximation using rough circular shapes

Average CCS of biomolecules are determined from IM experiments based on the amount of time required for the ion to traverse the region of inert buffer gas (usually helium or nitrogen) under the influence of a weak electric field^{43,45}. To use IM data in a structure prediction protocol, we developed Projection Approximation using Rough Circular Shapes (PARCS) in Rosetta. The schematic (a) and the illustration (b) in Fig. 6 demonstrate how the PARCS algorithm computes CCS from structure and estimates area of a projection, respectively. The PARCS algorithm, as shown in Fig. 6a, takes 3D atomic protein coordinates as input. Next, the structure is randomly rotated. For each rotation, the structure is projected on a 2D grid (grid cell area of 1 \AA^2) in the x - y , x - z , and y - z planes as shown in Fig. 6a (i). In the 2D grid, the projection of the protein is centered, and the grid extends 5 \AA beyond the most extreme atom in each direction. For each atom on the 2D grid (Fig. 6a

[ii]), the center grid cell is filled as denoted by the blue grid cell in Fig. 6b (ii). Then, eight additional cells (red grid cells in Fig. 6b [ii]) are also filled. The distance of these eight grid cells from the central cell (i.e., radius of the circular projection) is based on the sum of the radii of the projected atom and the buffer gas (r in Fig. 6b [ii]). An effective atomic cross-sectional radius of 1.91 \AA is used for heavy atoms (carbon, sulfur, oxygen, nitrogen, and phosphorous) and 1.21 \AA is used for hydrogen atoms. A buffer gas radius of 1.0 \AA and 1.82 \AA is used in the case of helium⁴³ and nitrogen⁸⁵, respectively. The eight points are positioned such that two adjacent points on the circumference form a 45° angle from the center point as shown in Fig. 6b (ii). This process is repeated for all atoms in the protein, filling the overall grid as shown in Fig. 6b (iii). Finally, the projection area (A) is derived by summing the areas of the filled grid cells. From the x - y , y - z , and x - z projections for each random rotation, three projection areas (A_i^{x-y} , A_i^{y-z} and A_i^{x-z}) are obtained. The CCS of the structure ($\text{CCS}_{\text{PARCS}}$) is then acquired from the average area of the total number of projections ($N = 3R$, where R is the total number of random rotations) as shown in Eq. 1.

$$\text{CCS}_{\text{PARCS}} = \frac{\sum_i^R (A_i^{x-y} + A_i^{y-z} + A_i^{x-z})}{N} \quad (1)$$

IM score function in Rosetta

CCS from experimental IM data were incorporated as a spatial restraint for integrative Rosetta modelling as it provides information about protein size and shape. Therefore, to integrate this information in Rosetta for protein structure prediction, a score term ($\text{IM}_{\text{Score_Term}}$) was developed to quantify agreement of protein structures with IM data, using CCS as the restraint. The evaluation score, IM_{Score} , was defined as a sum of the $\text{IM}_{\text{Score_Term}}$ score term with the Rosetta REF2015 score function⁶⁵ as shown in Eq. 2.

$$\text{IM}_{\text{Score}} = \text{RS} + \text{IM}_{\text{Score_Term}} \quad (2)$$

In Eq. 2, RS is the energy of the structure obtained from the Rosetta REF2015 score function. The $\text{IM}_{\text{Score_Term}}$ term is a penalty function (as

defined and shown in Eq. 3 and Supplementary Fig. 7, respectively) based on the absolute difference (ΔCCS) between $\text{CCS}_{\text{PARCS}}$ and CCS_{IM} . This function includes a lower bound (LB) and an upper bound (UB) cutoff (as shown in Eq. 3) to account for error²⁴. ΔCCS below LB (10 \AA^2) are not penalized and ΔCCS above UB (100 \AA^2) are given a maximum penalty of 100, with a fade function used in between. Conceptually, this scoring function penalizes structures with high deviation from experiment.

$$\text{IM}_{\text{ScoreTerm}} = \begin{cases} 0 & \text{if } \Delta\text{CCS} < \text{LB} \\ 100(2x^3 - 3x^2 + 1) & \text{if } \text{LB} < \Delta\text{CCS} < \text{UB} \\ 100 & \text{if } \Delta\text{CCS} > \text{UB} \end{cases} \quad (3)$$

$$x = -\left(\frac{\Delta\text{CCS} - \text{UB}}{\text{UB} - \text{LB}}\right)$$

IM datasets

In this work, our aim was to study predominantly globular and ordered proteins within all datasets. Values from $\text{CCS}_{\text{PARCS}}$ were compared to $\text{CCS}_{\text{IMPACT}}$ as well as evaluated for speed and precision on 4465 non-homologous protein structures (PARCS evaluation dataset) extracted from the protein databank (PDB)⁸⁶ (<http://www.rcsb.org/>) using the PISCES⁸⁷ webserver (<http://dunbrack.fccc.edu/piscs>). For this dataset (all PDB IDs used can be found in the Source Data file), the sequence identity was less than or equal to 10%, sequence length was between 40–250 residues, non-X-ray and CA-only entries were excluded and the PDBs were culled by chain. For CCS prediction and speed comparison, PARCS was benchmarked against IMPACT⁴³ (with flag ‘-H’ to include hydrogens) based on the calculations performed on the PARCS evaluation dataset. This dataset was also used to test the convergence of PARCS with respect to the number of rotations. In this convergence test, the standard deviation of 100 separate CCS calculations for each protein at varying numbers of rotations were obtained and assessed for the optimal number of random rotations required for calculations to converge.

To evaluate the ability of the score term (Eq. 3) to distinguish native from non-native protein models in the case of an error-free CCS prediction, a set of 60 proteins was selected from the PDB such that it contained all unique architectures (list of monomers shown in Supplementary Data 1) as classified by the CATH Protein Structure Classification database⁸⁸. The sequence length for proteins in this dataset ranged from 58 to 965. A set of structure prediction experiments (which will be described in detail in the following sections) was performed on this dataset, where the experimental CCS was simulated by predicting CCS of the native structure with PARCS. Therefore, this dataset was referred to as the ideal dataset. The simulated CCS ($\text{CCS}_{\text{ideal}}$) values ranged from 767 \AA^2 to 4130 \AA^2 for the 60 proteins in the ideal dataset. Furthermore, to address the effect of uncertainty in CCS values when using the IM score function (Eq. 2), varying degrees of noise were introduced to the simulated CCS data (as outlined in Supplementary Methods). The score function was also tested on actual experimental IM data, i.e., structures with CCS_{IM} (experimental dataset). The experimental dataset^{18,89–93} consisted of 25 monomeric proteins that also had structural information deposited in the PDB (as outlined in Supplementary Data 6). Sequence lengths ranged from 26–691 residues and CCS_{IM} values (for the lowest charge states) ranged from 588 \AA^2 to 4580 \AA^2 . Additionally, the proteins exhibited an average percent disorder of only 13.2% and 10.7% as calculated by the Rosetta ResidueDisorder^{94,95} application for the experimental and ideal dataset respectively.

Ab initio, comparative modelling, AlphaFold2, and RoseTTA-Fold protocols for structure prediction

To test whether shape and size information encoded in IM data were sufficient to discriminate between low and high RMSD (and TM-Score)

models of single-subunit proteins, we tested our algorithm on both the ideal and experimental dataset. For these two datasets, the Rosetta (v.3.1.3) ab initio protocol was used for proteins with sequence length less than 155 residues, otherwise the Rosetta multi-template comparative modelling (CM) protocol was used. The templates and weights associated with all proteins for CM are provided in Supplementary Data 11 and Supplementary Data 12 for the ideal and experimental dataset respectively. The 3mer and 9mer fragments required for both protocols were generated using the fragment picker tool⁹⁶ in Rosetta. The protocols (ab initio and CM) for both the ideal and experimental data set are further detailed in the Supplementary Methods. All structures generated from the ab initio and comparative modelling protocols (including the recovered structures for selected proteins in the experimental dataset as shown in Supplementary Data 13) were subjected to the Rosetta Relax protocol (term referred to as RS in Eq. 2). The IM data, ideal and experimental, were then used to score all the structures generated for each protein in Supplementary Data 1 and Supplementary Data 6, respectively. Radius of gyration has previously been used as proxy for IM data⁹⁷. Therefore, the IM score function was also benchmarked against a simple proxy score function (RG) that was solely based on radius of gyration (described in the Supplementary Methods section), and thus favored compact models for proteins in both datasets. The top scoring model was designated as the predicted structure. AlphaFold2 (v.2.0.0)⁹⁸ and RoseTTAFold (v.1.1.0)⁹⁹ (protocols detailed in the Supplementary Methods) were also benchmarked (with and without templates) for both the datasets to further assess the effectiveness of IM data.

Analysis metrics used for evaluating predictions

We quantitatively assessed the quality of our predicted models using several of the following metrics. The global RMSDs (root-mean-square deviations) of the predicted models to their native structures were calculated. Predictions with IM data where RMSD was within 0.5 \AA of the RMSD of the structure predicted without IM data were defined as unchanged. Next, P_{near} ¹⁰⁰, a goodness-of-energy funnel metric (at $k_B T$ and λ set to 10 and 1 \AA respectively), was used to compare the score versus RMSD distributions predicted with the RG, RS, and IM score functions. P_{near} ranges from 0 (a poor energy funnel) to 1 (a well-defined energy funnel). All predicted structures from both ideal and experimental datasets were further evaluated with the template modelling score (TM-Score)¹⁰¹. TM-Score was used to assess the topological similarity of the predicted structures to native structures using the TM-Score program¹⁰¹. The TM-Score metric ranges from 0 to 1, where scores below 0.17 indicate randomly chosen unrelated proteins and a score higher than 0.5 corresponds to structures being generally in the same fold and a score of 1 indicates a perfect match¹⁰¹.

Confidence metric used for identifying accurate and inaccurate predictions

A metric was developed (IM confidence score) to quantify confidence in predictions in the absence of known structure. The IM confidence score was defined as the average score of the top 100 scoring models predicted with IM data (using Eq. 2) divided by the sequence length (i.e., average score per residue). The specific metric was chosen because lower scores per residue are generally associated with more native-like structures. Thus, structures were defined as high confidence if the average residue score was less than -2.54 (above which structures were defined as low confidence). Instances where the RMSD of the prediction was less than 5.5 \AA (correspondingly above 0.5 TM-Score) and the average residue score was less than -2.54 were considered successful confidence measure cases. We chose an RMSD cutoff of 5.5 \AA (TM-Score cutoff of 0.5) since below that RMSD (and above that TM-Score), protein topologies are generally predicted correctly. The IM confidence score was further tested against two

other model quality programs Voronota (v1.22.3149)¹⁰² and P3CMQA (v.1.0.0)¹⁰³ (protocol outlined in Supplementary Methods).

Analysis of compactness of model selected with IM score function

We also investigated (using the ideal dataset) whether models predicted with the IM score function were biased towards a possible gas phase energy minimum which might show signs of structural compaction. To do this, we reduced the CCS_{ideal} by 2% to 30% and rescored with the IM score function. We subsequently analyzed the top scoring structures (using the reduced CCS_{ideal} as part of scoring) for compactness using radius of gyration. Next, we calculated the neighbor count (NC) of all residues with the Rosetta application `per_residue_solvent_exposure`²⁵ (using the sphere method with default parameters). A residue with high NC is approximated as buried, while a residue with low NC is thought to be solvent exposed. Therefore, we analyzed the NC of surface residues at varying reduced CCS_{ideal} . The surface residues were approximated as the top X % residues with the lowest NC in the native structure, where X varied from 5% to 20%.

Software usage for data analysis

Python v.3.7.3 was used for data analysis. Matplotlib v.3.1.2 was used for the creation of all scatter plots, line plots and violin distributions. PyMOL v.2.0.6 and Blender v.2.8.1 were used in combination to generate the figures of all proteins.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All crystal structures used in this study were retrieved from the Protein Data Bank (PDB) and can be accessed at <https://www.rcsb.org>. The accession codes of the structures in the ideal dataset are: 1I24, 1JMA, 1K7G, 1KB0, 1KTU, 1LII, 1N7V, 1NLT, 1OKC, 1QHU, 1QQC, 1R3F, 1RNE, 1SUU, 1T9F, 1YNF, 1YU0, 1ZVC, 2F3L, 2FM9, 2GD5, 2ISB, 2JAI, 2P8H, 2QGQ, 2QSD, 2ZZQ, 3A1Y, 3BXO, 3C7X, 3CPW, 3E8T, 3EB7, 3EF6, 3ENI, 3G9I, 3GUA, 3KYJ, 3M7M, 3N99, 3ODJ, 3POY, 3RST, 3S2O, 3SAE, 3VZ9, 3WPV, 4AII, 4D6F, 4E7G, 4IIM, 4QMJ, 4XTK, 5KIS, 5LB7, 5MIN, 5U69, 5VSK, 6AZZ, 6S2M. The accession codes of the structures in the experimental dataset are: 1BEB, 1BNI, 1CFD, 1DPX, 1EX3, 1FD3, 1FS3, 1HFX, 1HRC, 1J7N, 1LDS, 1LFG, 1OVA, 1QOY, 1UBQ, 1VXG, 1YTQ, 2MLT, 3INS, 3QYT, 3VWI, 4F5S, 4H2A, 6DAH, 6PTI. The CATH Protein Structure Classification database was used to determine the architectures of proteins and can be accessed here: <http://www.cathdb.info>. The processed simulation data in this work are available without any restriction from GitHub (<https://github.com/smturzo/IMMS/tree/v.1.0.0>) and Zenodo⁷⁵ (<https://doi.org/10.5281/zenodo.6726418>). Additionally, our GitHub repository contains instructions on how this work can be reproduced. Access to raw simulation data (not present in this repository due to size limitation) can be obtained by emailing the corresponding author (lindert.1@osu.edu). Processed simulation data can also be found in the Supplementary Data files. Source data are provided with this paper.

Code availability

The Collision Cross Section application (PARCS v.1.0.0) and the Ion Mobility score function are available for free to academic users through the Rosetta software suite at <https://www.rosettacommons.org/software/>. The current academic version of Rosetta (3.13) can be freely downloaded from <https://els2.comotion.uw.edu/product/rosetta> for academic users. The source code for PARCS and the Ion Mobility score function (which are part of the Rosetta codebase) are only made available to academic/non-profit/government entities and commercial entities with a Company Contributor License. While availability to the Rosetta codebase is free for academics/non-profit/

government entities, note that there is a Rosetta license fee for industry users to gain access to the source code and the applications in Rosetta (including the PARCS application and the Ion Mobility score function). Currently the University of Washington exclusively manages all Rosetta licensing. More information on Rosetta licensing can be found at <https://www.rosettacommons.org/about/faq>. Instructions to run the PARCS application in Rosetta can be found in Supplementary Note 1. Instructions on protein structure prediction with IM data can be found in Supplementary Note 2. Both of these instructions are accompanied by a step-by-step example. Additionally, instructions and results of modifying the AlphaFold source code such that it predicts models without the aid of templates can be found in Supplementary Note 3 and Supplementary Data 14, respectively.

References

- Huang, P.-S., Boyken, S. E. & Baker, D. The coming of age of de novo protein design. *Nature* **537**, 320–327 (2016).
- Coelho, E. D., Arrais, J. P. & Oliveira, J. L. From protein-protein interactions to rational drug design: are computational methods up to the challenge? *Curr. Top. Med Chem.* **13**, 602–618 (2013).
- Leelananda, S. P. & Lindert, S. Computational methods in drug discovery. *Beilstein J. Org. Chem.* **12**, 2694–2718 (2016).
- Kiss, G., Çelebi-Ölçüm, N., Moretti, R., Baker, D. & Houk, K. N. Computational enzyme design. *Angew. Chem. Int Ed. Engl.* **52**, 5700–5725 (2013).
- King, N. P. et al. Computational design of self-assembling protein nanomaterials with atomic level accuracy. *Science* **336**, 1171–1174 (2012).
- Khoury, G. A., Smadbeck, J., Kieslich, C. A. & Floudas, C. A. Protein folding and de novo protein design for biotechnological applications. *Trends Biotechnol.* **32**, 99–109 (2014).
- Wyttenbach, T. & Bowers, M. T. 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 (2011).
- Ruotolo, B. T. & Robinson, C. V. Aspects of native proteins are retained in vacuum. *Curr. Opin. Chem. Biol.* **10**, 402–408 (2006).
- Bleilholder, C. & Liu, F. C. Structure Relaxation Approximation (SRA) for Elucidation of Protein Structures from Ion Mobility Measurements. *J. Phys. Chem. B* **123**, 2756–2769 (2019).
- Artigues, A. et al. Protein Structural Analysis via Mass Spectrometry-Based Proteomics. *Adv. Exp. Med. Biol.* **919**, 397–431 (2016).
- Danev, R., Yanagisawa, H. & Kikkawa, M. Cryo-Electron Microscopy Methodology: Current Aspects and Future Directions. *Trends Biochem Sci.* **44**, 837–848 (2019).
- Leney, A. C. & Heck, A. J. R. Native Mass Spectrometry: What is in the Name? *J. Am. Soc. Mass Spectrom.* **28**, 5–13 (2017).
- Allison, T. M. et al. Computational Strategies and Challenges for Using Native Ion Mobility Mass Spectrometry in Biophysics and Structural Biology. *Anal. Chem.* **92**, 10872–10880 (2020).
- Matthew Allen Bullock, J., Schwab, J., Thalassinos, K. & Topf, M. The Importance of Non-accessible Crosslinks and Solvent Accessible Surface Distance in Modeling Proteins with Restraints From Crosslinking Mass Spectrometry. *Mol. Cell Proteom.* **15**, 2491–2500 (2016).
- Mendoza, V. L. & Vachet, R. W. Probing protein structure by amino acid-specific covalent labeling and mass spectrometry. *Mass Spectrom. Rev.* **28**, 785–815 (2009).
- Roberts, V. A., Pique, M. E., Hsu, S. & Li, S. Combining H/D Exchange Mass Spectrometry and Computational Docking To Derive the Structure of Protein-Protein Complexes. *Biochemistry* **56**, 6329–6342 (2017).
- Zhou, M. & Wysocki, V. H. Surface induced dissociation: dissecting noncovalent protein complexes in the gas phase. *Acc. Chem. Res.* **47**, 1010–1018 (2014).

18. Jurneczko, E. & Barran, P. E. 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 (2011).
19. Seffernick, J. T. & Lindert, S. Hybrid methods for combined experimental and computational determination of protein structure. *J. Chem. Phys.* **153**, 240901 (2020).
20. Kuhlman, B. & Bradley, P. Advances in protein structure prediction and design. *Nat. Rev. Mol. Cell Biol.* **20**, 681–697 (2019).
21. Alber, F., Forster, F., Korkin, D., Topf, M. & Sali, A. Integrating diverse data for structure determination of macromolecular assemblies. *Annu Rev. Biochem.* **77**, 443–477 (2008).
22. Saltzberg, D. J. et al. Using Integrative Modeling Platform to compute, validate, and archive a model of a protein complex structure. *Protein Sci.* **30**, 250–261 (2021).
23. Rajabi, K., Ashcroft, A. E. & Radford, S. E. Mass spectrometric methods to analyze the structural organization of macromolecular complexes. *Methods* **89**, 13–21 (2015).
24. Seffernick, J. T., Harvey, S. R., Wysocki, V. H. & Lindert, S. Predicting Protein Complex Structure from Surface-Induced Dissociation Mass Spectrometry Data. *ACS Cent. Sci.* **5**, 1330–1341 (2019).
25. Aprahamian, M. L., Chea, E. E., Jones, L. M. & Lindert, S. Rosetta Protein Structure Prediction from Hydroxyl Radical Protein Footprinting Mass Spectrometry Data. *Anal. Chem.* **90**, 7721–7729 (2018).
26. Aprahamian, M. L. & Lindert, S. Utility of Covalent Labeling Mass Spectrometry Data in Protein Structure Prediction with Rosetta. *J. Chem. Theory Comput.* **15**, 3410–3424 (2019).
27. Biehn, S. E. & Lindert, S. Accurate protein structure prediction with hydroxyl radical protein footprinting data. *Nat. Commun.* **12**, 341 (2021).
28. Hall, Z., Politis, A. & Robinson, C. V. Structural Modeling of Heteromeric Protein Complexes from Disassembly Pathways and Ion Mobility-Mass Spectrometry. *Structure* **20**, 1596–1609 (2012).
29. Kahraman, A. et al. Cross-link guided molecular modeling with ROSETTA. *PLoS one* **8**, e73411–e73411 (2013).
30. Tubb, M. R., Silva, R. A., Fang, J., Tso, P. & Davidson, W. S. A three-dimensional homology model of lipid-free apolipoprotein A-IV using cross-linking and mass spectrometry. *J. Biol. Chem.* **283**, 17314–17323 (2008).
31. Hauri, S. et al. Rapid determination of quaternary protein structures in complex biological samples. *Nat. Commun.* **10**, 192 (2019).
32. Saltzberg, D. J. 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* **121**, 3493–3501 (2017).
33. Zhang, M. M. 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.* **91**, 15709–15717 (2019).
34. Xie, B., Sood, A., Woods, R. J. & Sharp, J. S. Quantitative Protein Topography Measurements by High Resolution Hydroxyl Radical Protein Footprinting Enable Accurate Molecular Model Selection. *Sci. Rep.* **7**, 4552 (2017).
35. Politis, A., Park, A. Y., Hall, Z., Ruotolo, B. T. & Robinson, C. V. 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.* **425**, 4790–4801 (2013).
36. Degiacomi, M. T. On the Effect of Sphere-Overlap on Super Coarse-Grained Models of Protein Assemblies. *J. Am. Soc. Mass Spectrom.* **30**, 113–117 (2019).
37. Eschweiler, J. D., Frank, A. T. & Ruotolo, B. T. Coming to Grips with Ambiguity: Ion Mobility-Mass Spectrometry for Protein Quaternary Structure Assignment. *J. Am. Soc. Mass Spectrom.* **28**, 1991–2000 (2017).
38. Eschweiler, J. D., Farrugia, M. A., Dixit, S. M., Hausinger, R. P. & Ruotolo, B. T. A Structural Model of the Urease Activation Complex Derived from Ion Mobility-Mass Spectrometry and Integrative Modeling. *Structure* **26**, 599–606.e593 (2018).
39. Wang, H. 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.* **30**, 876–885 (2019).
40. Harvey, S. R. et al. Relative interfacial cleavage energetics of protein complexes revealed by surface collisions. *Proc. Natl Acad. Sci. USA* **116**, 8143–8148 (2019).
41. Graves, D. B. Transport properties of ions in gases by Edward A. Mason and Earl W. McDaniel, John Wiley and Sons, New York, 1988, 560 + xvi pp. *AIChE J.* **35**, 701–701 (1989).
42. Mason, E. A. & Schamp, H. W. Mobility of gaseous ions in weak electric fields. *Ann. Phys.* **4**, 233–270 (1958).
43. Marklund, E. G., Degiacomi, M. T., Robinson, C. V., Baldwin, A. J. & Benesch, J. L. P. Collision Cross Sections for Structural Proteomics. *Structure* **23**, 791–799 (2015).
44. Mesleh, M. F., Hunter, J. M., Shvartsburg, A. A., Schatz, G. C. & Jarrold, M. F. Structural Information from Ion Mobility Measurements: Effects of the Long-Range Potential. *J. Phys. Chem.* **100**, 16082–16086 (1996).
45. Ewing, S. A., Donor, M. T., Wilson, J. W. & Prell, J. S. Collidoscope: An Improved Tool for Computing Collisional Cross-Sections with the Trajectory Method. *J. Am. Soc. Mass Spectrom.* **28**, 587–596 (2017).
46. Larriba, C. & Hogan, C. J. Free molecular collision cross section calculation methods for nanoparticles and complex ions with energy accommodation. *J. Computational Phys.* **251**, 344–363 (2013).
47. Shvartsburg, A. A. & Jarrold, M. F. An exact hard-spheres scattering model for the mobilities of polyatomic ions. *Chem. Phys. Lett.* **261**, 86–91 (1996).
48. Bleiholder, C., Wyttenbach, T. & Bowers, M. T. 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 (2011).
49. Bleiholder, C. A local collision probability approximation for predicting momentum transfer cross sections. *Analyst* **140**, 6804–6813 (2015).
50. Mack, E. Average cross-sectional areas of molecules by gaseous diffusion methods. *J. Am. Chem. Soc.* **47**, 2468–2482 (1925).
51. Heck, A. J. R. Native mass spectrometry: a bridge between interactomics and structural biology. *Nat. Methods* **5**, 927–933 (2008).
52. Benesch, J. L. P., Ruotolo, B. T., Simmons, D. A. & Robinson, C. V. Protein Complexes in the Gas Phase: Technology for Structural Genomics and Proteomics. *Chem. Rev.* **107**, 3544–3567 (2007).
53. Snyder, D. T., Harvey, S. R. & Wysocki, V. H. Surface-induced Dissociation Mass Spectrometry as a Structural Biology Tool. *Chem. Rev.* <https://doi.org/10.1021/acs.chemrev.1c00309> (2021).
54. Rolland, A. D. & Prell, J. S. Computational Insights into Compaction of Gas-Phase Protein and Protein Complex Ions in Native Ion Mobility-Mass Spectrometry. *Trends Anal. Chem.* **116**, 282–291 (2019).
55. Rolland, A. D., Biberic, L. S. & Prell, J. S. Investigation of Charge-State-Dependent Compaction of Protein Ions with Native Ion Mobility-Mass Spectrometry and Theory. *J. Am. Soc. Mass Spectrom.* **33**, 369–381 (2022).
56. Webb, B. et al. Integrative structure modeling with the Integrative Modeling Platform. *Protein Sci.* **27**, 245–258 (2018).

57. Landreh, M. et al. Predicting the Shapes of Protein Complexes through Collision Cross Section Measurements and Database Searches. *Anal. Chem.* **92**, 12297–12303 (2020).
58. Kaldmäe, M., Sahin, C., Saluri, M., Marklund, E. G. & Landreh, M. A strategy for the identification of protein architectures directly from ion mobility mass spectrometry data reveals stabilizing subunit interactions in light harvesting complexes. *Protein Sci.* **28**, 1024–1030 (2019).
59. Leaver-Fay, A. et al. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol.* **487**, 545–574 (2011).
60. Lemay, J. K. et al. Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nat. Methods* **17**, 665–680 (2020).
61. Simons, K. T., Kooperberg, C., Huang, E. & Baker, D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.* **268**, 209–225 (1997).
62. Bradley, P., Misura, K. M. & Baker, D. Toward high-resolution de novo structure prediction for small proteins. *Science* **309**, 1868–1871 (2005).
63. Raman, S. et al. Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins* **77**, 89–99 (2009). **Suppl 9**.
64. Song, Y. et al. High-resolution comparative modeling with RosettaCM. *Structure* **21**, 1735–1742 (2013).
65. Alford, R. F. et al. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J. Chem. Theory Comput.* **13**, 3031–3048 (2017).
66. Paizs, B. A divide-and-conquer approach to compute collision cross sections in the projection approximation method. *Int. J. Mass Spectrom.* **378**, 360–363 (2015).
67. Devine, P. W. A. et al. Investigating the Structural Compaction of Biomolecules Upon Transition to the Gas-Phase Using ESI-TWIMS-MS. *J. Am. Soc. Mass Spectrom.* **28**, 1855–1862 (2017).
68. Pacholarz, K. J. et al. Dynamics of intact immunoglobulin G explored by drift-tube ion-mobility mass spectrometry and molecular modeling. *Angew. Chem. Int. Ed. Engl.* **53**, 7765–7769 (2014).
69. Scott, D., Layfield, R. & Oldham, N. J. Structural insights into interactions between ubiquitin specific protease 5 and its poly-ubiquitin substrates by mass spectrometry and ion mobility spectrometry. *Protein Sci.* **24**, 1257–1263 (2015).
70. Hansen, K. et al. A Mass-Spectrometry-Based Modelling Workflow for Accurate Prediction of IgG Antibody Conformations in the Gas Phase. *Angew. Chem. Int. Ed. Engl.* **57**, 17194–17199 (2018).
71. van der Spoel, D., Marklund, E. G., Larsson, D. S. & Caleman, C. Proteins, lipids, and water in the gas phase. *Macromol. Biosci.* **11**, 50–59 (2011).
72. Steinberg, M. Z., Elber, R., McLafferty, F. W., Gerber, R. B. & Breuker, K. Early structural evolution of native cytochrome c after solvent removal. *Chembiochem* **9**, 2417–2423 (2008).
73. Warnke, S., von Helden, G. & Pagel, K. Protein structure in the gas phase: the influence of side-chain microsolvation. *J. Am. Chem. Soc.* **135**, 1177–1180 (2013).
74. Bakhtiari, M. & Konermann, L. Protein Ions Generated by Native Electrospray Ionization: Comparison of Gas Phase, Solution, and Crystal Structures. *J. Phys. Chem. B* **123**, 1784–1796 (2019).
75. Turzo, S.M.B.A. smturzo/IMMS: v.1.0.0, Zenodo (smturzo/IMMS), <https://doi.org/10.5281/zenodo.6726418> (2022).
76. Gray, J. J. et al. Protein–Protein Docking with Simultaneous Optimization of Rigid-body Displacement and Side-chain Conformations. *J. Mol. Biol.* **331**, 281–299 (2003).
77. SE, B. & S, L. Protein Structure Prediction with Mass Spectrometry Data. *Ann. Rev. Phys. Chem.* <https://doi.org/10.1146/annurev-physchem-082720-123928> (2021).
78. Biehn, S. E., Picarello, D. M., Pan, X., Vachet, R. W. & Lindert, S. Accounting for Neighboring Residue Hydrophobicity in Diethylpyrocarbonate Labeling Mass Spectrometry Improves Rosetta Protein Structure Prediction. <https://doi.org/10.1021/jasms.1c00373> (2022).
79. Biehn, S. E., Limpikirati, P., Vachet, R. W. & Lindert, S. Utilization of Hydrophobic Microenvironment Sensitivity in Diethylpyrocarbonate Labeling for Protein Structure Prediction. <https://doi.org/10.1021/acs.analchem.1c00395> (2021).
80. Seffernick, J. T., Canfield, S. M., Harvey, S. R., Wysocki, V. H. & Lindert, S. Prediction of Protein Complex Structure Using Surface-Induced Dissociation and Cryo-Electron Microscopy. *Anal. Chem.* **93**, 7596–7605 (2021).
81. Leelananda, S. P. & Lindert, S. Using NMR Chemical Shifts and Cryo-EM Density Restraints in Iterative Rosetta-MD Protein Structure Refinement. *J. Chem. Inf. Model* **60**, 2522–2532 (2020).
82. Leelananda, S. P. & Lindert, S. Iterative Molecular Dynamics-Rosetta Membrane Protein Structure Refinement Guided by Cryo-EM Densities. *J. Chem. Theory Comput* **13**, 5131–5145 (2017).
83. Marzolf, D. R., Seffernick, J. T. & Lindert, S. Protein Structure Prediction from NMR Hydrogen-Deuterium Exchange Data. *J. Chem. Theory Comput* **17**, 2619–2629 (2021).
84. Nguyen, T. T., Marzolf, D. R., Seffernick, J. T., Heinze, S. & Lindert, S. Protein structure prediction using residue-resolved protection factors from hydrogen-deuterium exchange NMR. *Structure* **30**, 313–320.e313 (2022).
85. Ismail, A. F., Khulbe, K. C. & Matsuura, T. in *Gas Separation Membranes: Polymeric and Inorganic* (eds Ahmad Fauzi Ismail, Kailash Chandra Khulbe, & Takeshi Matsuura) 11–35 (Springer International Publishing, 2015).
86. Berman, H. M. et al. The Protein Data Bank. *Acta Crystallogr. Sect. D.* **58**, 899–907 (2002).
87. Wang, G. & Dunbrack, R. L. Jr PISCES: recent improvements to a PDB sequence culling server. *Nucleic Acids Res* **33**, W94–W98 (2005).
88. Sillitoe, I. et al. CATH: increased structural coverage of functional space. *Nucleic Acids Res* **49**, D266–D273 (2021).
89. Stiving, A. Q., Jones, B. J., Ujma, J., Giles, K. & Wysocki, V. H. Collision Cross Sections of Charge-Reduced Proteins and Protein Complexes: a Database for CCS Calibration. *Analytical Chem.* <https://doi.org/10.1021/acs.analchem.9b05519> (2020).
90. Allen, S. J., Giles, K., Gilbert, T. & Bush, M. F. Ion mobility mass spectrometry of peptide, protein, and protein complex ions using a radio-frequency confining drift cell. *Analyst* **141**, 884–891 (2016).
91. Faull, P. A. et al. Gas-phase metalloprotein complexes interrogated by ion mobility-mass spectrometry. *Int. J. Mass Spectrom.* **283**, 140–148 (2009).
92. Hoaglund-Hyzer, C. S., Counterman, A. E. & Clemmer, D. E. Anhydrous Protein Ions. *Chem. Rev.* **99**, 3037–3080 (1999).
93. Salbo, R. et al. Traveling-wave ion mobility mass spectrometry of protein complexes: accurate calibrated collision cross-sections of human insulin oligomers. *Rapid Commun. Mass Spectrom.* **26**, 1181–1193 (2012).
94. Kim, S. S., Seffernick, J. T. & Lindert, S. Accurately Predicting Disordered Regions of Proteins Using Rosetta ResidueDisorder Application. *J. Phys. Chem. B* **122**, 3920–3930 (2018).
95. Seffernick, J. T., Ren, H., Kim, S. S. & Lindert, S. Measuring Intrinsic Disorder and Tracking Conformational Transitions Using Rosetta ResidueDisorder. *J. Phys. Chem. B* **123**, 7103–7112 (2019).
96. Gront, D., Kulp, D. W., Vernon, R. M., Strauss, C. E. & Baker, D. Generalized fragment picking in Rosetta: design, protocols and applications. *PLoS One* **6**, e23294 (2011).
97. Calvo, F. et al. Statistical analysis of ion mobility spectrometry. II. Adaptively biased methods and shape correlations. *J. Am. Soc. Mass Spectrom.* **23**, 1279–1288 (2012).

98. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
 99. Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
 100. Bhardwaj, G. et al. Accurate de novo design of hyperstable constrained peptides. *Nature* **538**, 329–335 (2016).
 101. Zhang, Y. & Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins* **57**, 702–710 (2004).
 102. Olechnovič, K. & Venclovas, Č. VoroMQA: Assessment of protein structure quality using interatomic contact areas. *Proteins* **85**, 1131–1145 (2017).
 103. Takei, Y. & Ishida, T. P3CMQA: Single-Model Quality Assessment Using 3DCNN with Profile-Based Features. *Bioengineering* **8**, 40 (2021).
 104. Ohio Supercomputer Center. 1987. Ohio Supercomputer Center. Columbus OH: Ohio Supercomputer Center. <http://osc.edu/ark:/19495/f5s1ph73>.
- contributed to the development of the hypotheses, writing of the text as well as supervising this project. S.H. provided guidance in code development. A.D.R., M.T.D., J.S.P. and V.H.W. provided additional data, useful experimental insights and help during the preparation of this manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-022-32075-9>.

Correspondence and requests for materials should be addressed to Steffen Lindert.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022

Acknowledgements

We thank the members of the Lindert lab for many useful discussions. We would like to thank the Ohio Supercomputer Center¹⁰⁴ for valuable computational resources as well as their computational support team including Summer Wang, Samuel Khuvis, Zhiqiang You along with others for their help with issues related specific software installations and usage of the supercomputer. We also thank Alyssa Stiving and Benjamin Jones for the data they collected on beta lactoglobulin, carbonic anhydrase, ubiquitin, and serum albumin. Additionally, we would like to thank Vikram K. Mulligan, Rocco Moretti, Julia K. Leman, Steven Lewis along with many others in the Rosetta community for their help and guidance during the code development stage of the PARCS applications and IM score function as well as reviewing the final version of all code pushed into Rosetta. This work was supported by NIH (P41 GM128577 to V.H.W. and S.L., R21A125804 to J.S.P.). Additionally, integrative protein modelling was supported by a Sloan Research Fellowship to S.L.

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., J.S.P., V.H.W., and S.L.